



Katja Wengler, Dirk Reichardt, Dennis Pfisterer (Eds.)

DHBW AI Transfer Congress 2026

Proceedings

Imprint

Stuttgart, April 2026

DHBW AI Transfer Congress

ISSN: 2976-1735

Published by:

Duale Hochschule Baden-Württemberg
Baden-Wuerttemberg Cooperative State University
Friedrichstraße 14, 70174 Stuttgart
represented by: Prof. Dr. Martina Klärle

Photograph (page 2) by Michael Stifter

Photograph (page 3) by Fotofabrik Stuttgart

Table of Contents

Editors	1
President DHBW	2
Vicepresident DHBW	3
Keynote	4
‘Rethinking Mobility: AI-Based Pricing Systems To Promote Fair Mobility Participation In Cities’	5
<i>Madeleine Neumann, Vanessa Reit, Sarah Selinka, Sinu Thirukketheeswaran and Marc Kuhn</i>	
Systematic Analysis of the Evaluation of Video World Models as Data Engines	10
<i>Max Zeitler and Maximilian Pazer</i>	
Exploring Research Supervisors’ Perceptions of an Artificial Intelligence-Driven Virtual Reality Avatar Prototype: Insights from the REACH-ETHICS Project	21
<i>Marco Montalto, Lorna Bonnici West, Daren Scerri, Alan Gatt, David Deguara, David Debono, Marisa Galea Vella, Lisa Theuma, Francis Delicata and Shirley Ann Gauci</i>	
Assessing students’ performance and supporting building a professional identity by writing amidst the rise of AI	30
<i>Ema Jerkovic</i>	
On Hubness Effects in Retrieval-Augmented Generation Systems	37
<i>Alexandros Nanopoulos, Krisztian Buza and Margit Antal</i>	
An LLM-Assisted Framework for Improving the Robustness of Machine Learning Models to Label Noise	47
<i>Ziming Fang, Marcel Dix, Yuliang Ma and Andrey Morozov</i>	
From Ethical Principles to Design Practice: Implementing Trustworthy AI in the AUT-KI Project	55
<i>Daniela Rauen</i>	
Beyond Bandwidth: Profiling Software-Level Communication Bottlenecks in Distributed Large Language Model Training	63
<i>Maximilian Pazer, Max Zeitler and Hardik Soni</i>	
LLM-Based Knowledge Graph Construction for Curriculum Analysis - Automated Skill Extraction and ESCO Alignment	73
<i>Florian Pallmann and Franziska Schütz</i>	
From Prediction to Causation: A Comparison of SHAP and Double Machine Learning	83
<i>Markus Niedergesäss and Giacomo Welsch</i>	
The ADM-AI Model: Agile Development Model for AI Competences in European Higher Education Curricula	93
<i>Ulf Ehlers and Julia Marcia Mann</i>	

From Lead to Loyalty: A Systematic Review of AI Use Cases for SME along the Customer Journey	106
<i>Tessa Steinigke and Patrick Föll</i>	
Discovering skill requirements using AI: DataGEMS, an EU Horizon Project	116
<i>Shobhit Agarwal and Friedemann Schwenkreis</i>	
Author Index	126
Keyword Index.....	128
Program Committee	130
Book of Abstracts.....	131

Editors



Prof. Dr. Dennis Pfisterer
Program Chair
DHBW AI Transfer Congress



Prof. Dr. Dirk Reichardt
Program Chair
DHBW AI Transfer Congress



Prof. Dr. Katja Wengler
Program Chair
DHBW AI Transfer Congress

This volume presents the conference proceedings of the 4th AI Transfer Congress, organized by the Cooperative State University Baden-Württemberg (DHBW), held at the Bildungscampus in Heilbronn, Germany.

We would like to express our special thanks to Prof. Dr. Patrick Föll, who joined the congress for the first time as Local Chair and contributed significantly to the successful organization of the event on site in Heilbronn.

A distinctive feature of this year's congress is the large number of workshops explicitly designed to foster transfer between research and practice. These interactive sessions emphasize hands-on learning, direct exchange with practitioners, and the application of AI methods in real-world contexts. Topics include, among others, machine learning, embedded AI, generative AI, retrieval augmented generation, leadership in the age of AI, and AI applications in marketing.

Inaugurated in 2022, the conference aims to establish a transfer platform fostering exchange between academia and practical applications in the field of Artificial Intelligence (AI). The event features a conference track that explores current research in applied AI, and a workshop track that provides extensive opportunities for training, collaboration, and knowledge transfer. This structure strengthens the dialogue between academic and industrial researchers, as well as users and practitioners interested in AI.

The presentations within the conference track reflect a broad yet clearly structured spectrum of topics, particularly in the areas of business applications, data science, ethical considerations of AI, and educational aspects. This thematic focus highlights both the technological advancements and the societal implications of AI.

The review process was overseen by an internationally diverse program committee of experienced practitioners and scientists. We extend our gratitude to the authors and reviewers for their outstanding contributions.

Finally, we thank the members of the DHBW university administration, and the communication departments for their dedicated efforts in making the AI Transfer Congress a success.



Liebe Alle, die aus KI echte Anwendung machen,

Den AI-Transfer-Kongress gibt es nun seit vier Jahren an der DHBW. Dass wir ihn damals auf den Weg bringen konnten, freut mich heute mehr denn je. Was ich allerdings nicht erwartet habe: mit welcher Geschwindigkeit sich das Thema Künstliche Intelligenz entwickelt. Es ist kein Trend, es ist ein Dauerrausch – mit enormer Dynamik und ebenso großer Verantwortung.

Gerade in dieser Entwicklung liegt unsere besondere Stärke als DHBW: Wir sind die Transferhochschule. Wir verbinden Theorie mit Praxis, wissenschaftliche Erkenntnis mit konkreter Anwendung – gemeinsam mit unseren dualen Partnern. Genau hier entscheidet sich, ob aus technologischen Möglichkeiten echte Innovation wird.

Der AI-Transfer-Kongress ist dafür ein zentraler Ort geworden – für Austausch, für Einordnung und für gemeinsames Weiterdenken. Mein besonderer Dank gilt Herrn Prof. Dr. Dirk Reichardt, der diesen Kongress von Beginn an mit großem Engagement gestaltet und geprägt hat.

Dieses Kongressheft bündelt die Themen, Perspektiven und Impulse, die diesen Austausch tragen – und zeigt, mit welcher Dynamik und welchem Anspruch sich die DHBW der Künstlichen Intelligenz widmet.

Lassen Sie uns auf dem AI Transfer Congress dafür sorgen, dass aus Gedanken Lösungen werden – und aus Lösungen echte Anwendung

Ihre Prof. Dr. Martina Klärle

Präsidentin der DHBW



Dear participants of the AI Transfer Congress,

artificial intelligence is a key driver of innovation - especially for small and medium-sized enterprises, which face the challenge of unlocking its potential in a targeted and practical way.

A crucial factor in this process is the transfer of applied AI expertise from research into practice. Applied research plays a central role here: it addresses concrete challenges from industry and develops practical AI solutions in close collaboration with companies. In this way, direct added value and sustainable innovative capacity are created. In addition, continuing academic education and life-long learning is of great importance in equipping professionals and managers with the skills needed to apply AI competently and responsibly over the long term.

The Baden-Wuerttemberg Cooperative State University (DHBW) plays an important bridging role between academia, industry, and continuing education. Our AI Transfer Congress provides an ideal platform for exchange, networking, and new impulses. Its goal is to map current research results and educational offers to the concrete requirements of industry and society. Gaps can be identified and future collaborations can be initiated in order to close them.

I wish you inspiring discussions and many valuable practical insights.

Yours sincerely,

Dirk Brinkmann

Vice President for Research, Innovation and Internationalization
of the Baden-Wuerttemberg Cooperative State University



Keynote

Use AI to Stop the Information Flood – Smart Minds Need Space for Clear Decisions

Imagine you work in regulatory affairs in MedTech and your Monday morning starts with a full inbox. New regulations have appeared or been revised, and the organization needs to change immediately, but in which direction?

There might be a new regulation for medical devices in Europe, three FDA guidelines in the US and a revised ISO standard. In addition, product changes must be designed and implemented in order to continue marketing them on global markets. The same applies to products, such as the immediate change required to continue marketing them in specific markets. And the week hasn't even started yet! The flood is real, and it's not slowing down.

But here's what no one really talks about: the problem has never been the volume. The problem is what that volume does to us. It pulls the brightest minds in the room into an endless cycle of reading, sorting, and summarizing, even though their real value lies in thinking, deciding, and leading. At some point, regulatory intelligence was no longer about strategy, but about survival. That ends now.

We're going to talk openly about why the status quo is not only inefficient, but also a genuine strategic risk. We will examine what organizations that have mastered this challenge are doing differently. We will discuss how we can use AI so efficiently that regulatory teams can once again focus fully on their core tasks.

Because when you're no longer drowning in information, you can see the horizon again. That changes everything: how you plan, how you advise leadership, and how quickly you act when it matters.

It's not about keeping up. It's about getting ahead.

Sarah Haake-Schäfer

Head of Global Regulatory Strategic Innovation
Carl Zeiss Vision International GmbH

‘Rethinking Mobility: AI-Based Pricing Systems To Promote Fair Mobility Participation In Cities’

Neumann, Madeleine
Centre for Empirical Research
Baden-Wuerttemberg Cooperative State
University Stuttgart,
Stuttgart, Germany
Madeleine.Neumann@dhbw-
stuttgart.de

Thirukketheeswaran, Sinu
Centre for Empirical Research
Baden-Wuerttemberg Cooperative State
University Stuttgart,
Stuttgart, Germany
Sinu.Thirukketheeswaran@dhbw-
stuttgart.de

Reit, Vanessa
Centre for Empirical Research
Baden-Wuerttemberg Cooperative State
University Stuttgart,
Stuttgart, Germany
Vanessa.Reit@dhbw-stuttgart.de

Kuhn, Marc
Centre for Empirical Research
Baden-Wuerttemberg Cooperative State
University Stuttgart,
Stuttgart, Germany
Marc.Kuhn@dhbw-stuttgart.de

Selinka, Sarah
Centre for Empirical Research
Baden-Wuerttemberg Cooperative State
University Stuttgart,
Stuttgart, Germany
Sarah.Selinka@dhbw-stuttgart.de

Abstract—The study examines AI's role in pricing shared mobility services and its potential to promote fair mobility in urban areas. Urbanization exacerbates social inequalities, limiting mobility access for disadvantaged groups. AI-driven systems, such as shared autonomous vehicles (SAVs), are seen as solutions for enhancing accessibility and sustainability. This study explores how socio-demographic factors and social constructs like inclusion and participation influence AI pricing. Survey data from 89 German respondents reveals income, number of children, caregiver status, age, and travel purpose significantly impact pricing, while inclusion and participation have no effect. Higher income and age increase prices, while caregiving and more children reduce fares. These findings indicate AI prioritizes demographic data over nuanced social equity measures. Future research will use simulations to refine AI models and assess public acceptance of socially oriented pricing systems.

Keywords—*AI-driven Mobility Pricing, Social Equity, Shared Autonomous Vehicles*

I. INTRODUCTION (HEADING 1)

The ongoing urbanization leads to increasing social segregation and marginalization of certain population groups [1]. Socio-economically disadvantaged people, who often have less access to mobility services, are particularly affected [2,3]. This unequal participation exacerbates social inequalities in urban areas [4]. The introduction of AI-supported, shared mobility services, such as SAV, is seen as a solution to make mobility more efficient [5,6] and promote sustainability [7]. Research emphasizes that the transformative potential of AI for building a more inclusive and equitable society is significant when we focus on equality, social justice, and human welfare in its development and implementation [8]. On the other hand, there is research that highlights that the use of AI algorithms may further reinforce social inequalities, for example, by learning and perpetuating existing biases from historical data, applying models inaccurately across different contexts, and lacking standardized ethical evaluations, all of which can unintentionally disadvantage marginalized groups [9]. But how does AI currently make decisions? Do the underlying structures that trained the AI truly ensure that marginalized groups are granted greater participation, for

example, through a pricing model for the use of a mobility service adapted to social circumstances? And if so, what characteristics lead to certain groups being offered a lower price for mobility use? This work aims to investigate the relationship between social data and AI decision-making in urban mobility, as these factors can significantly influence the distribution of mobility services and their pricing. This paper therefore deals with the question of which social characteristics should be given particular importance in the context of AI-controlled pricing for shared mobility services. The theoretical background is therefore first discussed below, with a focus on the exploratory nature of the study. This is followed by a description of the survey method and the presentation of the regression results. Finally, the results are discussed, limitations are pointed out and the need for future research based on this explorative preliminary study is presented.

II. BACKGROUND

With AI-facilitated mobility, social inclusion and participation become increasingly important, as intelligent technologies could not only enhance access to mobility services but also create opportunities for social engagement and participation across all segments of the population, particularly for disadvantaged and mobility-impaired individuals. Hidayati et al. [3] highlight how mobility inequalities result from a combination of intrinsic (individual attributes such as income, gender, and age) and extrinsic (socio-cultural) factors.

Regarding socio-cultural factors that can influence mobility inequalities, social inclusion and participation appear to be interesting influencing factors.

Social inclusion describes the process by which individuals gain access to essential resources and opportunities for full participation in social, cultural, and economic life [10]. Social participation refers to the active involvement of individuals in societal activities, enhancing their sense of belonging, self-efficacy, and self-worth [11].

Research on sharing economy platforms has shown that these systems could promote social inclusion by facilitating access to key resources, helping users achieve satisfactory living standards through community integration [10]. Curtis et al. [12] seek to deepen the understanding of how sharing platforms impact society, arguing that while the sharing economy theoretically promotes social inclusion, the actual mechanisms to achieve this goal are not yet fully understood. Although these studies have made significant progress in exploring the general impact of the sharing economy, there remains a gap in understanding how shared mobility services, such as car sharing and ridesharing, specifically address social inequalities and promote broader social equity.

Socio-demographics are generally identified in research as an influencing factor for mobility use and participation [13]. This research therefore takes numerous socio-demographic variables into account. Of particular interest are those that can be an indicator of whether a person may be socially disadvantaged. Specifically, these are, for example, the variables household size, age, number of children in the household, income, caregiver status, educational and occupational background.

These and the previously described constructs of social inclusion [10] and social participation [11] serve as key factors for us to investigate the current decisions of an AI in relation to the pricing of shared mobility services

III. METHODOLOGY

The research was designed as an exploratory, impact orientated analysis. A questionnaire was developed and distributed online to a sample of the German population. In addition to the socio-demographic aspects, the constructs of social inclusion [10] and social participation [11] were operationalized using 12 and 5 corresponding items respectively. An overall index was then calculated for each construct based on the individual assessment of the items using a 5-point Likert scale (see Appendix).

After the completion of the field phase, the dependent variable of the potential mobility price was generated by using a two way approach in python. First, the data from the online survey, containing each respondents' data was automatically extracted, decoded and transformed and stored in an initial prompt file. A prompt is a text-based query, especially used for communication with large language models, such as GPT-4 [14]. We followed the current best practices for prompt engineering for creating the prompts [14]. After the first step, the initial prompt file was then processed through the GPT-4 API to analyze the social profile of each respondent through the AI. GPT-4 was chosen as a large language model due to its extensive language understanding capabilities for psychological constructs [15]. Afterwards the potential mobility price was determined. This price represents the dependent variable of the experiment and is analyzed and evaluated in the further course of the study.

IV. RESULTS

The quantitative data was analyzed with Jamovi 2.3. The basis of the data analysis is a sample of 89 test persons (78.8% female, 21.2% male). The average age is 36.31 years.

A linear regression was used to evaluate the research question (s. Table 1). The regression is based on 80 data sets. All requirements of the regression were tested (outlier analysis, multicollinearity, normal distribution, linearity, autocorrelation) and are fulfilled. The model test with total price of a shared autonomous shuttle as dependent variable and different sociodemographic factors as predictors is significant with $F(33, 46) = 6.823$, $p < 0.001$, the adjusted R^2 is $= 0.709$.

As assumed, there is a significant influence on pricing by AI for some influencing factors. Interestingly, the social constructs entered have no influence on the AI's price decision. The factors income, number of children, reason for mobility, caregiver and age are the factors that appear to have a significant influence on the AI's price decision, in order of strength of influence.

The higher the test person's income, the more expensive the journey. The more children a person has, the cheaper the ride is offered by the AI. For important and urgent reasons, the journey is more expensive than for normal appointments or reasons. The effect is particularly interesting for people who care for another person. These people have to pay less for the journey than people who are not carers. The older the person, the more the AI charges for the journey.

Predictor	Est.	SE	t	p	Stand. Est.
Intercept	-7.483	5.810	-1.288	0.204	
Gender:					
male - female	0.808	0.876	0.922	0.361	0.181
Formal School Graduation:					
Secondary School - Basic School	3.320	2.914	1.139	0.260	0.743
High School Diploma - Basic School	2.391	2.730	0.876	0.386	0.535
Formal Vocational Training:					
Vocational Training - No vocational train.	3.553	2.783	1.277	0.208	0.795
University Degree - No vocational train.	0.980	2.873	0.341	0.734	0.219
PhD - No vocational train.	-1.132	3.152	-0.359	0.721	-0.253

Current Occupational Status:						
Part-time Education	-	In	-0.676	3.125	-0.216	0.830 -0.151
Full-time/Self-employed Education	-	In	-2.211	3.192	-0.693	0.492 -0.495
Unemployed Education	-	In	-0.819	3.835	-0.213	0.832 -0.183
Retired Education	-	In	-2.296	4.283	-0.536	0.594 -0.514
Marital Status:						
Married/Registered Partnership - Single			0.614	1.008	0.609	0.546 0.137
Widowed - Single			-5.633	3.170	-1.777	0.082 -1.261
Divorced - Single			-2.940	3.014	-0.976	0.334 -0.658
People in Household:						
2 - 1			1.406	1.142	1.231	0.225 0.315
3 - 1			0.643	2.560	0.251	0.803 0.144
4 - 1			2.762	2.324	1.188	0.241 0.618
5 or more - 1			3.343	2.962	1.129	0.265 0.748
Number of Children:						
1 - 0			-2.845	2.791	-1.019	0.313 -0.637
2 - 0			-6.738	2.990	-2.254	0.029 -1.509
3 or more children - 0			-8.960	3.627	-2.471	0.017 -2.006
Children Living in Household:						
yes - no			3.469	2.834	1.224	0.227 0.777
Caregiver:						
yes - no			-3.305	1.130	-2.925	0.005 -0.740
Disability:						
yes - no			0.987	2.191	0.450	0.655 0.221
Member of a (non-profit) association:						
yes - no			-0.081	0.728	-0.112	0.912 -0.018

Travel Reason:						
Important reason - normal reason	-		3.067	0.974	3.150	0.003 0.687
Urgent reason - normal reason	-		5.127	0.985	5.205	< .001 1.148
Income:						
25.001-50.000 € - up to 25.000 €	-		4.727	1.657	2.853	0.006 1.058
50.001-75.000 € - up to 25.000 €	-		6.038	1.523	3.965	< .001 1.352
75.001-100.000 € - up to 25.000 €	-		10.881	1.853	5.873	< .001 2.436
over 100.000 € - up to 25.000 €	-		12.198	2.186	5.580	< .001 2.731
Age (years)			0.228	0.074	3.068	0.004 0.549
social_participation_avg			0.426	0.749	0.569	0.572 0.063
social_inclusion_avg			-0.263	0.605	-0.436	0.665 -0.045

Table 1: Results of regression with the total price of a shared autonomous shuttle as dependent variable and different (sociodemographic) factors as predictors

V. DISCUSSION

The primary aim of this study was to determine whether social participation and social inclusion, in addition to classic socio-demographic factors, influence AI-driven pricing for shared mobility services. Interestingly, the two theoretical constructs, which are intended to give an impression of the social situation or the respondents' own perception of it, show no significant influence. This finding suggests that the AI may prioritize quantifiable and structured data, such as income levels and household composition, over subjective measures. This raises the question of the extent to which current AI systems are able to integrate softer, more nuanced indicators of social context. The situation is different for some socio-demographic variables, which can also be seen as indicators of social circumstances. The issue of having children seems to be of particular importance for AI. Interestingly, however, it does not matter whether the children also live in the same household. The pure size of the household, which is presumably related to the number of children in many cases, does not appear to be a decisive factor either. Caring for relatives, on the other hand, again has a significant effect on the price model. The implicit idea of 'taking care' in the case of having children and caring for relatives seems to be a factor for AI that entitles users to be offered a socially assisted ride more favorably than others. Existing income, on the other hand, is perceived as a clear factor by the AI that a person has to pay more the more he or she actually earn. Age is also a highly significant influencing factor on price modeling by AI, suggesting that it recognizes financial capacity as a justification for differential pricing. This reinforces the perception of the AI as a tool for

progressive pricing, which might aim to balance affordability and revenue optimization. According to the model, the prices demanded rise with increasing age. This is interesting, because from a social perspective it could have been assumed that lower prices would be offered for older people, e.g. with regard to being a pensioner.

VI. CONCLUSIONS AND FURTHER RESEARCH

The present work is a preliminary study. It has been shown that some influencing factors have a significant impact on the AI-generated pricing of the mobility offer of interest. Contrary to expectations, factors social participation and inclusion do not appear to make a significant contribution to the pricing of a mobility service by AI. For the further process, it must be examined which other factors could be significant in addition to socio-demographics. It must also be evaluated whether the measurement methodology for social participation and inclusion may have been decisive for their non-significance. Based on the knowledge gained, the influencing factors will be investigated further as part of a simulation study. The aim is to make a socially oriented mobility offer tangible with the help of a driving simulator and to research the acceptance of AI-generated pricing by the test subjects. This step is important because the existing landscape is mainly based on pure surveys or scenarios that are to be evaluated by test persons [16,17]. However, it has already been shown that direct experience is important to consider the barriers to acceptance in AI technologies [18]. To overcome these barriers, comprehensive simulation environments can be used to investigate the holistic perspective of AI in a wide variety of fields of action, such as AI-driven pricing of shared mobility services. Compared to this preliminary study, even greater attention must be paid to ensuring that the sample is as representative as possible of the German population. For this case men are over-represented in the sample. A balanced relation between women and men is desirable for future research. Furthermore, the characteristics of age and income should also be better addressed in further work.

REFERENCES

[1] Yue, X., Wang, Y., Li, W., Wu, Y., Wang, Y., Zhang, H., & Ma, Z. (2024). Research Progress and Trends in Urban Residential Segregation. *Buildings*, 14(7), 1962. <https://doi.org/10.3390/buildings14071962>

[2] Guzman, L. A., Oviedo, D., & Rivera, C. (2017). Assessing equity in transport accessibility to work and study: The Bogotá region. *Journal of Transport Geography*, 58, 236–246. <https://doi.org/10.1016/j.jtrangeo.2016.12.016>

[3] Hidayati, I., Tan, W., & Yamu, C. (2021). Conceptualizing Mobility Inequality: Mobility and Accessibility for the Marginalized. *Journal of Planning Literature*, 36(4), 492–507. <https://doi.org/10.1177/08854122211012898>

[4] Van Ham, M., Tammaru, T., Ubarevičienė, R., & Janssen, H. (2021). Rising Inequalities and a Changing Social Geography of Cities. An Introduction to the Global Segregation Book. In M. Van Ham, T. Tammaru, R. Ubarevičienė, & H. Janssen (Hrsg.), *Urban Socio-Economic Segregation and Income Inequality* (S. 3–26). Springer International Publishing. https://doi.org/10.1007/978-3-030-64569-4_1

[5] Garikapati, D., & Shetiya, S. S. (2024). *Autonomous Vehicles: Evolution of Artificial Intelligence and Learning Algorithms* (arXiv:2402.17690). arXiv. <http://arxiv.org/abs/2402.17690>

[6] Mahrez, Z., Sabir, E., Badidi, E., Saad, W., & Sadik, M. (2022). Smart Urban Mobility: When Mobility Systems Meet Smart Data. *IEEE*

Transactions on Intelligent Transportation Systems, 23(7), 6222–6239. <https://doi.org/10.1109/TITS.2021.3084907>

[7] Munhoz, P. A. M. S. A., Da Costa Dias, F., Kowal Chinelli, C., Azevedo Guedes, A. L., Neves Dos Santos, J. A., Da Silveira E Silva, W., & Pereira Soares, C. A. (2020). Smart Mobility: The Main Drivers for Increasing the Intelligence of Urban Mobility. *Sustainability*, 12(24), 10675. <https://doi.org/10.3390/su122410675>

[8] Farahani, M. S., & Ghasemi, G. (2024). Artificial Intelligence and Inequality: Challenges and Opportunities. *Qeios*. <https://doi.org/10.32388/7HWUZZ>

[9] Ferrer, X., Nuenen, T. V., Such, J. M., Cote, M., & Criado, N. (2021). Bias and Discrimination in AI: A Cross-Disciplinary Perspective. *IEEE Technology and Society Magazine*, 40(2), 72–80. <https://doi.org/10.1109/MTS.2021.3056293>

[10] Davlembayeva, D., Papagiannidis, S., & Alamanos, E. (2020). Sharing economy: Studying the social and psychological factors and the outcomes of social exchange. *Technological Forecasting and Social Change*, 158, 120143. <https://doi.org/10.1016/j.techfore.2020.120143>

[11] Berger, U., Kirschner, H., Muehleck, J., Gläser, A., Werner, B., Kurz, M., Schwager, S., Wick, K., & Strauß, B. (2020). Kurz-Skala zur Erfassung wahrgenommener sozialer Teilhabe (KsT-5): faktorielle Struktur, interne Konsistenz, inhaltliche und konvergente Validität sowie Normwerte in einer repräsentativen Stichprobe. *PPmP - Psychotherapie · Psychosomatik · Medizinische Psychologie*, 70(09/10), 396–404. <https://doi.org/10.1055/a-1088-1354>

[12] Curtis, S. K., Singh, J., Mont, O., & Kessler, A. (2020). Systematic framework to assess social impacts of sharing platforms: Synthesising literature and stakeholder perspectives to arrive at a framework and practice-oriented tool. *PLOS ONE*, 15(10), e0240373. <https://doi.org/10.1371/journal.pone.0240373>

[13] O'Driscoll, C., Crowley, F., Doran, J., & McCarthy, N. (2024). How the relationship between socio-demographics, residential environments and travel influence commu. *REGIONAL STUDIES*.

[14] Ekin, S. (2023). Prompt Engineering For ChatGPT: A Quick Guide To Techniques, Tips, And Best Practices. <https://doi.org/10.36227/techrxiv.22683919.v2>

[15] Demszky, D., Yang, D., Yeager, D. S., Bryan, C. J., Clapper, M., Chandhok, S., Eichstaedt, J. C., Hecht, C., Jamieson, J., Johnson, M., Jones, M., Krettek-Cobb, D., Lai, L., Jones Mitchell, N., Ong, D. C., Dweck, C. S., Gross, J. J., & Pennebaker, J. W. (2023). Using large language models in psychology. *Nature Reviews Psychology*. <https://doi.org/10.1038/s44159-023-00241-5>

[16] Cloarec, J., Meyer-Waarden, L., & Munzel, A. (2022). The personalization–privacy paradox at the nexus of social exchange and construal level theories. *Psychology & Marketing*, 39(3), 647–661. <https://doi.org/10.1002/mar.21587>

[17] Yigitcanlar, T., Han, H., Kamruzzaman, Md., Ioppolo, G., & Sabatini-Marques, J. (2019). The making of smart cities: Are Songdo, Masdar, Amsterdam, San Francisco and Brisbane the best we could build? *Land Use Policy*, 88, 104187. <https://doi.org/10.1016/j.landusepol.2019.104187>

[18] Hong, A., Nam, C., & Kim, S. (2020). What will be the possible barriers to consumers' adoption of smart home services? *Telecommunications Policy*, 44(2), 101867. <https://doi.org/10.1016/j.telpol.2019.101867>

[19]

Appendix

Table 2: Constructs and Items

Construct	Items (5-point Likert scale)
[10] Social Inclusion	I have more than enough money to always afford enough food.
	I have unrestricted access to high-quality childcare and general care facilities.
	I can easily take out a loan of any amount at any time.
	I have unrestricted access to all public services at all times.
	I have full access to first-class healthcare.
	I can get any medical help I need immediately and without waiting.

	I can easily afford all transportation costs at any time.
	I have unrestricted access to all community facilities.
	I am fully and completely economically active.
	I have unrestricted access to all financial services at all times.
	I have unrestricted access to all available educational opportunities.

	I have access to all available transportation at all times and everywhere.
Social Participation [11]	I have the feeling that I am part of our society.
	I have an influence on how our society develops.
	I make a valuable contribution to our society.
	I feel well integrated in the social environment outside my family.
	I feel connected to the region in which I live.

AI Disclosure

The authors used GPT-4 (OpenAI) for data processing; all outputs were reviewed and validated by the authors

Systematic Analysis of the Evaluation of Video World Models as Data Engines

Max Zeitler

Baden-Wuerttemberg Cooperative State University Stuttgart
Hewlett Packard Enterprise
Boeblingen, Germany
zeitler@hpe.com

Maximilian Pazer

Baden-Wuerttemberg Cooperative State University Stuttgart
Hewlett Packard Enterprise
Boeblingen, Germany
maximilian.pazer@hpe.com

Abstract—Video World Models (VWMs) represent a paradigm shift in autonomous systems development, functioning as Data Engines to generate synthetic training data. However, a critical discrepancy exists: visual realism does not imply physical understanding. While models can generate visually stunning scenes, they often fail to obey fundamental physical laws, rendering them unreliable for safety-critical applications. This paper addresses the fragmented evaluation landscape through a Systematic Literature Review, synthesizing state-of-the-art approaches into a hierarchical four-level framework: (1) Perceptual Quality, (2) Spatiotemporal Coherence, (3) Plausibility & Controllability, and (4) Transferability & Utility. Our analysis highlights the limitations of traditional pixel-based metrics and documents a decisive shift toward model-based evaluation paradigms, such as AI Judges, which are essential for assessing complex properties like physical causality. This framework provides researchers and practitioners with a structured roadmap for holistic VWM evaluation, moving beyond superficial visual fidelity toward verifying behavioral correctness.

Index Terms—Video World Models, Data Engines, Generative AI, Evaluation Framework, Physical AI, Autonomous Systems, Systematic Literature Review, Benchmarks, FVD, AI Judges

I. INTRODUCTION

Recent breakthroughs in generative artificial intelligence are opening a fundamentally new path to realize a long-standing scientific vision. The research focus is shifting away from the labor-intensive manual creation of simulations toward the ability to learn complex and interactive worlds directly from visual data [1]. Central to this approach is the Video World Model (VWM), a specialized generative model designed to learn and simulate the underlying physical and causal dynamics of our world [2]. These models are trained predictively to forecast future outcomes based on past observations and actions [3], positioning them as a key technology for the next generation of autonomous systems.

In the paradigm of “Physical AI”, also referred to as “Embodied AI”, VWMs are expected to overcome a fundamental hurdle: the scarcity of high-quality, diverse, and safe training data. Rather than training robots or autonomous vehicles expensively and potentially dangerously in the real world, they could learn efficiently within the rich simulations provided by a VWM [4]. The capability to generate physically plausible synthetic data has thus become not merely a technical

milestone but a strategic necessity to accelerate progress in robotics and autonomous driving.

Recent advances, particularly through models such as OpenAI’s Sora [5], have demonstrated that the visual quality of AI-generated videos has reached an impressive level, sparking public debate about “World Simulators” [6]. However, current research literature paints a considerably more critical picture, revealing a significant gap between visual excellence and physical correctness. A comprehensive empirical study by Motamed et al. (2025) concludes that in today’s video models, visual realism does not correlate with physical understanding [7]. Models can thus generate scenes that *look* real but do not *behave* realistically. A further analysis by Kang et al. (2025) provides an explanation for this phenomenon: even massive scaling of data and model sizes does not lead models to learn generalizable physical laws [8]. Instead, they exhibit case-based generalization behavior, merely imitating known patterns from training data rather than abstracting physical principles. This study utilizes the concept of Out-of-Distribution (OOD) generalization to assess whether these models can perform beyond their training scenarios.

This fundamental discrepancy poses an evaluative challenge for the scientific community. In response, research has begun to extend evaluation practice beyond standard metrics such as Fréchet Video Distance (FVD) [9]. However, the current research design suffers from a critical weakness: it is highly fragmented [10]. Individual research groups develop isolated, domain-specific methods [11], benchmarks, or specific metrics for sub-problems such as object permanence or causality [12]. While these approaches represent important first steps, they remain isolated solutions [13]. A comprehensive, systematic framework that consolidates these scattered approaches is lacking [10].

The research gap is therefore not the complete absence of evaluation methods, but rather the absence of a systematic synthesis that creates, from the individual components, a coherent and applicable overall picture of VWM evaluation.

This paper addresses precisely this gap. The objective is deliberately not to add another isolated metric to an already confusing field. Given the fragmentation described above, the scientific contribution lies in synthesizing existing evaluation approaches for the first time into a coherent, multidimensional

framework through a Systematic Literature Review (SLR). This work thus creates the methodological foundation for fairly and purposefully evaluating the physical plausibility of current and future VWMs.

The central research question is: *What structured overall picture of evaluation dimensions and associated metrics can be systematically derived from the State-of-the-Art (SOTA) of VWMs?*

This investigation is supported by three sub-questions:

- **RQ1:** What are the relevant evaluation dimensions?
- **RQ2:** What metrics and benchmarks can be assigned to these dimensions?
- **RQ3:** What are the central limitations of current evaluation practice?

Beyond its scientific contribution, this framework serves an explicit transfer objective: it provides practitioners in industry with actionable guidance for selecting appropriate evaluation methods when deploying VWMs in safety-critical applications. By systematically mapping evaluation dimensions to their limitations and appropriate use cases, this work bridges the gap between research advances and practical industrial implementation, enabling informed decision-making in autonomous systems development.

II. THEORETICAL BACKGROUND

A. Generative Video Models: State of Research

Generative video models constitute a class of machine learning systems whose core function is learning the underlying distribution of video data to synthesize new, realistic video sequences [2]. The overarching goal of a Video Generation Model (VGM) is the synthesis of image sequences exhibiting high spatial and temporal consistency [14]. Precisely maintaining this spatiotemporal consistency represents the central technical challenge and simultaneously the crucial distinction from pure image synthesis [15].

Early approaches were predominantly characterized by Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs), which compress high-dimensional data into a latent space [14]. GAN-based methods utilized early Convolutional Neural Networks (CNNs), architectures that process visual data through hierarchical filters for detecting local patterns [16]. These were extended to spatio-temporal 3D-CNNs using techniques such as foreground-background decoupling or content-motion decomposition [17]. Despite these advances, GANs suffered from fundamental problems including training instability and limited resolution and long-term coherence [14], while VAEs often produced softer and blurrier outputs [15].

A paradigm shift occurred with the adaptation of Transformers for video generation [18]. Their self-attention mechanism enables modeling global dependencies across space and time, a capability that CNNs with their locally constrained receptive field possess only to a limited extent [19]. Approaches such as processing spatio-temporal tokens in transformer layers or

autoregressive generation could thereby effectively improve the coherence of longer sequences [20].

The years 2020–2022 marked the breakthrough of diffusion models, achieving image quality comparable to GANs [21]. Their core principle is a Markov process that learns to reverse a diffusion process: while a “forward process” destroys data through stepwise noise addition, a “reverse process” (denoising) is trained to iteratively reconstruct the original data from noise. Starting from pure noise (x_T), the model p_θ learns to reconstruct the original image (x_0) through iterative, learned denoising steps [22]. A decisive factor for their practicability was the development of more efficient sampling methods such as Denoising Diffusion Implicit Models (DDIM), which significantly accelerated the initially extremely slow iterative denoising process [23].

The most recent development phase since 2023 is characterized by the convergence of diffusion models and transformer architecture, a fusion resulting in so-called Diffusion-Transformers (DiTs) that set new standards in quality and coherence [21]. Prominent examples include Sora from OpenAI and Veo from Google. The architecture for videos is extended to Spatio-Temporal Diffusion Transformers (ST-DiTs) operating on spacetime patches [24]. Scaling these VGMs additionally leads to emergent properties, such as significantly improved generalization capability for longer video durations, higher resolutions, and variable aspect ratios.

Despite rapid progress, current research reveals a fundamental discrepancy: visual realism does not imply physical understanding [7]. This forces a reorientation of the actual research objective. The mere synthesis of aesthetically pleasing videos recedes in favor of developing models that can serve as functional World Models [25]. The focus shifts toward architectures explicitly trained to predict the consequences of actions in an environment, thereby learning a form of intuitive physical understanding [26].

B. World Models: Foundations and Formalization

The term World Model (WM) was mentioned as early as Sutton (1991) in the context of the Dyna architecture [27], but gained widespread adoption through the seminal work of Ha and Schmidhuber (2018) [28]. At its core, a WM’s task is to learn an internal model of environmental dynamics [28]. According to Huang (2025), a WM essentially learns a function that, based on the current state and an executed action, predicts the subsequent state of the environment [29].

WMs can be categorized according to a recent formal taxonomy by Kong et al. (2025) into two complementary paradigms based on their modeling objective [4]:

- 1) **Generative World Models:** Their goal is synthesizing plausible 3D/4D scenes from scratch or from partial observations, controlled through multimodal conditions.
- 2) **Predictive World Models:** These models predict the future development of a scene based on past observations (usually under action conditioning).

Formally, a WM is described as a learned approximation of environmental dynamics, typically embedded in a

Markov Decision Process (MDP) [30]. An MDP is the tuple (S, A, T, R, γ) : S denotes the state space, A the action space, $T : S \times A \rightarrow \Delta(S)$ the stochastic transition density, $R : S \times A \rightarrow \mathbb{R}$ the reward function, and $\gamma \in [0, 1)$ the discount factor [31]. Technically, the WM approximates the true dynamics $T(s_{t+1}|s_t, a_t)$ and reward $R(s_t, a_t)$ through learnable functions $T_\theta(s_{t+1}|s_t, a_t)$ and $R_\theta(s_t, a_t)$, optimized using collected experience tuples (s_t, a_t, r_t, s_{t+1}) via maximum likelihood estimation or related objective functions [32].

Early influential work was Dyna-Q, which combined real Reinforcement Learning (RL) with simulated rollouts in the model, thereby significantly increasing data efficiency [33]. Later works such as PETS [34], MBPO [35], and MuZero [36] refined this idea through ensemble estimations, model-based policy optimization, or embedded Monte Carlo planners. Common to all is that their learned models operate on compact state representations and do not aim to reconstruct visual raw data like pixels.

With the transition to real-world robotics and visual tasks, it became apparent that low-dimensional state representation no longer suffices: the model must “see” the environment directly from high-dimensional sensor data such as camera images. From this insight emerged the VWM class [37].

C. Video World Models: Classification and Demarcation

VWMs extend the classical paradigm of world modeling. As elaborated above, WMs can fundamentally pursue two goals: understanding the world through abstraction or generating and simulating its details. VWMs are a specialized class primarily dedicated to the second goal, aiming to simulate the world through video prediction methods [29].

1) Functional Classification

The functional categorization of video-based world models according to Kong et al. (2025) distinguishes three main categories differing in their task definition and interaction with the environment: *Data Engines*, *Action Interpreters*, and *Neural Simulators* [4].

a) Data Engines: These models focus on synthesizing diverse and controllable scenes from geometric and semantic conditions [4]. They function as the foundation for scalable data augmentation by generating high-resolution video sequences without being explicitly dependent on historical observations. Systems like Sora utilize DiT to extract representations from millions of hours of video material [5]. However, Data Engines exhibit a trade-off between generation quality and representation quality [38], and large datasets compiled from the internet can contain systematic biases [39].

b) Action Interpreters: These models extend the Data Engine approach through action-conditioned predictions of future states [4]. They translate action inputs such as trajectories or velocities into latent representations through mechanisms like VQ-VAE-based quantization [40] or learned adaptation layers [41]. Unlike Data Engines, they integrate dynamic action information for targeted state prediction.

c) Neural Simulators: These models implement closed-loop simulations by combining low-level motion control with high-level conditioning [4]. They utilize latent dynamics models such as Mixture Density Network-RNNs [28] or Neural Ordinary Differential Equations [42] to replicate long-term environmental interactions.

This tripartition illustrates that video-based WMs should not be interpreted as homogeneous systems but as specialized functional components addressing different aspects of world modeling. **This paper focuses exclusively on VWMs in their function as Data Engines.**

2) Technological Classification

Technologically, the current research landscape of VWMs can be divided into three main directions fundamentally differing in their world representation.

a) Latent VWMs: This class traces back to the foundational work of Ha and Schmidhuber (2018). They proposed a three-part architecture: an encoder $q_\phi(z_t|o_t)$ mapping observations o_t to latent states z_t , a transition model $p_\theta(z_{t+1}|z_t, a_t)$ predicting dynamics in latent space, and a decoder $p_\phi(o_t|z_t)$ reconstructing original images during training [43]. A prominent early example is PlaNet, using a Recurrent State-Space Model (RSSM) and optimizing the Evidence Lower Bound (ELBO) [44]. Dreamer demonstrated shortly afterward that the policy can be trained directly in latent space through imagined trajectories, enabling high data efficiency [45]. This line was further developed to DreamerV3, achieving SOTA results in diverse benchmarks [37]. The central advantage lies in computational efficiency achieved through strong compression of the world into a compact state [45].

b) Pixel-Space VWMs: Models like Cosmos-Predict from NVIDIA, Genie 3 from Google DeepMind, or OpenAI Sora, forgo strong compression and instead predict future frames in full resolution [46]. Cosmos-Predict illustrates this approach by first tokenizing videos into a compact but image-faithful latent grid [47]. A large diffusion model then operates on this representation to generate future frames [46]. Unlike earlier approaches targeting high semantic compression, the topological grid structure of the video is preserved here. This kind of latent space thus functions more as “advanced pixels,” enabling the generation of visually plausible data streams [48].

c) 3D/4D VWMs: This class constitutes a third research direction modeling world states directly in native 3D or 4D representations [49]. These methods typically build on explicit 3D representations such as Neural Radiance Fields (NeRFs) or Gaussian Splatting [50] and promise higher geometric consistency, though they remain at an early stage regarding scalability and controllability [51].

3) Demarcation from Video Generation Models

It is important to distinguish VWMs from pure text-to-video generators such as Imagen Video [52] or Stable Video Diffusion [53], as the term VWM is often used imprecisely in

literature [29]. Even Google and OpenAI present their models as WMs or at least “world simulators” and “physically accurate” [54]. However, there is a clearly demonstrable technical boundary between a mere VGM and a VWM [55].

Two works in particular define this boundary based on central capabilities extending beyond mere generation. Both Ding et al. (2025) [12] and Huang (2025) [29] emphasize the importance of **interactivity**, closely linked to controllability, as a decisive capability. A VWM must be able to simulate the consequences of external interventions and generate corresponding feedback, rather than merely processing a given input. Equally central for VWMs is **long-term coherence**. Unlike short, often inconsistent clips, they must be capable of making long-term predictions that remain faithful to the dynamic rules of the simulated environment over an extended period.

Huang (2025) proposes a more fundamental, technical demarcation. He argues that **causality** represents the first and essential step toward an authentic VWM. Causality, which requires adherence to an autoregressive, temporally correct sequence, is fundamental since bidirectional attention (as in many VGMs) excludes genuine interaction by definition.

D. Central Challenges in Evaluating Data Engines

The evaluation of VWMs in their role as Data Engines confronts central challenges arising directly from the specific functional requirements. VWMs must not only synthesize visually plausible scenes but ideally follow physical principles, maintain long-term spatiotemporal consistency, and produce causally correct predictions [12].

1) Visual Quality and Long-term Coherence

Research traditionally relies on established metrics from image and video processing, whose suitability for VWM requirements is increasingly questioned. Peak Signal-to-Noise Ratio (PSNR), while mathematically simple, correlates inadequately with human perception as it fails to capture structural distortions [56]. Structural Similarity Index Measure (SSIM) correlates better empirically but operates primarily at frame-level and remains insensitive to temporal inconsistencies, a core problem of generated videos [57]. Even perceptual metrics like Learned Perceptual Image Patch Similarity (LPIPS), utilizing deep learning features, evaluate videos essentially as sequences of isolated images, missing spatiotemporal dependencies [9].

2) Controllability and Domain-Specific Adaptation

Evaluation of controllability, often via text prompts, reveals that standard approaches like Contrastive Language-Image Pre-training (CLIP) Scores neglect the temporal dimension [58]. In domain-specific adaptation, particularly in complex areas like autonomous driving, studies reveal fundamental bias in current models. Research demonstrates that 80% of training videos exhibit low dynamics, causing models to systematically fail in highly dynamic scenarios [59].

3) Measuring Physical Understanding

This addresses whether VWMs actually learn internal world models or merely act as “sophisticated pixel predictors.” Specialized benchmarks empirically confirm the discrepancy between visual excellence and physical correctness, with top models achieving only 10–30% on physics understanding assessments [7].

III. METHODOLOGY

The methodological implementation of the research objectives follows a qualitative, concept-centric Systematic Literature Review (SLR) strictly adhering to the established procedure of Webster and Watson (2002) [60]. This approach provides the necessary structured framework not only to identify relevant literature but to synthesize it in a concept-centric manner.

A. Search Strategy and Selection

A multi-stage search strategy was conducted across databases including arXiv, Scopus, Google Scholar, and Springer Link. Keywords combined terms such as “World Model,” “Video World Model,” “Video Prediction,” “Evaluation,” “Metrics,” “Benchmark,” “Physics,” “Causality,” and “Data Engine.”

- **Timeframe:** 2018–2025, capturing the rise and evolution of VWMs since Ha and Schmidhuber’s foundational work on WMs.
- **Inclusion Criteria:** Papers presenting VWM architectures, evaluation methods, benchmarks, or systematic analyses of video generation quality.
- **Exclusion Criteria:** Papers focusing exclusively on image generation, pure RL without video components, or lacking evaluation components.
- **Selection Process:** After screening 369 initial results based on title, abstract, and full-text review, **69 papers** were selected for detailed analysis.

B. Inductive Synthesis Process

Evaluation dimensions were not predefined but derived inductively from the literature following a grounded theory approach. For each paper, evaluation approaches, metrics, benchmarks, and explicitly stated limitations were coded. Through iterative clustering and abstraction, 61 granular concepts were identified and consolidated into 9 overarching evaluation dimensions. These dimensions were then mapped to a hierarchical four-level framework based on their functional dependencies and conceptual relationships.

A concept matrix was constructed to track which papers addressed which dimensions, enabling frequency analysis and identification of research gaps. The synthesis followed the principle of moving from author-centric to concept-centric organization, ensuring that the framework captures the state of the field rather than individual perspectives.

IV. ANALYSIS OF EVALUATION DIMENSIONS

The SLR revealed nine distinct dimensions currently employed in VWM research, ranging from basic visual quality to complex downstream task utility. The frequency analysis shows that Visual Quality dominates (52/69 papers), followed by Temporal Consistency (47/69) and Physical Plausibility (39/69), while Generalizability (18/69) and Multi-View Consistency (21/69) remain underrepresented. Figure 1 visualizes this distribution, revealing a clear concentration on lower-level evaluation dimensions.

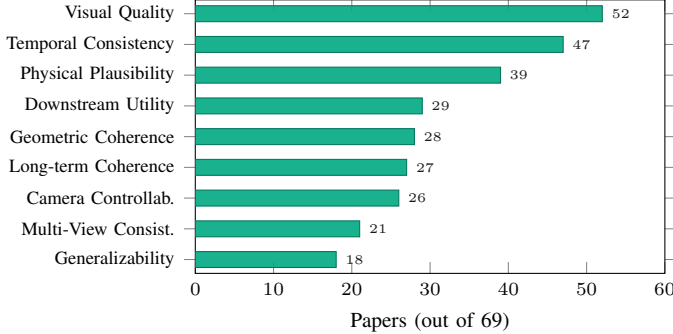


Fig. 1. Frequency of evaluation dimensions across the 69 reviewed papers. Lower-level dimensions (Visual Quality, Temporal Consistency) dominate, while higher-level dimensions (Generalizability, Multi-View Consistency) remain underrepresented.

A. Perceptual Quality

Visual Quality remains the most frequently evaluated dimension. The field has traditionally been dominated by distributional metrics: Fréchet Video Distance (FVD) and Fréchet Inception Distance (FID), both of which are distributional divergence measures. FVD measures the distance between feature distributions of real and generated videos using a spatiotemporal I3D network, while FID operates on individual frames using Inception features. However, these aggregate metrics are increasingly superseded by decomposed, model-based alternatives that allow finer diagnostic capability.

a) SOTA Metrics: The review identifies a decisive shift toward AI Judges and decomposed metrics. VBench Aesthetic Quality, LAION Aesthetic Predictor, DINOv2 L2-Distance, and MUSIQ for low-level artifacts represent the current state-of-the-art.

b) Central Limitations:

- FVD ($r = 0.67$) and FID ($r = 0.34$) correlate significantly more weakly with human perception than FVMD ($r = 0.85$).
- PSNR is insufficient for perceptual quality as it is purely pixel-based.
- SSIM becomes unreliable for long videos due to inevitable distributional shift.
- Standard metrics conflate visual quality with temporal consistency, making root-cause diagnosis impossible.

B. Spatiotemporal Coherence

This level encompasses four dimensions ensuring consistency across time, space, and viewpoints.

1) Temporal Consistency

This dimension measures motion smoothness, temporal flickering, and frame-to-frame coherence. Unlike aggregate scores, recent decomposed metrics enable distinguishing between flickering artifacts and motion quality degradation.

a) SOTA Metrics: VBench decomposes this into “Temporal Flickering” and “Motion Smoothness” sub-scores. FVMD specifically targets motion dynamics. CLIP feature similarity and frame interpolation reconstruction error provide additional perspectives.

b) Central Limitations:

- FVD conflates error accumulation and long-term consistency, precluding causal analysis.
- Distributional shift influences metric values more strongly than actual quality.
- Automated metrics favor short-term smoothness over long-term coherence.

2) Geometric Coherence

This dimension evaluates 3D structural integrity and semantic consistency of objects and scenes. While distributional metrics operate at the video level, geometric evaluation requires object-level identity tracking and 3D-aware representations.

a) SOTA Metrics: VBench-2.0 Subject Consistency (DINO-based), ArcFace for identity consistency, DINOv2 Scene Consistency, and Chamfer Distance for 3D structures.

b) Central Limitations:

- FVD/FID do not capture semantic identity drift of objects.
- Pixel-based metrics (PSNR/SSIM) are insensitive to geometric object deformations.
- Missing 3D ground truth complicates precise geometric validation.

3) Multi-View Consistency

Crucial for robotics applications, this dimension validates geometric relationships between generated views. These metrics differ fundamentally from temporal ones by operating in projective geometry rather than pixel space.

a) SOTA Metrics: VBench-2.0 Feature Matching Stability using Scale-Invariant Feature Transform (SIFT), Fast Library for Approximate Nearest Neighbors (FLANN), and Random Sample Consensus (RANSAC); Sampson Error (TSE, CSE), KPM Score, and Cross-View Attention mechanisms.

b) Central Limitations:

- Metrics are constrained to pairwise/local views; global multi-view consistency remains difficult to quantify.
- Stereo reconstruction via RANSAC is computationally intensive.

4) Long-term Coherence and Memory

This dimension addresses the “hallucination” problem in long videos, specifically assessing whether models can maintain consistent world states over extended periods. It represents the most challenging temporal evaluation, as errors compound over time.

a) SOTA Metrics: VBench Background/Subject Consistency, Scene Revisit Consistency (SRC), and View Recall protocols.

b) Central Limitations:

- FVD becomes problematic for long videos (>200 frames) due to inevitable distributional shift.
- Automated metrics fail to capture subtle temporal drift (e.g., object identity, scene logic).

C. Plausibility and Controllability

This level distinguishes a graphics engine from a physics engine.

1) Physical Plausibility

This represents the most critical capability gap in current VWMs. Unlike perceptual metrics, physics evaluation requires assessing causal correctness rather than statistical similarity.

a) SOTA Metrics: WorldModelBench Physics Adherence Score, Velocity Error, DreamSim for Physics Alignment, and Abnormal Ratio through human evaluation.

b) Central Limitations:

- Top models achieve only 10–30% Physics Alignment Score on WorldModelBench.
- Visual realism does not correlate with physical correctness.
- Scaling alone is insufficient for OOD generalization of physical rules.

Specialized benchmarks confirm this empirically. Physics-IQ reveals “severely limited understanding” even in SOTA models (Sora: 10.0%; VideoPoet: 29.5%). PhyGenBench corroborates this finding, with even the best models like Pika achieving only 32.4% on average across all physical phenomena tested.

2) Camera Controllability

This dimension measures adherence to trajectory prompts and geometric control specifications. It bridges physical plausibility and user intent, requiring precision without sacrificing visual quality.

a) SOTA Metrics: Rotation Error (RotError), Translation Error (TransError), Absolute Trajectory Error (ATE), Relative Pose Error (RPE), Geometric Alignment Error, and VBench-2.0 CoTracker-v2.

b) Central Limitations:

- Balance between control precision and generative flexibility proves difficult.
- Standardization for acceptable deviations is missing.

D. Transferability and Utility

The highest level evaluates the model’s practical value for downstream applications.

1) Generalizability

This dimension tests Out-of-Distribution (OOD) performance and compositional generalization. It reveals whether models learn transferable rules or merely memorize training patterns.

a) SOTA Metrics: Velocity Error for OOD/Combinatorial Generalization, Abnormal Ratio, and Control Latent Similarity for novel trajectories.

b) Central Limitations:

- Models prioritize case-based imitation over rule extraction.
- OOD Abnormal Ratio exceeds 60% in top models.

2) Utility for Downstream Tasks

The ultimate validation: training an agent on generated data and testing in the real world or trusted simulator. This dimension closes the loop from generation to application, revealing whether visual quality translates to functional value.

a) SOTA Metrics: nuScenes Detection Score (NDS) and mean Average Precision (mAP) for object detection, Success Rate/Collision Rate for planning, and mean Intersection over Union (mIoU) for Bird’s Eye View (BEV) segmentation.

b) Central Limitations:

- LPIPS ($r = 0.54$) correlates more weakly with task performance than semantic metrics like DINOv2.
- Pixel-based metrics fail to capture task-relevant dynamics.

V. SYNTHESIS: A HIERARCHICAL EVALUATION FRAMEWORK

We synthesize the findings into a four-level hierarchical framework with clear functional dependencies (Table I). The synthesis reveals that these levels form a dependency chain: Spatiotemporal coherence requires perceptual quality as its foundation; physical plausibility requires coherent spatiotemporal representations; and downstream utility depends on the reliability of the underlying physics simulation.

A. Grouping Dimensions into Conceptual Levels

The four levels represent increasing abstraction and complexity:

TABLE I
SYNTHESIZED EVALUATION FRAMEWORK FOR VIDEO WORLD MODELS AS DATA ENGINES

Dimension (RQ1)	SOTA Metrics (RQ2)	Central Limitations (RQ3)
1. Perceptual Quality		
Visual Quality	VBench Aesthetic Quality; LAION Aesthetic Predictor; DINOv2 L2 Distance; MUSIQ for low-level artifacts	FVD ($r=0.67$) and FID ($r=0.34$) correlate significantly weaker with human perception than FVMD ($r=0.85$) [9]; PSNR insufficient for perceptual quality as it is purely pixel-based [61], [72]; SSIM unreliable for long videos due to unavoidable distributional shift [62]; metrics conflate visual quality with temporal consistency [9]
2. Spatiotemporal Coherence		
Temporal Consistency	VBench: Temporal Flickering, Motion Smoothness; FVMD; CLIP Feature Similarity; Frame Interpolation Reconstruction Error	FVD conflates error accumulation and long-term consistency, no root cause separation possible [62]; distributional shift influences metric values more than actual quality [62]; automated metrics favor short-term smoothness over long-term coherence [12]; frame interpolation reconstruction error does not capture subtle temporal drift [63]
Geometry	VBench-2.0 Subject Consistency (DINO-based); ArcFace for identity consistency; DINOv2 Scene Consistency; Chamfer Distance for 3D structures	FVD/FID do not capture semantic identity drift of objects [11], [62]; pixel-based metrics (PSNR/SSIM) insensitive to geometric object deformations [4]; missing 3D ground truth hampers precise geometric validation [64]; DINO-based Subject Consistency limited to visually obvious inconsistencies [10]
Multi-View Consistency	VBench-2.0 Feature Matching Stability (SIFT/FLANN/RANSAC); Sampson Error (TSE, CSE); KPM Score; Cross-View Attention	Metrics (SIFT/FLANN, Sampson Error) limited to pairwise/local views; global multi-view consistency difficult to quantify [65], [66]; stereo reconstruction via RANSAC computationally intensive [47]; visual ambiguity in 3D inference from 2D projections [64]; Sampson Error sensitive to noise and outliers [61]
Long-Term Coherence	VBench: Background/Subject Consistency, Motion Smoothness; Scene Revisit Consistency (SRC, LPIPS-based); View Recall with Average Human Ranking	FVD problematic for long videos (>200 frames) due to unavoidable distributional shift [62]; automated metrics do not capture subtle temporal drift (e.g., object identity, scene logic) [12]; lack of standardization for long-term horizons [6], [62]; human evaluation resource-intensive and hard to scale [67]; spatial memory gaps during abrupt camera movements [67]
3. Plausibility & Controllability		
Physical Plausibility	WorldModelBench Physics Adherence Score; Velocity Error; DreamSim for Physics Alignment; Abnormal Ratio (Human Evaluation)	Top models achieve only 10–30% Physics Alignment Score on WorldModelBench [7]; visual realism does not correlate with physical correctness [7]; scaling alone insufficient for OOD generalization of physical rules [8]; 2D video projection complicates 3D physics extraction [64]; lack of formalized physics benchmarks for cross-domain evaluation [9], [11]
Camera Controllability	RotError/TransError; ATE / RPE (SLAM-based); Geometric Alignment Error; VBench-2.0 CoTracker-v2 for Motion Consistency	Difficult balance between control precision and generative flexibility: strict pose adherence can reduce visual plausibility [46], [68]; lack of standardization for acceptable deviations (thresholds for RotError/TransError) [47]; evaluation of complex trajectories (e.g., orbital, backward movements) limited [64]; Simultaneous Localization and Mapping (SLAM)-based metrics (ATE/RPE) prone to tracking errors [67]
4. Transferability & Utility		
Generalizability	Velocity Error for OOD/Combinatorial Generalization; Abnormal Ratio; Control Latent Similarity for novel trajectories; In-Context Learning Capability	FVD/LPIPS conflate error accumulation and generalization ability, no root cause analysis possible [62]; models prioritize case-based imitation over rule extraction [8]; lack of quantitative metrics for cross-domain generalization [41]; OOD Abnormal Ratio $> 60\%$ for top models [8]; in-context learning capability difficult to quantify [62]
Utility for Downstream Tasks	NDS/mAP for object detection; Success Rate/Collision Rate for planning; mIoU for BEV segmentation; DINOv2 CLIP Correlation	LPIPS ($r=0.54$) correlates weaker with task performance than semantic metrics (DINOv2 L2, CLIP Score) [68]; pixel-based metrics do not capture task-relevant dynamics [68]; frame comparison metrics unsuitable for long-horizon planning [62]; lack of standardization for downstream evaluation (which tasks are representative?) [64], [69]; disconnect between visual quality and utility [9]

1) Level 1 – Perceptual Quality

The foundational layer addressing whether generated content is visually acceptable. This is necessary but insufficient for Data Engine functionality.

2) Level 2 – Spatiotemporal Coherence

Builds on visual quality to ensure consistency across temporal, geometric, multi-view, and long-term dimensions. A video may look good frame-by-frame but fail at this level if objects morph or scenes become inconsistent.

3) Level 3 – Plausibility & Controllability

The critical differentiator between VGMs and VWMs. This level assesses whether the generated world obeys physical laws and responds appropriately to control inputs. Current models achieve high Level 1 scores but fail dramatically here.

4) Level 4 – Transferability & Utility

The ultimate validation measuring whether synthetic data actually improves downstream task performance. This closes

the loop from generation to application.

B. Major Findings from the Synthesis

1) Finding 1: Metric Conflation Problem

A major finding is the limitation of standard metrics. FVD mixes visual quality and temporal consistency into a single number, obscuring the root cause of errors. When a model receives a poor FVD score, it is impossible to determine whether the issue lies in visual artifacts, temporal inconsistency, or physical implausibility.

2) Finding 2: Paradigm Shift to Model-Based Evaluation

The trend is decisively moving toward **model-based evaluation**, where large pre-trained models act as “AI Judges” to quantify complex concepts like causality, object permanence, or physical law adherence that pixel-metrics fundamentally cannot capture. VBench, VBench-2.0, and WorldModelBench exemplify this shift.

3) *Finding 3: The Physics Gap*

Current VWMs excel at Level 1 (visual quality) but struggle significantly at Level 3 (physical plausibility). The “Uncanny Valley” of simulation is no longer visual but dynamic: videos look real, but objects do not consistently obey gravity or conservation of momentum.

4) *Finding 4: Case-Based vs. Rule-Based Generalization*

Models exhibit case-based generalization, imitating training examples rather than extracting abstract physical rules. This suggests that simply scaling data and parameters may not suffice to achieve genuine physical understanding; architectural priors or neuro-symbolic approaches may be required.

5) *Finding 5: Evaluation Fragmentation*

Despite 69 papers addressing VWM evaluation, no prior work has synthesized these approaches into a coherent framework. Individual benchmarks address specific sub-problems but remain “island solutions” disconnected from a holistic evaluation strategy.

VI. DISCUSSION

The synthesis from Section V yielded five central findings that collectively motivate a fundamental reorientation of VWM evaluation practice: the metric conflation problem (Finding 1), the paradigm shift to model-based evaluation (Finding 2), the persistent physics gap (Finding 3), the dominance of case-based over rule-based generalization (Finding 4), and the fragmentation of evaluation approaches (Finding 5). The following implications for research and practice are derived directly from these findings.

A. *Implications for Research*

The analysis highlights that for VWMs to succeed as Data Engines, the evaluation culture must fundamentally shift from optimizing aggregate scores like FVD toward rigorous, multi-dimensional testing of physical alignment and OOD generalization. The current practice of reporting a single FVD number obscures critical failure modes that could prove catastrophic when synthetic data is used to train autonomous systems.

Researchers should adopt the following principles:

- 1) **Adopt decomposed metrics** that explicitly separate visual quality, temporal consistency, and physical plausibility rather than relying on aggregate scores. When a model fails, practitioners need to know *why*—is it generating visual artifacts, temporal flickering, or physically impossible scenarios? The framework presented here enables such diagnostic capability.
- 2) **Incorporate AI Judges** for aspects that pixel-metrics fundamentally cannot capture, while remaining cognizant of their own biases and limitations. Model-based evaluation using Vision-Language Models (VLMs) or specialized physics classifiers can assess causality, object permanence, and law adherence at a semantic level inaccessible to pixel comparisons.

- 3) **Design explicitly for long-term coherence** by including evaluation protocols that test memory and scene revisit consistency. The “View Recall” paradigm, where a camera returns to a previously observed location, provides a powerful probe of whether models maintain genuine world state or merely generate plausible hallucinations.
- 4) **Validate through downstream tasks** as the ultimate proof of Data Engine quality. Improvement in agent performance when trained on synthetic data (e.g., mAP for detection, success rate for planning) closes the loop from generation to application value.
- 5) **Report OOD generalization** alongside standard benchmarks. The finding that models exhibit >60% abnormal ratio on out-of-distribution tests suggests that current evaluation practices are systematically overestimating true capability by testing within narrow training distributions.

B. *Implications for Practice*

A central objective of this work is facilitating knowledge transfer from research to industrial practice. For practitioners deploying VWMs in autonomous systems, this framework provides concrete guidance for responsible evaluation and deployment:

- 1) **Visual quality alone is insufficient and potentially misleading.** A model achieving excellent FVD scores may still generate physically impossible scenarios: objects defying gravity, vehicles passing through walls, or liquids behaving like solids. Training autonomous agents on such data could instill dangerous behaviors that only manifest in edge cases.
- 2) **Domain-specific validation is absolutely essential.** The DEVIL benchmark demonstrated dramatically that models performing exceptionally well on static scenes (93.4% quality) completely fail in dynamic scenarios (0.0%) [72]. Practitioners must ensure their evaluation scenarios match the dynamism and complexity of actual deployment environments.
- 3) **Test OOD generalization rigorously before deployment.** Current models exhibit >60% abnormal ratio on out-of-distribution tests [8], indicating they are primarily interpolating training data rather than learning transferable physical principles. Any novel scenario in deployment could trigger unpredictable failures.
- 4) **Establish clear acceptance criteria across all four framework levels.** Before using synthetic data for agent training, define minimum thresholds for visual quality, spatiotemporal coherence, physical plausibility, AND downstream utility. Passing only Level 1 evaluation is insufficient for safety-critical applications.
- 5) **Implement continuous monitoring.** As VWM architectures evolve rapidly, evaluation pipelines must be adaptable. The framework’s hierarchical structure enables incremental assessment as new capabilities emerge.

The gap between Level 1 performance and Level 3 performance documented in this study represents a fundamental risk for Physical AI applications. Practitioners should approach VWM-generated data with the same scrutiny applied to any simulation: verification against physical ground truth remains essential until evaluation methods mature sufficiently to provide reliable physics guarantees.

C. Limitations of This Work

The significance of the results is subject to specific methodological limitations that must be critically reflected upon. A central challenge of the methodology arose from the extreme dynamism of the research field. Since a significant portion of the relevant publications appeared as preprints within the last 12 months, this work builds upon a foundation that has often not yet fully undergone the traditional peer-review process. The inclusion of “grey literature” was essential to capture the actual State-of-the-Art (SOTA), but it carries the risk that individual benchmarks classified here might be replaced by more robust methods in the near future. Despite these limitations, the work successfully achieves a methodical balance between breadth and depth. On one hand, the research was designed comprehensively enough to identify fine-grained nuances, such as the various sub-dimensions of physical plausibility. On the other hand, this profound understanding enabled the consolidation of scattered and fragmented approaches from current research literature into a unified and systematic order.

Furthermore, a methodological limitation lies in the interpretative classification of ambiguous terms within the primary literature, such as “coherence” or “consistency,” whose definitions often vary and require clear framing by the authors. Additionally, for the sake of generalizability, a detailed elaboration of test protocols for individual niche domains was omitted; this gives the work the character of a structural roadmap but cannot cover specific practical requirements in ultimate detail. The compressed format of this conference paper necessitated condensing several aspects, particularly the detailed concept matrices and per-dimension metric breakdowns available in the underlying study. The concrete design for specialized use cases is left for future research.

D. Future Research Directions

Several critical gaps warrant future investigation, representing high-priority opportunities for advancing VWM evaluation:

- 1) **Standardized Long-term Benchmarks:** Rigorous protocols for evaluating coherence beyond 200 frames are urgently needed, with clear success criteria, standardized test scenarios, and reproducible evaluation pipelines. Current ad-hoc approaches make cross-study comparison impossible.
- 2) **Unbiased and Calibrated AI Judges:** Model-based evaluators may inherit systematic biases from their training data, potentially favoring certain visual styles or dynamics. Developing robust, calibrated judges with known

failure modes and confidence intervals is essential for trustworthy evaluation.

- 3) **Physics-Informed Architectures:** The finding that models exhibit case-based rather than rule-based generalization suggests that pure data scaling may be fundamentally insufficient. Future work should explore incorporating explicit physical priors, differentiable physics engines, or neuro-symbolic reasoning modules to enable genuine physical understanding.
- 4) **Cross-Domain Transfer Metrics:** Quantifying how well synthetic data from one domain (e.g., autonomous driving) transfers to another (e.g., indoor robotics) remains an open challenge. Such metrics would enable more efficient resource allocation and identify fundamental vs. domain-specific capabilities.
- 5) **Causal Intervention Benchmarks:** Moving beyond observational evaluation to test whether models correctly predict the consequences of hypothetical interventions would provide stronger evidence of genuine world modeling capability.
- 6) **Real-World Validation Studies:** Ultimately, the value of VWMs as Data Engines must be validated through comprehensive studies comparing agent performance when trained on purely real, purely synthetic, and optimally mixed combinations of data across diverse tasks and conditions.

VII. CONCLUSION

This work systematically structured the fragmented field of VWM evaluation through a Systematic Literature Review of 69 state-of-the-art publications. The analysis identified nine evaluation dimensions organized into a four-level hierarchical framework: Perceptual Quality, Spatiotemporal Coherence, Plausibility & Controllability, and Transferability & Utility.

The central finding is that while visual fidelity has reached impressive levels, physical plausibility remains the critical bottleneck for deploying VWMs as Data Engines in Physical AI applications. Current models can generate scenes that appear visually authentic yet behave unrealistically. Consequently, the primary challenge has shifted from achieving static visual quality to ensuring dynamic and physical correctness.

The proposed framework provides researchers and practitioners with a roadmap to move beyond superficial metrics like FVD and adopt a holistic evaluation strategy incorporating geometric consistency, physical law adherence, long-term memory, and downstream task utility. The documented shift toward model-based evaluation (AI Judges) reflects the field’s recognition that pixel-metrics are fundamentally insufficient for assessing whether a VWM has learned to simulate or merely to imitate.

For VWMs to fulfill their promise as engines of synthetic data for autonomous systems, the evaluation paradigm must evolve from measuring “does it look real?” to “does it behave correctly?”. This framework aims to guide this essential transformation.

ACKNOWLEDGMENT

The authors acknowledge the use of AI-assisted tools during the preparation of this work. Elicit and Open Knowledge Maps were used to support the literature search, identify relevant prior work, and verify forward and backward search results during the SLR. Google Gemini was used for language refinement, including improvements to grammar, spelling, and phrasing, as well as for rapid content analysis of PDF sources to locate relevant passages efficiently. DeepL Write was used for linguistic polishing of author-written text. A ChatLLM was used to scan research literature for additional relevant sources and to critically review the accuracy of cited references. All AI-assisted outputs were reviewed, validated, and edited by the authors; the research design, literature analysis, synthesis, interpretation of results, and final wording remain the sole responsibility of the authors.

REFERENCES

- [1] J. Bruce et al., “Genie: Generative Interactive Environments,” in *Proc. of the 41st International Conference on Machine Learning (ICML)*, PMLR 235:4603–4623, 2024.
- [2] A. Melnik et al., “Video Diffusion Models: A Survey,” *arXiv preprint*, 2024.
- [3] S. Ali, P. Arcaini, and A. Arrieta, “Foundation Models for the Digital Twin Creation of Cyber-Physical Systems,” in *Proc. of the International Symposium on Leveraging Applications of Formal Methods (ISoLA)*, 2024.
- [4] L. Kong et al., “3D and 4D World Modeling: A Survey,” *arXiv preprint*, 2025.
- [5] T. Brooks et al., “Video generation models as world simulators,” OpenAI Research, 2024.
- [6] Z. Zhu et al., “Is Sora a World Simulator? A Comprehensive Survey on General World Models and Beyond,” *arXiv preprint*, 2024.
- [7] S. Motamed et al., “Do generative video models understand physical principles?,” in *Proc. of the International Conference on Computer Vision (ICCV)*, 2025.
- [8] B. Kang et al., “How Far is Video Generation from World Model: A Physical Law Perspective,” in *Proc. of the 42nd International Conference on Machine Learning (ICML)*, 2025.
- [9] H. Luo et al., “FVMD: Fréchet Video Motion Distance,” in *Proc. of the 41st International Conference on Machine Learning (ICML)*, 2024.
- [10] F. Baldassarre et al., “Back to the Features: DINO as a Foundation for Video World Models,” in *ICML Workshop on Video Generation*, 2025.
- [11] D. Agarampalli, “Synthetic Data Pipelines for Training AI Models in Data-Scarce Domains,” *International Journal of Innovative Research in Science Engineering and Technology*, vol. 14, no. 04, 2024.
- [12] J. Ding et al., “Understanding World or Predicting Future? A Comprehensive Survey of World Models,” *ACM Comput. Surv.*, vol. 58, no. 3, pp. 1–38, 2025.
- [13] P. F. Zhang et al., “A Step Toward World Models: A Survey on Robotic Manipulation,” *arXiv preprint*, 2025.
- [14] N. Aldausari et al., “Video Generative Adversarial Networks: A Review,” *ACM Computing Surveys*, vol. 55, no. 2, 2022.
- [15] Team Wan et al., “Wan: Open and Advanced Large-Scale Video Generative Models,” *arXiv preprint*, 2025. [Accepted to NeurIPS 2025 as Wan-Move]
- [16] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.
- [17] C. Vondrick, H. Pirsiavash, and A. Torralba, “Generating Videos with Scene Dynamics,” *NeurIPS*, 2016.
- [18] S. Tulyakov, “Three and a Half Generations of Video Generation Models,” in *Proc. ACM International Conference on Multimedia Retrieval*, 2025.
- [19] G. Bertasius, H. Wang, L. Heng, and L. Torresani, “Is Space-Time Attention All You Need for Video Understanding?,” *ICML*, 2021.
- [20] A. Arnab et al., “ViViT: A Video Vision Transformer,” *ICCV*, 2021.
- [21] Y. Wang et al., “Survey of Video Diffusion Models: Foundations, Implementations, and Applications,” *arXiv preprint*, 2025.
- [22] J. Ho, A. Jain, and P. Abbeel, “Denoising Diffusion Probabilistic Models,” *NeurIPS*, 2020.
- [23] J. Song, C. Meng, and S. Ermon, “Denoising Diffusion Implicit Models,” in *Proc. ICLR*, 2020.
- [24] Y. Yin et al., “Towards Precise Scaling Laws for Video Diffusion Transformers,” in *Proc. CVPR*, 2025.
- [25] B. Camburn, “Universal Physics Simulation: A Foundational Diffusion Approach,” *arXiv preprint*, 2025.
- [26] M. Q. Ali et al., “Humanoid World Models: Open World Foundation Models for Humanoid Robotics,” *arXiv preprint*, 2025.
- [27] R. S. Sutton, “Dyna, an Integrated Architecture for Learning,” *SIGART Bulletin*, 1991.
- [28] D. Ha and J. Schmidhuber, “World Models,” *arXiv preprint*, 2018.
- [29] X. Huang, “Towards Video World Models,” 2025, https://www.xunhuang.me/blogs/world_model.html.
- [30] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, MIT Press, 2020.
- [31] C. Dong and Y. Wang, “Towards Generalized Inverse Reinforcement Learning,” *arXiv preprint*, 2024.
- [32] Y. Li and P. Wang, “GPTMORE: Generative Pre-trained Transformer for Model-based Offline Reinforcement Learning,” in *Proc. Asia Conference on Cognitive Engineering and Intelligent Interaction (CEII)*, 2023.
- [33] R. Faulkner and D. Precup, “Dyna Planning using a Feature Based Generative Model,” in *Proc. of the Advances in Neural Information Processing Systems (NIPS) Workshop on Deep Learning and Unsupervised Feature Learning*, 2010.
- [34] K. Chua et al., “Deep Reinforcement Learning in a Handful of Trials using Probabilistic Dynamics Models,” *NeurIPS*, 2018.
- [35] M. Janner et al., “When to Trust Your Model: Model-Based Policy Optimization,” *NeurIPS*, 2019.
- [36] J. Schrittwieser et al., “Mastering Atari, Go, Chess and Shogi,” *Nature*, 2020.
- [37] D. Hafner et al., “Mastering diverse control tasks through world models,” *Nature*, vol. 640, no. 8059, pp. 647–653, 2025.
- [38] C.-P. Chang et al., “Seeing Clearly, Forgetting Deeply: Revisiting Fine-Tuned Video Generators for Driving Simulation,” *arXiv preprint*, 2025.
- [39] M. Assran et al., “V-JEPA 2: Self-Supervised Video Models Enable Understanding, Prediction and Planning,” *arXiv preprint*, 2025.
- [40] S. Ye et al., “Latent Action Pretraining from Videos,” in *Proc. of the International Conference on Learning Representations (ICLR)*, 2025.
- [41] M. Rigter et al., “AVID: Adapting Video Diffusion Models to World Models,” in *Proc. of the Reinforcement Learning Conference (RLC)*, 2025.
- [42] C. Hao et al., “Neural Motion Simulator: Pushing the Limit of World Models in Reinforcement Learning,” in *Proc. CVPR*, 2025.
- [43] D. Hafner et al., “Mastering Atari with Discrete World Models,” *arXiv preprint*, 2022.
- [44] D. Hafner, T. Lillicrap, I. Fischer et al., “Learning Latent Dynamics for Planning from Pixels,” *ICML*, 2019.
- [45] D. Hafner, T. Lillicrap, J. Ba et al., “Dream to Control,” *ICLR*, 2020.
- [46] M.-Y. Liu et al., “World Simulation with Video Foundation Models for Physical AI,” 2025, https://research.nvidia.com/publication/2025-09_world-simulation-video-foundation-models-physical-ai.
- [47] NVIDIA et al., “Cosmos World Foundation Model Platform for Physical AI,” *arXiv preprint*, 2025.
- [48] S. Dieleman, “Generative modelling in latent space,” 2025, <https://sander.ai/2025/04/15/latents.html>.
- [49] K. Gao et al., “NeRF: Neural Radiance Field in 3D Vision: A Comprehensive Review (Updated Post-Gaussian Splatting),” *arXiv preprint*, 2025.
- [50] P. J. Ball et al., “Genie 3: A New Frontier for World Models,” 2025.
- [51] Q. Miao et al., “Advances in 4D Generation: A Survey,” *arXiv preprint*, 2025.
- [52] J. Ho, W. Chan et al., “Imagen Video,” *arXiv preprint*, 2022.
- [53] A. Blattmann et al., “Stable Video Diffusion,” *arXiv preprint*, 2023.
- [54] T. Wiedemer et al., “Video models are zero-shot learners and reasoners,” *arXiv preprint*, 2025.
- [55] R. Ming et al., “A Survey on Video Prediction: From Deterministic to Generative Approaches,” *arXiv preprint*, 2024.
- [56] A. Horé and D. Ziou, “Image Quality Metrics: PSNR vs. SSIM,” *ICPR*, 2010.
- [57] MSU Graphics and Media Lab Video Group, “Ways of cheating on popular objective metrics,” <https://videoprocessing.ai/metrics/ways-of-cheating-on-popular-objective-metrics.html>.

- [58] K. Guan et al., “ETVA: Evaluation of Text-to-Video Alignment via Fine-grained Question Generation and Answering,” *arXiv preprint*, 2025.
- [59] M. Liao et al., “Evaluation of Text-to-Video Generation Models: A Dynamics Perspective,” in *Proc. of the Advances in Neural Information Processing Systems (NeurIPS 37)*, 2024.
- [60] J. Webster and R. T. Watson, “Analyzing the Past to Prepare for the Future: Writing a Literature Review,” *MIS Quarterly*, vol. 26, no. 2, pp. xiii–xxiii, 2002.
- [61] X. Li et al., “DriVerse: Navigation World Model for Driving Simulation via Multimodal Trajectory Prompting and Motion Alignment,” in *Proc. of the ACM International Conference on Multimedia (ACM MM)*, 2025.
- [62] T. Chen et al., “Learning World Models for Interactive Video Generation,” in *Proc. of the Advances in Neural Information Processing Systems (NeurIPS)*, 2025.
- [63] Y. Zhang et al., “Matrix-Game: Interactive World Foundation Model,” *arXiv preprint*, 2025.
- [64] Aether Team et al., “Aether: Geometric-Aware Unified World Modeling,” in *Proc. of the International Conference on Computer Vision (ICCV)*, 2025.
- [65] D. Zheng et al., “VBench-2.0: Advancing Video Generation Benchmark Suite for Intrinsic Faithfulness,” *arXiv preprint*, 2025.
- [66] X. Wang et al., “DriveDreamer: Towards Real-World-Drive World Models for Autonomous Driving,” in *Proc. ECCV*, pp. 55–72, 2024.
- [67] T. Wu et al., “Video World Models with Long-term Spatial Memory,” in *Proc. of the Advances in Neural Information Processing Systems (NeurIPS)*, 2025.
- [68] Y. Shang et al., “RoboScape: Physics-informed Embodied World Model,” in *Proc. of the Advances in Neural Information Processing Systems (NeurIPS)*, 2025.
- [69] W. Wu et al., “DriveScape: High-Resolution Driving Video Generation by Multi-View Feature Fusion,” in *Proc. CVPR*, 2025.
- [70] R. Zhang et al., “The Unreasonable Effectiveness of Deep Features as a Perceptual Metric,” *CVPR*, 2018.
- [71] F. Meng et al., “Towards World Simulator: Crafting Physical Commonsense-Based Benchmark for Video Generation,” in *Proc. of the International Conference on Machine Learning (ICML)*, 2025.
- [72] W. Liao et al., “DEVIL: Dynamic Video Evaluation Benchmark,” in *Proc. of the Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Exploring Research Supervisors' Perceptions of an Artificial Intelligence-Driven Virtual Reality Avatar Prototype: Insights from the REACH-ETHICS Project

Marco Montalto
Malta College of Arts, Science and
Technology
Paola, Malta
marco.montalto@mcast.edu.mt

Lorna Bonnici West
Malta College of Arts, Science and
Technology
Paola, Malta
lorna.bonnici.west@mcast.edu.mt

Daren Scerri
Malta College of Arts, Science and
Technology
Paola, Malta
daren.scerri@mcast.edu.mt

Alan Gatt
Malta College of Arts, Science and
Technology
Paola, Malta
alan.gatt@mcast.edu.mt

David Deguara
Malta College of Arts, Science and
Technology
Paola, Malta
david.deguara@mcast.edu.mt

David Debono
Malta College of Arts, Science and
Technology
Paola, Malta
david.d.debono@mcast.edu.mt

Marisa Galea Vella
Malta College of Arts, Science and
Technology
Paola, Malta
marisa.vella@mcast.edu.mt

Lisa Theuma
Malta College of Arts, Science and
Technology
Paola, Malta
lisa.theuma@mcast.edu.mt

Francis Delicata
Malta College of Arts, Science and
Technology
Paola, Malta
francis.delicata@mcast.edu.mt

Shirley Ann Gauci
Malta College of Arts, Science and
Technology
Paola, Malta
shirley.ann.gauci@mcast.edu.mt

Abstract—This study explores research supervisors' perceptions of an artificial intelligence-driven virtual reality avatar developed within the REACH-ETHICS project to support research ethics training. Using a qualitative focus group design supplemented by quantitative content evaluation, four supervisors from health studies and animal studies and one member from technical development interacted with the prototype by posing ethics-related prompts and rating the avatar's responses for relevance, correctness, completeness, and clarity. Descriptive analysis showed consistently high median scores across all evaluation dimensions, while thematic analysis revealed strong perceived relevance, factual accuracy, and pedagogical clarity. Participants highlighted the system's potential to enhance engagement and standardization in ethics supervision, while also noting the need for greater contextual depth and discipline-specific tailoring. Overall, findings suggest that artificial intelligence-driven virtual reality avatars can serve as valuable supplementary tools within multimodal research ethics training frameworks, particularly when supervisors are actively involved in their evaluation and iterative refinement. Importantly, the proposed system adopts a multimodal interaction design, allowing users to engage via both speech and text, while receiving primarily text-based responses optionally augmented with speech output. Although implemented within a virtual reality environment to enhance immersion and contextualized learning, the underlying conversational artificial intelligence framework is platform-independent and can also be accessed through non-immersive desktop interfaces. The use of virtual reality avatars is therefore intended not as a replacement for conventional chat-based systems,

but as an extension that enables more situated, dialogical, and experiential engagement with research ethics scenarios.

Keywords—research ethics, supervision, virtual reality, artificial intelligence, education technology

I. INTRODUCTION

Research ethics form a cornerstone of responsible research practice, especially in domains that directly affect humans and animals. In recent years, increasing concern has emerged around the adequacy and effectiveness of traditional methods of research ethics training. This has led to calls for more engaging, context-specific, and practical frameworks—especially in disciplines with high ethical stakes like healthcare and animal studies [1, 2, 3]. While student engagement has often been studied, the perspective of research supervisors, as key mediators and educators of ethical conduct, remains an underexplored but vital area for investigation.

The foundational role of research ethics is to safeguard the dignity, rights, and well-being of participants, whether human or animal [1, 3]. Over the past century, milestones such as the Nuremberg Code, Declaration of Helsinki, and European Code of Conduct for Research Integrity have solidified the importance of ethical rigor in scientific endeavors [1, 4]. However, despite longstanding educational efforts, several studies highlight a persistently low level of ethics awareness among researchers and postgraduate students, especially in low- to middle-income settings [5, 6, 7].

Project Research Ethics financed by Xjenza Malta, for and on behalf of the Foundation for Science and Technology, through the FUSION: R&I Research Excellence Programme under grant agreement number REP-2024-006.

Research supervisors are crucial in fostering ethical research behaviors among students. Supervisors are not only expected to ensure regulatory compliance but also serve as ethical role models. As [8] argues, teaching research ethics goes beyond disseminating rules—it entails modelling and guiding good science. Reference [9] reinforces this, finding that students feel burdened by overly bureaucratic approaches and prefer guidance that makes ethics practical and relatable.

Yet, evidence suggests supervisors themselves often lack structured support or training to fulfil this role effectively [10, 11, 12]. A systematic framework that assists supervisors in navigating complex ethical dilemmas and guiding students more confidently is therefore urgently needed.

Several studies have critiqued traditional didactic ethics education for being overly theoretical and insufficiently contextual [13, 14]. The use of collaborative case-based learning [15], online Moodle platforms [16], and even narrative-based learning [17] have been explored to make ethics training more practical and relevant. However, the integration of supervisors in the co-creation or evaluation of these tools remains minimal in the literature. The lack of supervisor-centric studies and feedback mechanisms means that program designers often miss vital insights from these frontline mentors.

There is growing recognition of the need to design research ethics training that incorporates supervisors' feedback. For example, the Path2Integrity and INTEGRITY EU-funded projects introduced dialogical methods for fostering ethical reasoning but largely focused on student outcomes and did not actively seek supervisors' insights [18]. Therefore, the current research (REACH-ETHICS project) aims to fill this gap by including focus groups with supervisors as part of its formal evaluation. This is in line with calls for more participatory ethics education, where both educators and learners shape the training experience [15].

Innovative technologies such as artificial intelligence (AI) and virtual reality (VR) have been suggested as tools for increasing engagement in challenging topics. While AI-driven VR for ethics training is novel, early applications show promise. For instance, [19] found that VR enhanced interactive learning, and ethical critical thinking and decision-making, in bioethics. The AI-driven VR avatar prototype which is being trained under the current REACH-ETHICS project represents an extension of this innovation into supervisor training support as well. Content to train the AI-driven VR avatar has been selected to ensure widespread coverage of the various contexts pertaining to research ethics in healthcare. The content includes documents from widely regarded organizations such as the World Health Organization, the United Nations, the European Commission, national regulations and legislation, the sustainable developmental goals, leading European guidelines and publications on healthcare and research ethics guidelines, and documents on rights across the life span considering varying vulnerabilities.

For supervisors, such a system can provide standardized ethical scenarios for training students, real-time demonstrations of ethical dilemmas, and a collaborative interface for guiding team discussions around ethics. By incorporating immersive tools and multimodal delivery, supervisors may find reduced burden in creating training content, more dynamic methods to engage students, and

enhanced confidence in facilitating ethically complex discussions.

Research ethics education has traditionally relied on lectures, seminars, and written guidelines to promote responsible research conduct. However, several studies have highlighted limitations in these approaches, particularly their tendency to present ethical principles in abstract or overly theoretical ways that may not adequately prepare learners for complex real-world dilemmas. Empirical work in responsible conduct of research (RCR) education indicates that interactive and contextualized learning methods are more effective in promoting ethical reasoning and awareness among researchers [20, 21]. Similarly, [22] argues that ethics education should move beyond rule-based instruction toward pedagogical approaches that foster reflective thinking and discussion about the underlying values of research integrity.

To address these limitations, researchers have increasingly explored digital tools and technology-enhanced learning environments for ethics training. Online platforms and case-based learning environments have been used to simulate ethical dilemmas and encourage collaborative discussion among learners. Large-scale initiatives such as Path2Integrity [18] and the VIRT2UE [11, 12, 23] programmes have also introduced dialogical teaching frameworks aimed at strengthening research integrity competencies among researchers and supervisors. While these approaches promote discussion and reflection, they are typically delivered through conventional online or classroom formats and rarely exploit immersive technologies.

Recent advances in conversational AI have further expanded opportunities for interactive learning. Conversational agents and educational chatbots have been applied across a range of educational contexts to provide personalized guidance, scaffold learning activities, and facilitate engagement [24, 25]. A systematic review [26] found that such systems can support inquiry-based learning and provide on-demand feedback to students. However, their effectiveness depends heavily on domain-specific knowledge grounding, and poorly contextualized responses may lead to overly generic or simplified explanations.

Parallel developments in immersive technologies have demonstrated the potential of VR as a powerful tool for experiential learning. VR environments allow learners to interact with simulated scenarios that mirror real-world situations, thereby promoting deeper engagement and reflective decision-making. A systematic review [27] found that immersive VR is particularly effective for training that requires contextual understanding and practical application of knowledge. Similarly, [28] highlights that immersive VR can increase learners' sense of presence and emotional performance, factors that affect learning outcomes. Studies exploring human behavior in immersive environments further show that VR can enhance users' awareness of ethical dilemmas and stimulate reflection [29, 30].

Despite these developments, the integration of conversational AI with immersive VR for research ethics education remains relatively unexplored. Existing frameworks typically employ either dialogue-based systems or immersive simulations independently, rather than combining the strengths of both approaches. While conversational AI systems are often implemented as text-based chatbots in educational settings [24, 25, 26], the present work adopts a

multimodal approach in which the same underlying AI agent can be accessed through different interaction modalities, including text, speech, and immersive VR interfaces. The inclusion of a VR-based avatar was motivated by the need to support contextualized and scenario-driven ethics training, where interaction can more closely resemble real supervisory dialogue. Unlike traditional chat interfaces, immersive environments enable a sense of presence and embodied interaction, which have been shown to enhance engagement and reflective learning, and contextual understanding in educational settings [27, 28, 29]. Importantly, the VR component should be understood as an additional interaction layer rather than a strict requirement of the system, as the prototype can also function in non-immersive desktop settings. The REACH-ETHICS project contributes to this emerging research area by integrating a conversational AI agent with an immersive VR environment, allowing users to engage in real-time dialogue with an expert avatar within simulated research ethics scenarios.

To conclude, the literature suggests a clear need for contextualized ethics training, especially in healthcare and animal research. This demands greater inclusion of supervisor perspectives in evaluating training effectiveness. Embracing interactive, AI-enhanced solutions to complement existing pedagogical methods is also advocated. The REACH-ETHICS project is well-positioned to contribute meaningful data on these points, especially as it directly engages supervisors in assessing the usefulness of its novel framework. Therefore, the aim of this study was to qualitatively investigate supervisors’ insights on the usefulness of a multimodal ethics training framework, with particular emphasis on its applicability in research supervision and its capacity to enhance student understanding of ethical principles.

II. METHODOLOGY

A. System Architecture and AI Components

The system integrates multiple AI models to deliver its full pipeline. Conversational processing was handled using the Qwen 2.5 (7B) large language model, deployed locally via Ollama, enabling reasoning and response generation with configurable temperature and endpoint settings. Retrieval-Augmented Generation (RAG) was implemented using GPT4All embeddings, which transformed source documents into semantic vectors stored within a local scikit-learn–based vector database.

For speech synthesis, the system employed the Tacotron2-DDC model via the Coqui TTS framework, trained on the LJSpeech dataset, with optional support for XTTS v2 to enable streaming, multi-speaker, and voice cloning

capabilities, accelerated through DeepSpeed where available. Audio transcription was performed using the Whisper model (base variant), converting spoken input into text prior to processing within the reasoning pipeline.

All models were executed locally, with GPU acceleration enabled through PyTorch with CUDA 12.6.

B. AI-driven VR Avatar Prototype Development

The VR project architecture was divided into three interconnected phases: the development of ethics knowledgebases, the deployment of a remote AI agent for responsive dialogue, and the creation of an immersive three-dimensional (3D) VR user interaction environment. Ethics knowledgebases were systematically developed through the collection, preprocessing, embedding, and storage of relevant ethical documentation and guidelines within a robust, queryable vectorstore. This foundational phase ensured accurate, context-aware, and appropriately constrained responses from the AI agent.

The second phase employed a remote backend server incorporating the ReAct AI agent framework. It integrated speech-to-text and text-to-speech functionality to enable dynamic verbal interaction. The AI agent drew upon the embedded ethical knowledgebase and maintained contextual awareness throughout interactions, thereby enhancing the realism and educational value of the dialogues.

It is important to note that in addition to speech-based interaction within VR, the system also supports text-based input, allowing users to submit prompts via a conventional interface when operating outside immersive environments. Regardless of the input modality, responses generated by the AI agent are primarily delivered in written form, ensuring consistency and traceability of information, with optional text-to-speech functionality enabling the avatar to vocalize responses. This design reflects a multimodal architecture in which VR serves as an enhanced interaction layer rather than the sole access point to the system.

The final phase comprised the delivery of two distinct 3D VR scenarios—a virtual boardroom and a didactic environment—developed using Unity with VR and extended reality (XR) toolkits. Each environment hosted an expert avatar capable of real-time interaction, synchronized speech, and corresponding lip movements. Users engaged with these avatars through direct spoken input via the Meta Quest headset, participating in scenario-driven dialogues designed to simulate authentic ethical deliberation. Fig. 1 provides a visualization of these three interconnected phases in the prototype development.

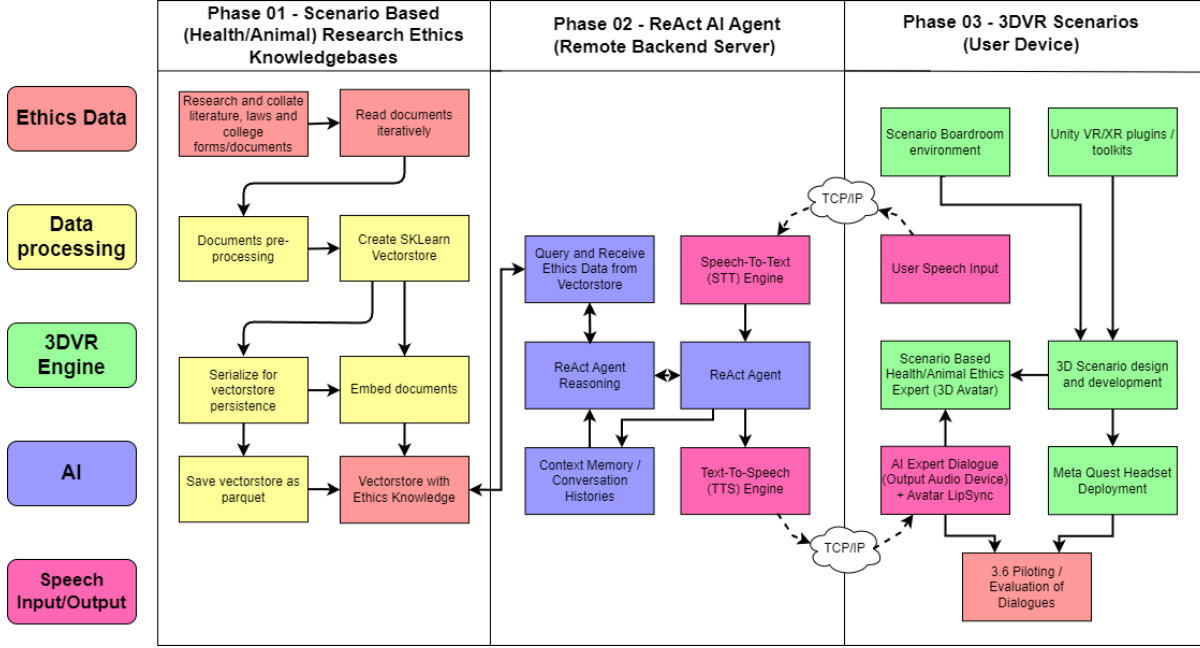


Fig. 1 Development pipeline. Source(s): Authors' own work

This interactive approach promoted deeper engagement with ethical concepts and supported improved retention and application of complex ethical standards across both animal research and human health contexts. In addition, the modular development approach allowed for scalability beyond the current implementation, permitting the creation of additional scenarios through flexible combinations of knowledgebases, avatars, and 3D environments. Furthermore, although the prototype was evaluated in a VR setting using a head-mounted display, the system architecture is platform-agnostic and can be deployed on standard desktop or laptop systems without requiring immersive hardware. This flexibility enables comparison with more conventional interfaces (e.g., chatbot or web-based systems) in future iterations. Fig. 2 gives a snapshot of one of the VR scenarios implemented in the prototype.

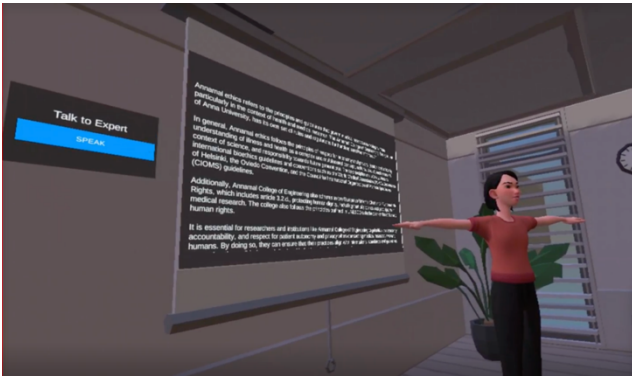


Fig. 2 Virtual Reality Scenario

C. Study Design

This study employed a qualitative focus group design supplemented by quantitative content evaluation to explore the perceived usefulness and educational potential of the REACH-ETHICS research ethics training framework. The session aimed to assess the feasibility of a multimodal, AI-driven ethics training approach integrating VR and Large Language Model (LLM) technologies.

D. Participants

A purposive sample of five (5) participants, all members of staff at the Malta College of Arts, Science and Technology (MCAST, Malta), participated in a focus group, reflecting the interdisciplinary nature of the project:

- Two (2) supervisors from health studies;
- Two (2) supervisors from animal studies;
- One (1) member of the technical development team responsible for the AI-driven VR prototype.

This composition enabled triangulated perspectives—academic, technical, and ethical—on the avatar's performance and its pedagogical implications.

E. Ethical Considerations

Ethical approval was granted by the MCAST Research Ethics Committee (Approval No. I001_2024). Informed consent was obtained before participation. Data was securely stored on password-protected institutional drives and deleted at the end of the project. All data collected was anonymized to protect identities.

F. Procedure

The approximately sixty-minute-long session took place in a collaborative meeting room equipped with an interactive digital board connected via High-Definition Multimedia Interface (HDMI) to a laptop running the AI-driven VR prototype developed as part of the REACH-ETHICS project. Participants were first briefed on the prototype's objectives and capabilities. They were then guided to:

1. Collectively formulate prompts within predefined thematic domains of research ethics (e.g., informed consent, confidentiality, dissemination of findings): to ensure comparability across interactions while preserving the authenticity of supervisory inquiry, participants were asked to formulate prompts within predefined thematic domains of research ethics. These domains reflected the principal areas addressed

in the prototype’s knowledge base and included: (1) ethical principles in human research (e.g., informed consent, confidentiality, anonymity, protection from harm, and research with vulnerable participants); (2) methodological ethics across research designs (e.g., quantitative, qualitative, ethnography, grounded theory, phenomenology, and observational methods); and (3) ethical and regulatory considerations in animal research. Within these domains, supervisors were free to pose either procedural questions (e.g., clarification of ethical requirements or regulatory obligations) or scenario-based questions involving practical research situations. Most prompts focused on clarifying ethical procedures or principles, while a smaller subset presented applied scenarios related to human/animal experimentation and field research contexts. Although the wording of prompts was not strictly standardized, their scope was constrained by these predefined ethics topics. This semi-structured prompting approach enabled the study to capture realistic supervisory questions while ensuring that all interactions addressed core areas of research ethics relevant to both human and animal studies;

2. Interact with the VR avatar, submitting these prompts to receive AI-generated responses;
3. Evaluate each response using a structured evaluation form, rating the outputs on a scale of 1 to 10 (scores 1 to 4: for hallucinations, off-topic, incomplete answers, etc.; scores 5 to 10: for acceptable responses of good to excellent quality) for Relevance, Correctness, Completeness, and Clarity;
4. Discuss their impressions, perceived benefits, and challenges of using the AI-driven VR avatar for ethics training.

Anonymized, detailed notes were taken during the discussion. Quantitative ratings were entered into an Excel database for descriptive analysis.

G. Data Preprocessing

Prior to analysis, participant ratings on each evaluative dimension of the AI-driven VR avatar prototype were screened for univariate outliers using the $1.5 \times$ interquartile range (IQR) criterion. A few observations were identified below the lower fence on two dimensions (Correctness [5 cases when using the IQR generated by the QUARTILE.INC function and corresponding lower fence = 6.5; 1 case when using the IQR generated by the QUARTILE.EXC function and corresponding lower fence = 5.25] and Clarity [1 case when using the IQR generated by the QUARTILE.INC function and corresponding lower fence = 5.5]). Inspection of the raw data and study notes confirmed that these represented valid participant responses rather than entry or procedural errors. To minimize their disproportionate influence while maintaining the full sample (health studies supervisors [$n = 2$] gave a consolidated score for health studies related prompts [$n = 8$], while animal studies supervisors [$n = 2$] gave individual scores to animal studies prompts [prompts $n = 5$, total $n = 10$], and prompts on general research ethical considerations [$n = 15$] were scored collectively, hence total sample size $n = 33$) and comparability across dimensions, the identified values were Winsorized—replaced with the lower fence boundary values. Because the rating scale comprised discrete integer points from 1 to 10, the decimal fence values were retained for

computation and reporting purposes. Descriptive statistics (medians and interquartile ranges [IQRs]) were recomputed on the Winsorized dataset. Comparisons with the original, unadjusted results showed no differences, indicating that the central tendency and dispersion of ratings were robust to the treatment of outliers.

H. Data Analysis

Quantitative data (ratings) were analyzed descriptively to obtain median value and the IQR for each evaluation dimension. IQRs were computed by calculating quartiles including the minimum and maximum values (inclusive) and also by calculating quartiles excluding the minimum and maximum values (exclusive) using the two functions available in Microsoft Excel for Microsoft 365 (QUARTILE.INC [inclusive] and QUARTILE.EXC [exclusive]). As the primary objective of this study was to assess the overall performance and educational adequacy of the AI-driven VR prototype, rather than to compare disciplinary perspectives, all supervisors’ evaluations were aggregated into a single dataset. This approach enabled a holistic appraisal of the system’s effectiveness across both health and animal studies contexts, which were treated as complementary case studies within the broader REACH-ETHICS framework.

Qualitative data (written justifications and salient verbatim notes taken during the group discussion) were subjected to thematic analysis following the six-step framework of [31]. First, the written justifications and notes recorded during the focus group were read repeatedly for familiarization. Second, initial codes were developed inductively from the content of the written justifications and notes. Third, related codes were grouped to identify candidate themes. Fourth, the themes were reviewed and refined to ensure internal consistency and distinction. Fifth, each theme was clearly defined and named. Finally, representative extracts were organized and integrated into the report.

III. RESULTS

A. Quantitative Findings

A comparison of descriptive statistics before and after Winsorization indicated that the adjustment of the low outlier scores had no effect on the overall results. The medians and IQRs for all dimensions did not change at all ($\Delta M = 0$; $\Delta IQR = 0$), and the rank order of the dimensions remained identical. These findings suggest that the central tendency and variability of participant ratings were robust to the treatment of outliers.

To evaluate the AI-driven VR prototype’s overall pedagogical and technical performance, participants’ ratings were pooled and analyzed collectively, rather than disaggregated by disciplinary background. The rationale for this approach was that the focus of the analysis was on the avatar’s general capability to provide relevant, accurate, and clear guidance on research ethics, rather than on disciplinary variations in perception. The inclusion of both health studies and animal studies supervisors served primarily to provide diverse but complementary case contexts, ensuring that the evaluation captured a range of ethical scenarios representative of the broader research ethics landscape.

Participants provided ratings for twenty-eight (28) AI-generated responses to ethics-related prompts (Health Studies $n = 8$; Animal Studies $n = 5$; General Research Ethical Considerations $n = 15$). The overall performance of the AI-

driven VR avatar was rated highly across all dimensions, as detailed in Table 1 below.

TABLE I.

Evaluation Dimension	Sample Size (n), Median (M) scores and Interquartile Ranges (IQR) for each dimension			
	n	M	IQR (using QUANTILE. INC)	IQR (using QUANTILE. EXC)
Relevance	33	9	2	2
Correctness	33	9	1	1.5
Completeness	33	8	2	2
Clarity	33	8	1	1.5

B. Qualitative Findings

Three recurrent patterns were identified across the supervisors' feedback:

- Relevance and focus of responses;
- Accuracy and completeness of information;
- Clarity and suitability for ethics learning.

All themes reflected participants' immediate qualitative impressions of the AI-driven VR avatar's responses to ethics-related prompts.

1) Theme 1: Relevance and focus of responses

Participants consistently highlighted that the avatar's responses were highly relevant to the questions posed. Typical comments included: "The answer is highly relevant to the user's question." This phrase was repeated across nearly all evaluated prompts, indicating strong alignment between the participants' prompts and the avatar's generated answers.

The consistency of this feedback suggests that the prototype effectively maintained topic focus across diverse research ethics themes such as confidentiality, informed consent, vulnerable participants, and dissemination practices.

2) Theme 2: Accuracy and completeness of information

Participants frequently noted that the information presented by the avatar was mostly correct and factually sound. This was not necessarily the case with the citation details provided by the avatar, which were at times judged by the participants to be incorrect. Other minor reservations regarding the accuracy and completeness of the avatar's responses were aired by the participants, these concerning occasional lack of contextual depth, insufficient examples, and overgeneralization (one participant commented that "most of the prompts are correct but a bit generic, and often repetitive between the different submitted research scenarios", and that "practical steps are often generic and not really applicable" and that "mostly, the content is correct, but very generic and simplified"). On the whole, though, most participants thought that the information provided by the avatar was correct, with one participant describing the answer given by the avatar as "factually correct and follows a logical structure". Participants proposed the future training of the AI-driven VR avatar on research conducted within the College to enhance contextual relevance.

The uniformity of these comments across prompts indicates confidence in the factual reliability of the prototype's content, with no recorded cases of hallucinated, misleading, or off-topic responses (one participant did note, though, that one of the answers given by the avatar to one of

the prompts felt somewhat ambiguous: "Reference to consent forms may be misleading here as they are not relevant"). The remarks also highlight the importance of refining the prototype to better differentiate between disciplinary contexts and to avoid overgeneralized formulations. Participants suggested that this could be achieved in future iterations of the AI-driven VR avatar by incorporating training data derived from research conducted within the College.

3) Theme 3: Clarity and suitability for ethics learning

A further recurring feature in the written justifications and notes was an emphasis on clarity and explanatory strength. Supervisors repeatedly observed that the answers were clear, well explained, and structured. One participant stated that "considering the broadness of the question the answer provides salient points that need to be considered in the context of [discipline]". Similarly, another participant stated that "salient points are addressed and clearly explained". Such feedback from the participants reinforces the perception that the prototype's language and presentation were accessible and educationally appropriate.

The participants' remarks suggest that the AI-driven VR avatar communicates complex ethical principles in a manner suitable for training and supervisory contexts.

4) Summary of Qualitative Insights

Overall, the qualitative feedback from participants reflected a consistently positive appraisal of the AI-driven VR avatar's capacity to address ethics-related prompts with relevance, factual accuracy, and clarity. Participants agreed that the system reliably stayed on topic and produced responses aligned with the intent of their questions, demonstrating the prototype's ability to maintain thematic focus across varied areas of research ethics such as confidentiality, informed consent, and dissemination practices.

Taken together, these findings indicate that the AI-driven VR avatar demonstrates high baseline performance in delivering relevant and accurate content, while also revealing areas for iterative improvement—particularly in enhancing contextual depth and tailoring examples to discipline-specific research settings. The overall tone of the participants' feedback was affirming, suggesting that the prototype offers a strong pedagogical foundation for ethics learning and supervision, and holds promise as a supportive tool within the REACH-ETHICS multimodal training framework.

IV. DISCUSSION

The purpose of this study was to explore research supervisors' perceptions of the REACH-ETHICS multimodal ethics training framework, particularly its AI-driven VR component. The findings indicate that supervisors viewed the prototype as both pedagogically valuable and technically capable of addressing key themes in research ethics education. Consistent with previous research advocating for interactive, immersive learning [19, 15], participants in this study found the avatar's responses to be highly relevant, clear, and accurate, suggesting that AI-based systems can meaningfully complement existing ethics training frameworks.

Participants consistently described the avatar's answers as relevant and clearly aligned with their prompts, reinforcing its potential to sustain learners' attention and engagement. This resonates with findings by [19], which reported that VR enhances comprehension and critical reflection in bioethics

learning. The REACH-ETHICS prototype similarly appeared to transform abstract ethical principles into accessible, conversational exchanges, which participants felt could make ethics more approachable for students.

The strong performance across all rated dimensions—relevance, correctness, completeness, and clarity—demonstrates the feasibility of using AI-driven VR as a teaching aid for ethics education. The combination of conversational AI and immersive VR also offers distinctive pedagogical advantages for ethics education. Immersive environments enable learners to engage with realistic scenarios that require reflection and decision-making, while conversational AI allows users to explore ethical questions dynamically through dialogue. This interaction encourages active learning rather than passive reception of information, enabling supervisors and students to interrogate ethical principles in context and receive immediate feedback. Research on immersive learning environments indicates that the sense of presence created by VR can increase cognitive engagement and support deeper understanding of complex topics [27, 28]. When coupled with AI-driven dialogue systems capable of responding to user questions, such environments may therefore facilitate reflective learning and ethical reasoning more effectively than traditional didactic instruction alone. It is important to note, however, that the present study does not position VR as a replacement for conventional chatbot or web-based interfaces. Rather, VR is explored here as a complementary modality that may enhance engagement and contextual understanding in specific learning scenarios. While text-based conversational systems remain effective for rapid information retrieval, immersive interaction may be particularly valuable in domains such as research ethics, where discussion, interpretation, and situational judgment are central. Future research should therefore include comparative evaluations across different interaction modalities to better understand their respective pedagogical affordances. By maintaining topical accuracy while simplifying complex ethical constructs, the prototype aligns with the growing call for multimodal learning experiences that merge cognitive and experiential engagement [32, 33].

While participants largely affirmed the correctness and coherence of the avatar’s answers, several highlighted limitations regarding contextual depth and practical specificity. Comments describing some responses as “generic”, “repetitive”, or “not really applicable” reflect a concern that the system occasionally lacked disciplinary nuance. These observations align with earlier critiques of traditional didactic learning tools that risk insufficient contextualization when handling specific ethical scenarios [14, 13].

The feedback also highlights the need for adaptive response algorithms that recognize the distinct ethical frameworks governing healthcare and animal studies. Future iterations of the avatar should ensure that the system is made more capable to generate more tailored, scenario-specific guidance.

The observation that one answer was potentially “misleading” due to its reference to consent forms reinforces the importance of precision and contextual appropriateness in AI-generated educational content. Reference [34] shows that failures to adhere to ethical principles in research may result either from actions undertaken without awareness that such

actions are unethical or—more importantly, and pertinent to the present discussion—from technological limitations.

Participants consistently praised the clarity and structure of the avatar’s responses. This perceived communicative clarity is central to the framework’s pedagogical promise. Reference [22] highlights “the importance of sufficiently verbalizing what characterizes research integrity and ethics” (p. 17), this in order that students understand “not only the ‘what,’ but also the ‘why’ of ethics and integrity” (p. 17). Participants’ comments that the avatar “provides salient points that need to be considered in the context of [discipline]” suggest that the prototype may facilitate reflective learning by prompting supervisors and students to contextualize ethical principles collaboratively.

Collectively, these findings point to the potential of AI-driven VR as a supplementary training resource for supervisors rather than a replacement for human instruction. The avatar’s capacity to deliver consistent, on-demand information aligns with supervisors’ dual needs for efficiency and standardization in ethics guidance [12, 11]. By integrating such a system within supervision, educators could devote more time to mentoring students on situational and reflective aspects of ethics, while relying on the avatar to cover baseline procedural or definitional content.

This integration also aligns with the goals of Responsible Research and Innovation (RRI), which emphasize inclusiveness and reflexivity in research training [35]. By embedding AI-VR technology within a multimodal framework—including seminars, written materials, and collaborative group sessions utilizing visualization walls—the REACH-ETHICS model can enhance participatory engagement between supervisors and students.

This study adds to a limited but growing body of evidence on supervisor-centric ethics training. Most existing frameworks, such as the EU-funded project Path2Integrity [18], prioritize student learning outcomes, whereas the REACH-ETHICS project deliberately positions supervisors as evaluators and co-designers. The current findings suggest that involving supervisors in both the testing and refinement of such frameworks enhances credibility and ensures alignment with real-world supervisory practices.

Importantly, the uniformly positive evaluations observed here demonstrate early proof-of-concept for the feasibility and acceptability of immersive, AI-based tools in ethics education—a domain traditionally resistant to digital innovation.

V. LIMITATIONS

While the study provides promising evidence, several limitations should be acknowledged.

First, the participant sample size ($n = 5$) was small and drawn from a single institution (MCAST), which constrains the generalizability of findings. The focus-group format prioritized depth over breadth, but larger and more diverse samples are needed to test whether these perceptions are shared across institutions and cultural contexts.

Second, the qualitative data collected in this initial evaluation consisted primarily of written comments and brief session notes. Although these provide valuable insights, they limit the ability to capture richer discursive dynamics and non-

verbal cues that might emerge in extended focus-group discussions or longitudinal observations.

Third, the prototype's AI outputs were evaluated using structured prompts designed by the same group of supervisors. While this ensured content relevance, it may have introduced some bias toward favorable assessments. Future evaluations could include independent prompts generated by external experts or students to better assess the system's adaptability.

Fourth, the prompts used to evaluate the AI-driven VR prototype were generated by the participating supervisors during the session. While this approach ensured ecological validity by reflecting authentic supervisory questions, it also introduced some variability in wording and scope. To mitigate this, prompts were guided within predefined thematic domains of research ethics, ensuring that all interactions addressed comparable ethical topics across human and animal research contexts.

Fifth, the study did not include a comparative evaluation between different interaction modalities (e.g., VR-based interaction versus text-based chatbot or web interfaces). As a result, the added value of immersive VR relative to more conventional approaches cannot be quantitatively established within the present study. Future work should therefore incorporate controlled comparisons to assess differences in engagement, learning outcomes, and usability across modalities.

Sixth, a key limitation of this study lies in the selection and configuration of the underlying AI models. The system relies on specific model families (e.g., Qwen 2.5, GPT4All embeddings, Tacotron2, and Whisper), each of which introduces inherent constraints related to performance, accuracy, and generalizability. Different large language models or embedding techniques may yield varying results in terms of response quality, contextual understanding, and domain adaptation. Furthermore, the use of locally deployed models, while beneficial for privacy and control, may limit scalability and access to more advanced or continually updated cloud-based models. As such, the findings should be interpreted in light of these technological choices, and future work could explore comparative evaluations across alternative AI model families.

Finally, the study did not include student users, whose perspectives are essential for assessing the framework's impact on engagement and learning outcomes. Subsequent phases of the REACH-ETHICS project will therefore triangulate supervisor and student experiences to obtain a holistic picture of pedagogical effectiveness.

VI. CONCLUSION

This study explored research supervisors' perceptions on the performance of the REACH-ETHICS research ethics AI-driven VR prototype which is part of a multimodal research ethics training framework. The findings reveal that participants perceived the system as relevant, accurate, and clear, highlighting its potential to support both ethics instruction and supervisory practice. Despite some critiques regarding generic or overgeneralized responses, the avatar was widely regarded as factually reliable, pedagogically suitable, and communicatively effective. These insights suggest that AI-VR systems can serve as valuable supplementary tools in ethics education—enhancing

engagement, consistency, and accessibility—when integrated within a broader multimodal training framework.

The study contributes to ongoing discussions about modernizing research ethics education through interactive, participatory, and technology-enhanced methods. By embedding supervisors' feedback early in the development process, the REACH-ETHICS project demonstrates an inclusive approach aligned with the principles of Responsible Research and Innovation.

Future research should expand on this pilot by involving larger, cross-disciplinary samples, testing student learning outcomes, and refining the AI model to deliver more context-sensitive and discipline-specific guidance. Through iterative co-design and evaluation, the framework holds strong potential to strengthen ethics education and foster a culture of research integrity across diverse research environments.

ACKNOWLEDGMENT

AI Disclosure: The authors declare that no artificial intelligence (AI) or AI-assisted technologies were used in the preparation of this manuscript.

REFERENCES

- [1] ALLEA, The European Code of Conduct for Research Integrity – Revised Edition 2023. Berlin, 2023.
- [2] M. F. Fathalla and M. M. Fathalla, "A practical guide for health researchers," World Health Organization, Regional Office for the Eastern Mediterranean, 2004.
- [3] World Health Organization, Standards and Operational Guidance for Ethics Review of Health-Related Research with Human Participants. Switzerland: WHO Press, 2011.
- [4] US Department of Health & Human Services. Ethical Codes & Research Standards. Updated April 29, 2024. Accessed March 17, 2026. <https://www.hhs.gov/ohrp/international/ethical-codes-and-research-standards/index.html>
- [5] J. Ana, T. Koehlmoos, R. Smith, and L. L. Yan, L. L., "Research misconduct in low-and middle-income countries," PLoS medicine, 10(3), e1001315, March 2013.
- [6] Liu, X., et al., "Knowledge and attitudes of Chinese medical postgraduates toward research ethics and research ethics committees: a cross-sectional study," BMC Medical Education, 23(1), 482, June 2023.
- [7] M. M. Than, H. Htike, and H. J. Silverman, "Knowledge, awareness, attitudes, and practices towards research ethics and research ethics committees among Myanmar post-graduate students," Asian bioethics review, 12(4), pp. 379-398, December 2020.
- [8] M. Yarborough and L. Hunter, "Teaching research ethics better: focus on excellent science, not bad scientists," Clinical and Translational Science, 6(3), pp. 201-203, June 2013.
- [9] R. R. Lotto, "Post-graduate students' perception of research ethics training: a qualitative study," Innovations in Practice, 12(2), pp. 24-33, December 2018.
- [10] K. Chatfield and E. Law, E., "I should do what? Addressing research misconduct through values alignment," Research Ethics, 20(2), pp. 251-271, February 2024.
- [11] N. Evans, et al., "VIRT2UE: A European train-the-trainer programme for teaching research integrity," Research Ethics, 20(2), pp. 187-209, April 2024.
- [12] N. Evans, et al., "Virtue-based ethics and integrity of research: train-the-trainer programme for upholding the principles and practices of the European Code of Conduct for Research Integrity (VIRT2UE)," Research Ideas and Outcomes 7: e68258, June 2021.
- [13] L. Naimi, "Strategies for teaching research ethics in business, management and organisational studies," Electronic Journal of Business Research Methods, 5(1), pp. 29-36, July 2007.
- [14] V. Burr and N. King, "'You're in cruel England now!': Teaching research ethics through reality television," Psychology Learning & Teaching, 11(1), pp. 22-29, January 2012.

- [15] A. Tammeleht, M. J. Rodríguez-Triana, K. Koort, and E. Löfström, "Collaborative case-based learning process in research ethics," *International Journal for Educational Integrity*, 15(1), 6, October 2019.
- [16] A. Halkoaho, M. Matveinen, V. Leinonen, K. Luoto, and T. Keränen, "Education of research ethics for clinical investigators with Moodle tool," *BMC medical ethics*, 14(1), 53, December 2013.
- [17] K. Sim, M. Y. Sum, and D. Navedo, "Use of narratives to enhance learning of research ethics in residents and researchers," *BMC Medical Education*, 15(1), 41, March 2015.
- [18] Research Ethics and Integrity. European Commission. Accessed March 17, 2026. https://research-and-innovation.ec.europa.eu/strategy/support-policy-making/scientific-support-eu-policies/research-ethics-and-integrity_en
- [19] A. L. Harfouche and F. Nakhle, "Creating bioethics distance learning through virtual reality," *Trends in Biotechnology*, 38(11), pp. 1187-1192, November 2020.
- [20] D. K. Plemmons and M. W. Kalichman, "Reported goals for knowledge to be learned in responsible conduct of research courses," *Journal of Empirical Research on Human Research Ethics*, 2(2), pp. 57-66, June 2007.
- [21] A. L. Antes, et al., "A meta-analysis of ethics instruction effectiveness in the sciences," *Ethics and Behavior*, 19(5), pp. 379-402, September 2009.
- [22] H. Hyytinen and E. Löfström, "Reactively, proactively, implicitly, explicitly? Academics' pedagogical conceptions of how to promote research ethics and integrity," *Journal of Academic Ethics*, 15(1), pp. 23-41, March 2017.
- [23] Virtue based ethics and Integrity of Research: Train-the-Trainer program for Upholding the principles and practices of the European Code of Conduct for Research Integrity. European Commission. Accessed March 17, 2026. <https://cordis.europa.eu/project/id/787580>
- [24] R. Winkler and M. Söllner, "Unleashing the potential of chatbots in education: A state-of-the-art analysis," *Academy of Management Proceedings*, 2018(1), 15903, July 2018.
- [25] C. W. Okonkwo and A. Ade-Ibijola, "Chatbots applications in education: A systematic review," *Computers and Education: Artificial Intelligence*, 2, 100033, January 2021.
- [26] S. Wollny, et al., "Are we there yet?-a systematic literature review on chatbots in education," *Frontiers in Artificial Intelligence*, 4, 654924, July 2021.
- [27] J. Radianti, T. A. Majchrzak, J. Fromm, and I. Wohlgenannt, "A systematic review of immersive virtual reality applications for higher education: Design elements, lessons learned, and research agenda," *Computers and Education*, 147, 103778, April 2020.
- [28] M. Coban, Y. I. Bolat, and I. Goksu, "The potential of immersive virtual reality to enhance learning: A meta-analysis," *Educational Research Review*, 36, 100452, June 2022.
- [29] X. Pan and A. F. D. C. Hamilton, "Why and how to use virtual reality to study human social interaction: The challenges of exploring a new research landscape," *British Journal of Psychology*, 109(3), pp. 395-417, August 2018.
- [30] M. Slater and M. V. Sanchez-Vives, "Enhancing our lives with immersive virtual reality," *Frontiers in Robotics and AI*, 3, 74, December 2016.
- [31] V. Braun and V. Clarke, "Using thematic analysis in psychology," *Qualitative Research in Psychology*, 3(2), pp. 77-101, January 2006.
- [32] H. T. Crogman, V. D. Cano, E. Pacheco, R. B. Sonawane, and R. Boroon, "Virtual reality, augmented reality, and mixed reality in experiential learning: Transforming educational paradigms," *Education Sciences*, 15(3), 303, February 2025.
- [33] A. D. Samala, et al., "Virtual reality in education: global trends, challenges, and impacts—game changer or passing trend?" *Discover Education*, 4(1), pp. 1-45, December 2025.
- [34] A. C. Cairns, et al., "A phenomenographic study of scientists' beliefs about the causes of scientists' research misconduct," *Research Ethics*, 17(4), pp. 501-521, September 2021.
- [35] RRI Tools - Fostering responsible research and innovation. CaixaResearch. Accessed March 17, 2026. <https://caixaresearch.org/en/caixaresearch-rri-tools>

Assessing students' performance and supporting building a professional identity by writing amidst the rise of AI.

Ema Jerkovic
Faculty of Social Work
Baden-Wuerttemberg State Cooperative
University
Stuttgart, Germany
ema.jerkovic@dhbw-stuttgart.de

Abstract— This article conceptualizes academic writing as a culturally situated social practice central to disciplinary enculturation, epistemic learning, and professional identity formation in higher education. Drawing on cultural anthropology, writing studies, and higher-education pedagogy, it understands writing tasks as formative sites where students acquire disciplinary norms, develop authorial voice, and construct a professional writing habitus. In light of the rapid diffusion of generative artificial intelligence (GenAI), the paper examines how AI-mediated writing reshapes these functions. While research highlights gains in efficiency, linguistic support, and motivation, significant concerns arise regarding epistemic outsourcing, reduced metacognitive engagement, and weakened authorial agency. Based on a systematic literature review, the article argues for redesigning writing tasks toward process-oriented, reflective, and transparent formats that integrate critical AI literacy. Such approaches aim to preserve writing as a practice of learning, identity work, and professional becoming within and beyond academia.

Keywords— academic literacies; disciplinary enculturation; generative artificial intelligence; epistemic agency; writing pedagogy; professional identity; AI literacy; higher education

I. INTRODUCTION AND METHODOLOGICAL APPROACH

Academic Writing has long been a central practice of knowledge production, assessment, and socialization in higher education. Through writing, students do not merely demonstrate what they know; they learn how to participate in disciplinary communities, adopt epistemic norms, and develop a professional identity as academic and, later, workplace writers. Recent advances in generative artificial intelligence (GenAI), however, profoundly challenge this established role of writing. AI-supported text generation alters how academic texts are produced, revised, and evaluated, raising fundamental questions about learning, epistemic agency, authorship, and identity formation in higher education.

This paper approaches these developments from an explicitly interdisciplinary perspective that brings together cultural anthropology, writing studies and writing didactics, as well as educational and higher-education research. From a cultural-anthropological viewpoint, academic writing is understood as a situated social practice embedded in institutional, historical, and discursive contexts. Writing tasks function as rites of passage through which learners are enculturated into disciplinary ways of thinking, arguing, and positioning themselves. Writing-didactic and educational research complements this perspective by conceptualizing writing as a process of epistemic engagement, metacognitive reflection, and competence development rather than a mere technical skill or assessment product.

Against this background, the article examines how the increasing availability of GenAI tools intersects with the formative functions of academic writing. While current debates often focus on questions of plagiarism or efficiency, this contribution argues that the more profound challenge lies in the potential transformation of writing as a learning practice, as a medium of disciplinary socialization, and as a site of professional identity construction. The paper therefore asks how writing tasks in higher education can be rethought in the age of AI so that they continue to support deep learning, epistemic agency, and inclusive academic socialization.

Methodologically, the article is based on a systematic, literature-driven analysis of international and German-language research from writing studies, higher-education pedagogy, educational psychology, and related fields. Rather than presenting new empirical data, the contribution aims to conceptually clarify current developments and to derive pedagogically grounded perspectives for transforming academic writing tasks under conditions of AI-mediated writing.

II. THE ROLE OF WRITING TASKS IN ACADEMIA

Writing in academia is often conceived as a means for assessment, reproduction of knowledge, or demonstration of what has been learned. This proposes a very narrow view. However, a growing body of research argues that writing tasks serve a far deeper and more formative role, namely as instruments of **academic socialization** and **disciplinary enculturation**, through which students internalize the norms, genres, epistemic practices, and discursive conventions of their field. Empirical and theoretical contributions from writing studies, writing-didactics, and higher-education pedagogy substantiate this broader function.

A. Theoretical perspective: Writing and disciplinary enculturation

Writing is central to the process by which graduate students become members of disciplinary communities. Negotiating and demonstrating disciplinary knowledge, textual genres and identities is done by course papers, theses, grant applications and similar texts bound to academia [1]. Writing is understood not merely as transcription of thought but as literate activity embedded in social, institutional, and historical contexts.

In this sense, writing tasks in higher education become sites of enculturation which means of learning how to think, argue and write like a member of the discipline [2].

B. Empirical support for writing-based learning

Empirical studies show that writing tasks foster both disciplinary literacy and subject learning when embedded in instruction. Knowledge about academic writing often emerges through interaction (e.g. feedback, draft

discussions), highlighting writing as a social practice [3]. Discipline-specific writing tasks combined with process-oriented support can enhance writing performance, content knowledge, and students' epistemic engagement [4].

C. Institutional support and writing pedagogy

For writing tasks to fulfill their socializing function, institutional support is essential. Writing centers play a crucial role in guiding students, especially L2 writers, through non-linear, dynamic processes of academic socialization [5]. Similar approaches in German-speaking contexts show that writing instruction integrated with subject expertise supports theory-based learning, including in fields such as sports education and STEM disciplines [6][7].

D. Diversity and Inclusivity

Disciplinary socialization through writing is shaped by students' heterogeneous linguistic and educational backgrounds. Research emphasizes the need for inclusive writing task design, awareness of multilingualism, and differentiated support to ensure equitable access to disciplinary practices [5][8].

E. Implications for higher education

Writing assignments should be systematically integrated into subject teaching, designed as guided processes with feedback and revision, and supported by institutional structures. Lecturers act as mediators of disciplinary writing cultures, as writing is central not only to learning but also to students' disciplinary identity formation.

F. Writing tasks as disciplinary initiation

When properly designed and supervised, writing tasks in academia serve as rites of passage into the discipline: they are more than just exercises or exams; they are a means of social and epistemic integration. They enable students to position themselves as part of an academic community, learn disciplinary ways of thinking and expressing themselves, and develop their professional identity. In this sense, academic writing is not just an output, but a central component of input and socialization in higher education.

III. IMPACT OF AI ON WRITING SKILLS

The advent of generative artificial intelligence (GenAI) tools (e.g., large language models, writing assistants) is reshaping academic writing practices. What once were human-centered processes of drafting, revising, and reflecting, can now be supported or in some cases largely substituted by algorithmic assistance. This raises fundamental questions: Does AI support improve writing skills, or does it risk undermining deep learning, metacognitive engagement, and academic integrity? This chapter addresses these questions by synthesizing recent empirical and theoretical research.

A. Potential benefits of GenAI for academic writing

Empirical research indicates that GenAI can substantially increase writing productivity and text quality. Studies report significant reductions in writing time alongside improved assessment outcomes, particularly for less confident writers [9]. GenAI also offers notable support for L2 writers by facilitating idea generation, translanguaging practices, and iterative revision processes, thereby improving linguistic quality and revision skills [10][11][12]. In addition, students often report increased motivation and perceived self-efficacy

when using AI tools in a controlled manner [13][14]. These findings highlight the growing importance of GenAI literacy as a key competence encompassing reflective, ethical, and self-regulated use of AI in writing [12][14].

B. Risks and Challenges

Despite these benefits, research identifies significant risks. Uncritical adoption of AI-generated text can reduce metacognitive engagement and lead to superficial writing; studies show that learning gains depend on active revision and critical interaction with AI output [15][16]. Insufficient AI literacy may also create an "agency gap," reducing students' ability to write independently without AI support [17]. Ethical concerns remain central, as GenAI use raises issues of authorship, originality, and academic integrity, especially when used opaquely [18][19][20]. Moreover, AI feedback often remains surface-oriented and may neglect disciplinary depth [12].

Beyond integrity, GenAI poses challenges to the development of writing identity and epistemic agency. Academic writing is an identity-forming practice through which writers develop voice, stance, and disciplinary positioning [21][22][23]. Overreliance on AI risks diluting authorial voice and outsourcing epistemic writing processes that are essential for knowledge construction and conceptual development [24][25][26][27]. These effects may be unevenly distributed, as students' ability to benefit from AI depends on prior skills, access, and AI literacy, potentially reinforcing existing educational inequalities [28][29][30].

From a cognitive perspective, extensive AI use may lead to long-term skill atrophy through cognitive offloading, particularly in planning, coherence building, and rhetorical decision-making [31][32]. Finally, AI-generated feedback can disrupt feedback literacy and peer learning by bypassing dialogic and process-oriented writing practices [33].

C. Implications for writing pedagogy

Taken together, these findings suggest that GenAI should not be treated solely as a threat but as a catalyst for rethinking writing pedagogy and assessment [34][35]. From an Academic Literacies perspective, writing instruction must integrate critical AI literacy, ethical reflection, and task designs that preserve epistemic engagement while making AI use transparent [36][37]. Process-oriented assessment, reflective commentary, and genre-aware tasks may help align AI-supported writing with the goals of disciplinary learning and identity development.

D. Transforming academic writing tasks in the age of AI

AI-based text generation challenges traditional writing assessments by obscuring writing processes, learning trajectories, and identity formation. As a result, writing tasks need to be redesigned to foreground writing as a process of learning, reflection, and professional development rather than solely as a final product.

a) *Portfolios and process-oriented assessment:* E-portfolios offer a powerful alternative by documenting artefacts such as drafts, revisions, reflections, and multimodal materials, thereby making learning processes visible [38]. They enable the integration of formative and summative assessment, foster learner agency, and support self-directed learning [39][40]. In AI-supported contexts, portfolios

counteract product-oriented outsourcing by revealing how texts are developed and revised. Their bridge function connects learning, assessment, and identity development, and they have proven feasible as examination formats in higher education [41][42]. When embedded within competence-oriented curricula and constructive alignment, portfolios show strong pedagogical potential despite higher workload demands [43].

b) Iterative writing, collaboration and feedback: Beyond portfolios, smaller iterative tasks (e.g. reading journals, wikis, staged assignments) emphasize drafting, peer feedback, revision, and reflection. Such formats make epistemic writing processes visible, strengthen metacognition, and support writing as knowledge construction and disciplinary positioning.

E. Integrating AI literacy into writing tasks

Rather than prohibiting AI use, writing instruction should integrate critical AI and media literacy. Tasks can explicitly require AI-generated drafts followed by critical revision, source checking, and reflective documentation, positioning AI as a learning tool rather than a shortcut [44]. This approach requires clear institutional frameworks, transparent assessment criteria, and appropriate digital infrastructure, as emphasized in e-portfolio implementation guidelines [45].

F. Writing identity and institutional conditions

Process-oriented formats create space for students to develop a reflective writing identity by engaging with strategies, argumentation, uncertainty, and feedback. However, their success depends on institutional support, assessment literacy, and a cultural shift toward openness and reflection. Despite increased workload, such formats help preserve writing as a site of learning and identity formation in an AI-rich academic environment and call for a reassessment of product-focused academic assessment.

IV. IMPACT ON BUILDING A PROFESSIONAL IDENTITY

The process of writing in academic contexts does not merely serve to demonstrate knowledge or satisfy assessment criteria but also plays a decisive role in forming a professional identity. With the rise of AI and corresponding shifts in assessment formats, it is important to reconsider how such identity formation is supported or potentially undermined.

A. Writing as Identity Work and Acculturation into Academic Community

An influential study by Amanda French shows that academic writing in higher education can be conceptualized as identity-work. French argues that through repeated engagement in writing practices, from the first undergraduate essays through postgraduate dissertations, individuals gradually form what she terms a professional writing in higher education habitus [46].

This habitus encompasses not only technical competence, but dispositions, expectations, and tacit cultural norms of what it means to be a member of an academic community [46]. As such, writing becomes a site where discipline-specific belonging, and professional self-conception are mediated.

Similarly, work on writer identity in academic writing more broadly argues that writing is not just content transmission but a construction of identity through discourse. For example, Roz Ivanic shows that writers (including students) adopt and negotiate subject-positions through language and discourse conventions, thereby aligning (or resisting) the dominant norms of their academic or social communities [21].

In short, when students write, they also become integrated into a community; they learn to become members of that community, not only in terms of knowledge, but also in terms of identity, style, authority, and voice.

B. From Student to (Prospective) Professional: Authorial Voice, Agency and Ownership

Part of professional identity formation lies in developing an authorial voice through writing. Research among EFL (English as Foreign Language) learners demonstrates that identity construction in academic writing often involves a shift from a weak, tentative authorial presence toward a more confident, self-positioned scholarly voice. For example, in a recent empirical study of Chinese English Learners the authors show how novice writers, initially rely on direct intertextual copying, but gradually move toward indirect intertextuality, signaling their own ideas and positioning themselves as scholarly authors [47].

This transformation, which goes from imitation to appropriation, from passive reproduction to active authoring, is a core dimension of academic self-formation. As such, writing tasks, exposure to reading and writing practices, and guided cycles of reading, writing, feedback and revision foster not just competence but identity.

Moreover, educational research suggests that writing (and reading) practices shape not only the writer's identity but also their epistemic stance. This means their sense of what counts as knowledge, how knowledge is produced, and what it means to belong to a scholarly discipline [49]. As noted in a review on academic literacies and disciplinary identity, "[s]tudents use writing to demonstrate their positions with respect to discipline-specific knowledge and norms." [48].

Therefore, academic writing is practice **and** socialization.

C. Relevance under AI: Why Process-Oriented and Reflective Writing Formats Matter for Identity Development

With the rise of AI-based generation tools (LLMs, etc.), the traditional final-product-oriented writing tasks (essays, theses, exams) risk becoming shallow exercises in requirement fulfillment. There is the possibility that in the end students only create a polished text on paper, but lack an authentic engagement with the writing process, intellectual struggle, reflection, and authorial decision-making. This threatens the identity-work dimension of academic writing.

Transforming writing tasks toward process-oriented, reflective, and iterative formats (e.g., portfolios, reading journals, wikis, staged writing with feedback loops) can counter this risk, by making visible and evaluating not only final output, but also the writing journey, decisions, re-workings, and reflections, as shown above. In doing so, such formats support the emergence of a professional writing identity grounded in practice, self-awareness, and authorial agency rather than mere production.

Moreover, by integrating iterative writing, reflection, and possibly peer-review or collaborative writing, students engage not only with content but with the norms, values, and responsibilities of academic authorship. This supports a deeper internalization of what it means to belong to an academic community and fosters a sense of ownership, belonging, and professional identity.

D. Broader Professional Identity beyond Academia:

Writing Competence as Professional Capital

Beyond academic identity, writing competence is a crucial component of professional identity in many non-academic but still professional contexts. For instance, German-language research on “writing professionals” highlights that many employed individuals engage in professional writing at work, yet lack formal training in writing competence and thus remain under-studied in writing didactics [50].

Similarly, literature on professional identity formation emphasizes that professional identity is not static or purely role-based; it evolves through experiences, reflection, and alignment with values, responsibilities and social recognition [51] [52].

Thus, embedding reflective, process-oriented writing practices and supporting identity work during academic education can serve as a foundation for broader professional identity formation — equipping students with writing competence, self-conception, professional values, and authorial agency useful in academic and non-academic professional contexts.

E. Implications and Challenges

- Pedagogical implication: If academic institutions adopt more process-oriented and reflective writing tasks (portfolios, staged writing, peer review, reading/writing journals), they actively support students’ development of a writing-based professional identity, not just competence.
- Institutional implication: This requires institutional commitment to not only redesign assessments, but to provide guidance, scaffolding, feedback, and space for reflection. It also calls for recognition of writing identity formation as a legitimate educational goal, not just output.
- Equity and inclusion consideration: For learners coming from diverse linguistic, cultural or educational backgrounds (e.g., non-native speakers, first-generation students), supporting identity construction through writing can help bridge differences in academic “habitus” and foster inclusion. The identity-construction studies of EAL students suggest that tailored support can strengthen their authorial voice and sense of belonging [47] [53].
- Risk: If writing becomes too standardized, formulaic or automated (e.g., AI-generated essays without reflection), the opportunity for identity work is lost. This can lead to negative consequences for students’ sense of agency, authorship, belonging, and professional self-conception.

To sum up, writing in higher education is not merely a means to demonstrate knowledge but also a central medium through which students learn to think, argue, belong, and position themselves as members of academic and professional communities. As research shows, academic writing is identity work, shaped by habitus, discourse communities, practices, and socio-cultural contexts [21][46][49].

Therefore, and especially in the context of rising AI, it is urgent to reconceive writing tasks not only as assessment instruments but as formative, reflective, identity-building practices. Transforming writing assignments accordingly can support the development of professional writing identity, epistemic agency, and belonging. These outcomes matter deeply for both academic and broader professional trajectories.

V. ACADEMIA AND WORKING WORLD COLLIDING THROUGH WRITING SKILLS AND IDENTITY

The transition from academic studies into the professional world often involves more than a change of topic or employer. Instead it confronts graduates with different expectations, genres, and discursive contexts for writing and knowledge production. With writing at the intersection of academia and workplace, the nature, purpose, and precarity of identity shaped through writing are exposed to rupture and re-negotiation. This collision reveals both the fragility of academic-professional identity and its potential as a resource for bridging academic and non-academic life.

Classical writing-studies research has long argued that academic writing is embedded in a distinct socio-cultural context, one that differs significantly from writing in professional workplaces. In a seminal comparative study [54], analyses writing practices across university courses and corresponding professional workplaces (e.g., law, public administration, social work, architecture). The study demonstrates that writing in the academic sense serves particular social motives, involves particular genres and institutional logics, and socializes writers into disciplinary and academic discourses whereas in workplaces writing is deployed in functionally different ways, namely as a tool to achieve organizational tasks, as part of collective activities, and embedded in institutional, social and material settings with different audiences, constraints, and power relations.

In this sense, when students graduate and enter professional contexts, they may experience a genre shock: the genres, conventions, purposes, and rhetorical expectations they learned at university may not align with those required at work. The shift thus pressures their academic writing habitus and challenges their identity as writers. They are no longer strictly scholars, but practitioners, communicators, or writing professionals.

A. Writing Competence as Employability and Social Participation

Writing competence is a key prerequisite for professional and societal participation, also across many sectors beyond traditional writing professions as a central argument in a study on writing and professionals [50] shows. It documents that the vast majority of employed person perform writing tasks in their jobs, including reports, documentation, communication, internal/external texts, but seldom receive formal training in professional writing.

This suggests that the writing skills acquired in academia constitute a form of professional capital: they enable graduates to navigate the communicative demands of the workplace, to author texts under different institutional and social constraints, and to contribute to knowledge work beyond academia. The academic-professional boundary thus becomes porous, with writing competence serving as a bridging competence.

Given societal transformations, as is the shift from production to knowledge- and service-based economies, this bridging becomes even more critical. Now writing competence becomes part of employability, lifelong learning, and civic/organizational participation. In the mentioned study [50] it is argued that writing in its broad sense (not only creative or journalistic) deserves more attention in continuing education and professional development.

B. Identity Re-negotiation: From “Academic Writer” to “Writing Professional”

For many graduates, the move into the workplace entails a transformation of writing identity. The “academic writer” identity which is shaped by disciplinary conventions, scholarly values, and institutional belonging, may no longer fit. Instead, individuals may need to develop a new identity as writing professionals in contexts of professional communication, corporate or administrative writing, technical or organizational texts, or cross-disciplinary knowledge work.

This identity renegotiation requires not only competence, but reflection, adaptability, and often a reorientation of what good writing means. Because academic writing pedigrees do not always align with workplace demands, individuals may experience a form of dislocation which can be treated as an opportunity for identity crafting. The comparative study offered in *Worlds Apart* [54] demonstrates that the social motives, institutional logics, and communicative contexts shift dramatically; to succeed, writers must learn to situate their writing in new activity systems, address different audiences, and adopt new genres.

Moreover, these shifts often involve not just form or style, but social roles: from solitary academic authorship to collaborative workplace documentation; from rhetorical autonomy to institutional accountability; from theoretical argumentation to pragmatic communication. These are identity-relevant transformations.

C. The Collision Intensified by AI: New Challenges and Potentials

With the rise of generative AI and other digital writing supports, the collision between academic and working-world writing becomes more complex. On one hand, AI may flatten differences by offering writing as service which may lower the threshold to produce texts, making writing tasks more uniform across contexts; this risks eroding the distinctiveness of academic writing identity, and may encourage viewing writing as mere output rather than as a meaning-making, identity-shaping practice.

On the other hand, AI also brings opportunities: if future professionals (whether in academia or outside) learn to use AI responsibly and reflexively, writing competence becomes a more valuable form of professional capital. This includes the learning of meta-cognitive awareness, genre sensitivity, and ethical/epistemic judgment. Drawing on writing competence from academia, individuals may reconfigure their writing-professional identity in synergy with AI-mediated practices, bridging academic literacy and workplace demands.

Recent empirical work hints at this complexity. For example, interviews with writing professionals reveal strategies of both resistance and appropriation in response to generative AI [55]. The study shows that some professionals double down on human-centered writing craft to preserve

credibility and identity, others integrate AI to streamline workflow and redefine their professional role.

Thus, the collision between academia and the working world becomes a triad: academic writing tradition ↔ workplace writing practices ↔ AI-mediated writing environments. Identity formation and re-negotiation occur within this dynamic interplay.

D. Implications for Higher Education, Writing Didactics and Career Preparation

Given this collision, higher education institutions and in particular writing-didactic settings have a responsibility to prepare students not only for academic writing, but for writing-professional life. This requires a shift in writing pedagogy and assessment, as argued previously, but also an expanded view of writing competence as part of employability and professional identity formation.

Concretely, this implies:

- Incorporating genres and tasks in curricula that mirror professional writing settings (reports, documentation, collaborative writing, workplace-relevant genres).
- Offering training for writing professionals (not only future scholars). As examples via seminars/workshops, writing centres or continuing-education offers outside of academia, as described in [50].
- Developing students’ genre awareness, audience sensitivity, and writing adaptability so they can successfully transition from academic to workplace contexts.
- Fostering reflection on writing identity: encouraging students to consider how their writing roles may change, how to negotiate institutional and social norms, and how to maintain agency and authorship under new conditions (e.g., collaborative, AI-mediated, institutional writing).

In doing so, higher education contributes not only to academic socialization, but also to lifelong writing competence. Only then are graduates equipped to become reflective, adaptable, and ethically aware writing professionals.

VI. CONCLUSION

This paper has argued that academic writing must be understood as far more than a neutral vehicle for assessment or textual output. Drawing on cultural anthropology, writing studies, and educational research, it conceptualized writing tasks as central instruments of disciplinary enculturation, epistemic learning, and professional identity formation. Writing in higher education functions as a social practice through which students learn not only what to know, but how to think, argue, and position themselves within academic and professional communities.

The rapid diffusion of generative AI fundamentally unsettles this role of writing. While empirical research demonstrates clear benefits of AI-supported writing in terms of efficiency, language support, and motivation, the literature also reveals substantial risks such as diminished metacognitive engagement, epistemic outsourcing, erosion of authorial voice, integrity concerns, and the amplification of existing educational inequalities through uneven AI literacy. These risks are not incidental side effects but touch the core pedagogical function of writing as a site of learning and identity work.

Rather than framing AI primarily as a threat to academic standards, this article has proposed a reframing, namely that AI should be treated as a catalyst for rethinking writing pedagogy and assessment. The central challenge is not whether AI is used, but how writing tasks are designed to preserve epistemic agency, reflective engagement, and disciplinary socialization under AI-mediated conditions. Process-oriented, reflective, and transparent writing formats emerge from the literature as particularly promising responses. One of these formats that has been discussed in more depth in this article are portfolios. They shift attention from polished end products to learning trajectories, rhetorical decision-making, and identity development, thereby counteracting the risk of writing becoming an act of epistemic outsourcing.

At the same time, the analysis highlights that transforming writing tasks is not merely a didactic adjustment but an institutional and cultural undertaking. Supporting meaningful writing development in the age of AI requires infrastructural resources, assessment literacy among instructors, clear institutional frameworks for AI use, and a shared understanding of writing as a formative practice. Without such conditions, even well-intentioned innovations risk reinforcing workload pressures or reproducing inequalities.

Beyond higher education, the paper has shown that writing competence constitutes a form of professional capital that bridges academia and the workplace. As graduates move between academic, professional, and AI-mediated writing environments, their writing identity is continuously renegotiated. Higher education therefore carries a responsibility not only to socialize students into academic genres, but to equip them with transferable writing competence, genre awareness, and reflective agency that remain viable beyond the university.

In conclusion, writing is not simply a means of producing text, but a central practice of learning, knowing, and becoming. This meaning of writing has been long emphasized in writing research but is made visible by the rise of generative AI. If academic writing tasks are redesigned to foreground process, reflection, and epistemic responsibility, AI need not undermine writing competence or identity formation. Instead, it can coexist with pedagogical frameworks that foster critical AI literacy, inclusive academic socialization, and the development of resilient professional writing identities. The future of academic writing, therefore, lies not in defending traditional formats, but in deliberately shaping writing practices that remain educational convincing in an AI-saturated world.

REFERENCES

- [1] P. Prior, *Writing/Disciplinarity. A Sociohistoric Account of Literate Activity in the Academy*, London: Routledge, 1999.
- [2] P. Prior, R. Bilbro, "Academic Enculturation. Developing Literate Practices and Disciplinary Identities", in *University Writing. Selves and Texts in Academic Societies*, M. Castelló, C. Donahue, Eds. Leiden: Brill, 2012, pp. 19-31.
- [3] K. Kaufhold, "The dynamics of building academic writing knowledge in interaction," *Journal of English for Academic Purposes*, vol. 75, Nr. 101518, n.p., April 2025.
- [4] L. Holdinga, J. van Drie, G. Rijlaarsdam, "Enhancing disciplinary literacy through integrated writing and process instruction: an intervention study", *Reading and Writing*, n.p., February 2025.
- [5] O. Solmaz, "The Role of Writing Center in Academic Writing Socialization of Second Language Graduate Students", *Acta Educationis Generalis*, Vol. 11, Iss. 3, p. 1-22, December 2021.
- [6] E. Langelahn, A. Menze-Sonneck "Zur Verknüpfung wissenschaftlichen Schreibens und fachlichen Lernens in der Sportlehrramtsausbildung – eine Studie zur Überprüfung der Wirksamkeit effektiver Schreibaufgaben", *Zeitschrift für Studium und Lehre*, Vol. 3, Iss. 1, p. 5-16, 2020.
- [7] A. Hirsch-Weber, S. Scherer, Eds., *Wissenschaftliches Schreiben in Natur- und Technikwissenschaften. Neue Herausforderungen der Schreibforschung*. Wiesbaden: Springer, 2016.
- [8] M. Brinkschulte, I. Barczaitis, E. Grieshammer, M.-E. Stoian, *Mehrsprachiges Schreiben im akademischen Kontext unterstützen, anleiten, begleiten*. Bielefeld: wbv, 2022.
- [9] A. E. Conell Pensky, J. Usdan, H. Chang, "Generative AI's Impact on Graduate Student Professional Writing Productivity and Quality", *International Journal of Artificial Intelligence in Education*, n.p., November 2025.
- [10] W. Yang, C. Lin, "Translanguaging with generative AI in EFL writing. Students' practices and perceptions", *Journal of Second Language Writing*, Vol. 67, January 2025.
- [11] M. Mekheimer, "Generative AI-assisted feedback and EFL writing: a study on proficiency, revision frequency and writing quality", *Discover Education*, Vol. 4, June 2025.
- [12] S. Li, "Generative AI and Second Language Writing", *Digital Studies in Language and Literature*, Vol. 2, Iss. 1, p. 122-152, June 2025.
- [13] J. Zare, A. Al-Issa, F. Ranjbaran Madiseh, "Interacting with ChatGPT in essay writing. A study of L2 learners' task motivation", *ReCall*, Vol. 27, Iss. 3, p. 385-402, September 2025.
- [14] J. Xu, "Harnessing AI in academic writing. The complex interplay of AI Literacy and self-directed learning among university L2 learners", *System*, Vol. 135, December 2025.
- [15] K. Yang, M. Raković, Z. Liang, Z. Lixiang, Z. Zeng, Y. Fan et al., "Modifying AI, Enhancing Essays: How Active Engagement with Generative AI Boosts Writing Quality", *arXiv*, Nr. 2412.07200., December 2024.
- [16] C. Wang, W. Xu, X. Tan, "Beyond the Human-AI Binaries. Advanced Writers' Self-Directed Use of Generative AI in Academic Writing", *arXiv*, Nr. 2505.12165., May 2025.
- [17] Y. Jin, K. Yang, R. Martinez-Maldonado, D. Gasević, L. Yan, "The Agency Gap: How Generative AI Literacy Shapes Independent Writing after AI support", *arXiv*, Nr. 2507.04398, November 2024.
- [18] P. D. Deep, Y. Chen, "The Role of AI in Academic Writing. Impacts on Writing Skills, Critical Thinking, and Integrity in Higher Education", *Societies*, Vol. 15, Iss. 9, Nr. 247, September 2025.
- [19] K. Bittle, O. El-Gayar, "Generative AI and Academic Integrity in Higher Education. A Systematic Review and Research Agenda", *Information*, Vol. 16, Iss. 4, Nr. 296, April 2025.
- [20] S. Brommer, J. Berendes, U. Bohle-Jurok, I. Buck, K. Girgensohn, E. Grieshammer et al., "Wissenschaftliches Schreiben im Zeitalter von KI gemeinsam verantworten. Eine schreibwissenschaftliche Perspektive auf Implikationen für Akteur*innen an Hochschulen", *Hochschulforum Digitalisierung*, Discussion paper, Iss. 27, November 2023.
- [21] R. Ivanič, *Writing and Identity. The Discoursal Construction of Identity in Academic Writing*. Amsterdam: John Benjamins, 1998.
- [22] K. Hyland, "Authority and Invisibility: Authorial Identity in Academic Writing", *Journal of Pragmatics*, Vol. 34, p. 1091-1112, February 2002.
- [23] O. Kruse, *Lesen und Schreiben. Der richtige Umgang mit Texten im Studium*. 2nd edition. Konstanz/Stuttgart: uvk, 2015.
- [24] C. Bereiter, M. Scardamalia, *The Psychology of Written Composition*. Hillsdale, NJ: Lawrence Erlbaum.
- [25] P.D., Klein, "Knowledge Construction in Collaborative Science Writing: Strategic Simplicity, Distributed Complexity, and Explanatory Sophistication", in *Writing as a Learning Activity*, Vol. 28, P.D. Klein, P. Boscolo, L. Kirkpatrick, C. Gelati, Eds., Leiden: Brill, 2014, pp. 300-326.
- [26] O. Kruse, K. Seifried, M. Ulmi, Eds., *Prozessorientierte Schreibdidaktik. Schreibtraining für Schule, Studium und Beruf*. Köln Haupt, 2006.
- [27] T. Pohl, T. Steinhoff, "Textformen als Lernformen", in *Kölner Beiträge zur Sprachdidaktik*, Vol. A, H. Günther, U. Bredel, M. Becker-Mrotzek, Eds. Duisburg: Gilles & Francke, 2010, pp. 5-26.

- [28] N. Selwyn, *Is Technology Good for Education?*. Polity Press: Cambridge, 2016.
- [29] L. Pangrazio, N. Selwyn, "Towards a school-based 'critical data education'", *Pedagogy, Culture and Society*, Vol. 29, Issue 3, pp. 431-448, 2021.
- [30] K. Mayrberger, *Partizipative Mediendidaktik. Gestaltung der (Hochschul-)Bildung unter den Bedingungen der Digitalisierung*. Weinheim/Basel: Juventa, 2019.
- [31] A. Clark, D. J. Chalmers, "The Extended Mind", *Analysis*, Vol. 58, Issue 1, pp. 7-19, January 1998.
- [32] M. Steig, *Handlungskompetenz. Kompetenzmodelle in der pädagogischen Praxis*. Schotten: STG, 2000.
- [33] D. Carless, D. Boud, "The development of student feedback literacy: enabling uptake of feedback", *Assessment & Evaluation in Higher Education*, Vol. 48, Issue 8, pp. 1315-1325, 2018.
- [34] T. Fawns, D. Liu, M. Cowling, A. Bridgeman, "Responding to Generative AI in Australian Higher Education", working paper, October 2023.
- [35] O. Zawacki-Richter, V. I. Marín, M. Bond, F. Gouverneur, "Systematic review of research on artificial intelligence applications in higher education – where are the educators?", Vol. 16, No. 39, October 2019, pp. 1-27.
- [36] L. Mary, B. V. Street, "Student writing in higher education. An academic literacies approach", *Studies in Higher Education*, Vol. 23, Iss. 2, 1998, pp. 157-172.
- [37] T. Lillis, M. Scott, "Defining academic literacies research. Issues of epistemology, ideology and strategy", *Journal of Applied Linguistics*, Vol. 4, Iss. 1, September 2008, 5-32.
- [38] P. Baumgartner, K. Himpel-Gutermann, "Implementierungsstrategien für E-Portfolios an (österreichischen) Hochschulen", in *Kontrolle und Selbstkontrolle. Zur Ambivalenz von E-Portfolios in Bildungsprozessen*, S. Münte-Goussar, C. Schwalbe, Eds. Wiesbaden: Springer, 2011, pp. 203-223.
- [39] P. Mak, "Portfolio Assessment", in *Research Questions in Language Education and Applied Linguistics. A Reference Guide*, G. Mohebbi, C. Coombe, Eds. Cham: Springer, 2021, pp. 373-377.
- [40] J. Beckers, D. Dolmans, J. van Merriënboer, "e-Portfolios enhancing students' self-directed learning. A systematic review of influencing factors", *Australasian Journal of Educational Technology*, Vol. 32, Iss. 2, June 2016, 32-46.
- [41] T. Häcker, "Portfolios als Instrument der Kompetenzdarstellung und reflexiven Lernprozesssteuerung", *Berufs- und Wirtschaftspädagogik Online*, Iss. 8, 2005, pp. 1-11.
- [42] A. Richter, "Portfolios als alternative Form der Leistungsbewertung", in *Neues Handbuch Hochschullehre. Lehren und Lernen effizient gestalten*, B. Berendt, A. Fleischmann, N. Schaper, B. Szczyrba, J. Wildt, Eds. Berlin: Raabe, 2004, Chapter H 4.2. n.p.
- [43] S. Fischer, S. Jung, K. Fisch, "Portfolio-Prüfungen– die Patentlösung für kompetenzorientiertes Prüfen? Ergebnisse aus der Evaluation von sechs Pilotveranstaltungen", in *Digital gestützte Lehre. Innovative Konzepte zur digitalen und analogen Lehre an Hochschulen*, U. Fahr, P. Riegler, Eds. Wiesbaden, Springer, 2025, pp. 97-123.
- [44] A. Faust, E. Mayweg-Paus, "Empowering Tomorrow's Minds: A Comprehensive Framework for Teaching Non-Technical Students to Think Critically about AI", in *Artificial Intelligence Applications in Higher Education: Theories, Ethics, and Case Studies for Universities*, H. Crompton, D. Burke, Eds. London Routledge, 2024, pp. 73-89.
- [45] P. Eggensberger, I. Kleiber, R. Klöber, S. M. Lorenz, A. Schindel, *Virtuelle Hochschullehre. Ein Handbuch in 50 Fragen und Antworten*. Heidelberg: heiBooks, 2023.
- [46] A. French, "Academic writing as identity-work in higher education: Forming a 'professional writing in higher education habitus'", *Studies in Higher Education*, Vol. 45, Iss. 8, 2020, pp. 1605-1617.
- [47] L. Zhang, J. Wang, "To be like a 'scholar': a study on the construction of authorial identity of Chinese EFL learners in academic writing: an intertextuality perspective", *Frontiers in Psychology*, January 2024, Nr. 14:1297557.
- [48] M. Maguire, A. E. Reynolds, B. Delahunty, "Reading to be. The role of academic reading in emergent academic and professional student identities", *Journal of University Teaching & Learning Practice*, Vol. 17, Iss. 2, 2020, pp. 1-12, citation p. 2.
- [49] M.R. Lea, B. Stierer, "Changing academic identities in changing academic workplaces: learning from academics' everyday professional writing practices", *Teaching in Higher Education*, Vol. 16, Iss. 6, 2011, pp. 605-616.
- [50] S. Marek, "Schreibende Professionals. Typologie einer Zielgruppe für Weiterbildungen im beruflichen Schreiben", *HERMES – Journal of Language and Communication in Business*, Vol. 64, December 2024, pp. 313-330.
- [51] A. Spona, "Forschungsmethodologie zum Wesen und Inhalt der beruflichen Identität der Hochschullehrkräfte", *International Dialogue on Education*, Vol. 4, Iss. 1, February 2017, pp. 181-190.
- [52] T. Miller, R. Roland, V. Vonau, P. Pfeil, Eds., *Professionelle Identitätsentwicklung in der Sozialen Arbeit. Perspektiven auf ein berufsbegleitendes Studium*. Wiesbaden: Springer, 2021.
- [53] H. Fangzhin, "Identity Construction in Academic Writing of Student Writers Who Use English as an Additional Language: A Literature Review", *Chinese Journal of Applied Linguistics*, Vol. 43, Iss. 4, November 2025, n.p.
- [54] P. Dias, A. Freedman, P. Medway, A. Par, *Acting and Writing in Academic and Workplace Contexts*. New York: Routledge, 1999.
- [55] R.A. Varanasi, B.M. Wiesenfeld, O. Nov, "AI Rivalry as a Craft: How Resisting and Embracing Generative AI Reshape Writing Professions", *arXiv*, Nr. 2503.09901, March 2025, n.p.

On Hubness Effects in Retrieval-Augmented Generation Systems

Alexandros Nanopoulos
Baden-Württemberg
Cooperative State University
Mosbach, Germany
alexandros.nanopoulos@mosbach.dhbw.de

Krisztian Buza
Budapest University of
Economics and Business
Budapest, Hungary
buza.krisztian@uni-bge.hu

Margit Antal
Sapientia Hungarian
University of Transylvania
Targu Mures, Romania
manyi@ms.sapientia.ro

Abstract—Retrieval-Augmented Generation (RAG) systems rely on dense vector retrieval to select relevant documents that condition answer generation. While retrieval quality is typically evaluated using aggregate ranking metrics, the geometric properties of high-dimensional embedding spaces are rarely examined. In this paper, we study the impact of hubness—a phenomenon in which certain points appear disproportionately often in nearest-neighbor lists—on dense retrieval within RAG-based question answering systems. Using a standard benchmark dataset for question answering (QA), we empirically show that hub documents frequently occupy top retrieval positions despite being irrelevant, displacing correct documents. We further demonstrate that these hub-induced ranking distortions negatively affect the quality of generated answers, as assessed by an independent language model evaluator. Finally, we introduce a simple hub-aware ranking baseline as an oracle-informed diagnostic scheme, showing that mitigating these effects can yield consistent improvements in both retrieval performance and answer quality. Our findings highlight hubness as an overlooked failure mode in dense retrieval for RAG systems and motivate further investigation of hub-aware retrieval strategies.

Index Terms—Retrieval-Augmented Generation, Embedding-Based Retrieval, Hubness, Nearest-Neighbor Search, Document Ranking

I. INTRODUCTION

Retrieval-Augmented Generation (RAG) has emerged as a dominant paradigm for integrating large language models with external knowledge sources. By retrieving a small set of relevant documents and conditioning generation on them, RAG systems aim to improve factual accuracy and reduce hallucinations in generative models [1]. As a result, dense vector retrieval based on semantic embeddings has become a core component of many modern question answering and knowledge-intensive generation pipelines [2].

Most existing evaluations of RAG systems focus on end-to-end performance, typically measured through answer correctness or generation quality. Retrieval quality is often summarized using aggregate metrics such as Recall@k or Mean Reciprocal Rank (MRR), while the internal behavior of nearest-neighbor retrieval in high-dimensional embedding spaces is largely treated as a black box. As a result, it remains unclear how structural characteristics of the embedding space—such as the relative distribution of document representations and their concentration around certain regions—may systematically in-

fluence retrieval rankings and, in turn, downstream generation quality.

One concrete manifestation of such structural effects is *hubness*, a phenomenon that emerges in high-dimensional embedding spaces and reflects systematic irregularities in nearest-neighbor relationships. As dimensionality increases, distances between points tend to concentrate, and small differences in relative position can cause certain representations to lie closer to many others. As a result, a limited number of points—referred to as *hubs*—appear disproportionately often in the k -nearest-neighbor lists of other points [3]. Hubness has been shown to distort nearest-neighbor structure and to negatively affect performance in a range of kNN-based learning and retrieval tasks [4], [5]. Despite the widespread use of dense embeddings and nearest-neighbor search in RAG systems, the implications of hubness for retrieval rankings and downstream generative performance have not yet been systematically examined.

In this paper, we present an exploratory analysis of hubness effects in a RAG-based question answering (QA) setting. We use a standard QA benchmark dataset with known ground-truth documents for the corresponding answers [6]. This allows us to analyze the distribution of document occurrences in k -nearest-neighbor retrieval and identify the presence of hubs in the embedding space. We further distinguish between hubs and *precedence* hubs, where the latter are documents that are frequently ranked in top retrieval positions before the correct documents, despite being irrelevant to the query.

Our empirical findings show that precedence hubs can displace relevant documents from top-ranked retrieval positions, leading to degraded retrieval. More importantly, we demonstrate that these hub-induced ranking distortions propagate downstream, resulting in lower-quality answers generated by a RAG pipeline, as assessed by an independent large language model evaluator. To illustrate the practical relevance of this effect, we introduce a simple hub-aware ranking heuristic that penalizes precedence hubs and observe consistent improvements in both retrieval and generation quality.

The contributions of this paper are threefold:

- We present a systematic empirical analysis of hubness in document embedding spaces used for retrieval-augmented

question answering, revealing its prevalence and structural characteristics.

- We demonstrate that precedence hubs induce systematic distortions in retrieval rankings, leading to measurable degradations in retrieval effectiveness reflected in answer generation quality.
- Using an oracle-informed re-ranking strategy, we further show that mitigating these effects can improve both retrieval and generation performance, highlighting hubness as a diagnosable and actionable source of error.

Our results highlight hubness as an overlooked failure mode in dense retrieval for RAG systems and suggest that hub-aware diagnostics and mitigation strategies deserve further attention in the design and evaluation of retrieval-augmented generation pipelines.

The remainder of the paper is organized as follows. Section II reviews relevant background and related work. Section III describes the experimental setting and evaluation setup. Section IV analyzes hubness and rank-sensitive retrieval biases in retrieval-augmented generation systems. Section V presents an empirical evaluation of their impact on retrieval performance and answer generation quality. Section VI discusses and interprets the experimental findings. Finally, Section VII concludes the paper and outlines directions for future work.

II. BACKGROUND AND RELATED WORK

A. Retrieval-Augmented Generation

Retrieval-Augmented Generation (RAG) integrates a retrieval component with a generative language model to ground generation in external sources, improving factuality and coverage for knowledge-intensive tasks [1], [2], [7]. In typical RAG pipelines a retriever (often a dense vector index) returns a small set of passages or documents that are used as additional context for a sequence-to-sequence or decoder-only model; this architecture has become a standard approach for open-domain question answering, summarization, and knowledge-grounded dialog.

Recent research summarize the rapid development of RAG and related Retrieval-Augmented Large Language Model (RALLM) systems, as well as the open challenges they face [8]–[10]. These challenges include retrieval quality, evaluation gaps, scalability, robustness, grounding and hallucination, and the methodology used to assess end-to-end system behavior. A common thread across these works is that modern RAG pipelines rely predominantly on dense, embedding-based retrieval; that evaluation is typically conducted using aggregate retrieval metrics (such as Recall@k and mean reciprocal rank) or overall answer correctness; and that detailed diagnostic analyses of internal retrieval failure modes remain comparatively underexplored.

In dense retrieval, queries and documents are represented as continuous vector embeddings and nearest-neighbor search is used to select candidate documents. Dense retrieval has been the focus of extensive recent work aimed at improving

encoder architectures, training objectives, and hybrid retrieval strategies [2], [10]. Despite this progress, retrieval is typically treated as a modular component and evaluated primarily through aggregate ranking metrics. As a result, structural characteristics of embedding-based retrieval—such as uneven neighborhood relationships and the tendency of certain documents to appear disproportionately often across retrieval results—are rarely examined as explicit sources of systematic retrieval errors that may propagate to downstream generation.

In summary, while the RAG literature offers a growing body of engineering advances and evaluation practices, comparatively little attention has been devoted to diagnostic analyses that connect internal retrieval behaviors to observable degradation in end-to-end generation quality. In particular, the impact of structural characteristics of embedding-based retrieval on downstream generation remains insufficiently understood. This work addresses this gap by examining hubness as a specific retrieval-side phenomenon and by analyzing how it affects both retrieval rankings and the quality of generated answers.

B. Hubness and High-Dimensional Retrieval

Hubness is a phenomenon observed in high-dimensional spaces, where certain points—referred to as *hubs*—appear disproportionately often in the k -nearest-neighbor lists of other points due to the concentration of distances [3], [5]. Hubness is typically quantified through k -occurrence, i.e., the number of times a point is among the k nearest neighbors of other points.

Prior work has shown that hubness can distort neighborhood structure and negatively affect a variety of tasks that rely on nearest-neighbor relationships, including classification, clustering, regression, and retrieval. A comprehensive empirical comparison of hubness reduction methods has been conducted in the context of nearest-neighbor learning, evaluating multiple unsupervised approaches across diverse data sets to improve neighborhood symmetry and reduce hub occurrence [4]. In natural language semantic embedding spaces, hubness has been shown to degrade nearest-neighbor accuracy, and targeted hubness reduction techniques have been demonstrated to improve semantic representations, such as in Sentence-BERT embeddings where hubness reduction can reduce hub frequency and improve kNN classification performance [11].

In the information retrieval literature, hubness has also been addressed in cross-modal and multimodal retrieval settings. For example, adversarial hubness work investigates how embedding spaces can be manipulated to create universal retrieval hubs that adversely affect multi-modal retrieval outcomes [12]. Other approaches, such as hub-aware training objectives, have been proposed to balance hub centrality and improve retrieval quality in cross-modal embeddings [13]. Hubness mitigation losses have also been studied in text–image matching tasks to improve retrieval robustness [14].

Most of this literature focuses on either the statistical properties of hubness in high-dimensional spaces or on improving retrieval and embedding quality in specific tasks (e.g., semantic search or multimodal matching). However, hubness

effects in dense retrieval components of RAG pipelines—and especially their impact on downstream generative answer quality—remain largely unexplored.

C. Re-ranking Methods in RAG

In RAG systems, document re-ranking plays a key role in prioritizing the most relevant retrieved documents, thereby improving answer quality and reducing hallucinations. Existing re-ranking approaches are commonly categorized as point-wise, pairwise, and list-wise. Point-wise methods (e.g., MonoT5 [15]) score documents independently and are computationally efficient but ignore inter-document relationships. Pairwise methods (e.g., PRP [16], EcoRank [17]) compare document pairs to improve relative ranking at increased computational cost. List-wise approaches (e.g., RankGPT [18], ListT5 [19]) jointly consider multiple documents, capturing richer interactions and often achieving higher effectiveness, albeit with greater resource requirements. Recent work [20] shows that while LLM-based re-rankers can achieve strong performance, particularly in list-wise settings, their generalization to unseen queries remains inconsistent and often comes with significant computational overhead.

D. Scope and Relation to Prior Work

Prior work has shown that hubness in high-dimensional embedding spaces can distort nearest-neighbor retrieval. However, its impact on retrieval-augmented generation (RAG) systems—particularly on document ranking and downstream answer generation—remains largely unexplored. This work addresses this gap by empirically analyzing hubness in dense retrieval and its association with retrieval rankings and generated answer quality in a RAG setting. Moreover, in contrast to the aforementioned re-ranking approaches, which aim to optimize relevance estimation, our work focuses on diagnosing structural biases in retrieval—specifically hubness—and analyzing how such biases affect ranking behavior and downstream generation quality. Notably, such structural biases may persist even after the application of re-ranking methods, suggesting that their interaction with re-ranking strategies warrants further investigation.

III. EXPERIMENTAL SETTING

This section describes the experimental setup used to analyze hubness effects in dense retrieval for RAG systems. The overall pipeline is intentionally kept simple to isolate the influence of retrieval behavior on downstream generation and evaluation. Figure 1 provides an overview of the complete process, from question embedding and document retrieval to answer generation and evaluation, and also illustrates the diagnostic re-ranking employed later in the analysis.

A. Dataset

We conduct our experiments using the Stanford Question Answering Dataset (SQuAD), a widely used benchmark for machine reading comprehension that provides question-answer pairs grounded in reference documents.

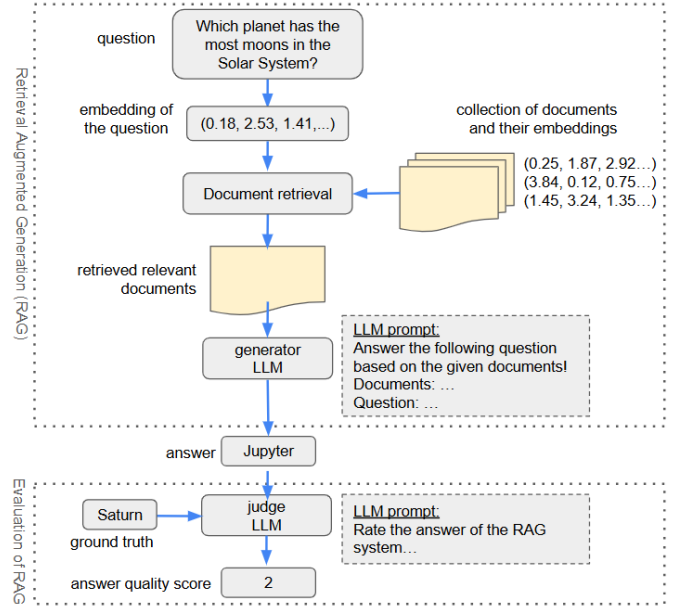


Fig. 1. Overview of the experimental RAG pipeline. A question is embedded and used to retrieve the top- k most similar documents based on cosine similarity. The retrieved documents condition a generator LLM, whose output is subsequently evaluated by an independent judgment LLM against the ground-truth answer.

In particular, we use the SQuAD 2.0 dataset¹ as a benchmark for question answering evaluation. Specifically, we work with the development (dev) split of SQuAD 2.0 and restrict our experiments to answerable questions only, excluding unanswerable ones introduced in this version of the dataset. The resulting subset consists of 1,204 contexts and 5,928 questions, providing a well-established and carefully curated evaluation setting [6], [21].

In SQuAD 2.0, long documents are segmented into passages following standard preprocessing practices to ensure manageable input lengths for embedding and retrieval. For simplicity, we will refer to these passages as documents from now on. Each question is associated with the document that contains the correct answer which allows us to determine whether the correct document appears among the top retrieved results and at which rank position. This property is central to our analysis of how retrieval rankings are affected by hubness.

B. Embedding and Retrieval

To enable embedding-based retrieval in RAG systems, both questions and documents are mapped into a shared continuous embedding space using a fixed neural encoder. In this paper, we use E5-BASE², which is a dense-text embedding model from Microsoft designed for retrieval and semantic similarity tasks. It produces 768-dimensional vector embeddings and supports input sequences of up to 512 tokens, reflecting its BERT-base architecture. The model is trained using weakly supervised contrastive learning on large-scale text pair datasets,

¹<https://rajpurkar.github.io/SQuAD-explorer/>

²<https://huggingface.co/intfloat/e5-base>

where queries and passages are prefixed with instructions like “query:” and “passage:” to unify different embedding tasks. This training approach enables E5-BASE to generate high-quality, task-agnostic embeddings that work well for search, clustering, and matching across domains [22].

The embedding is applied without task-specific fine-tuning to avoid confounding effects from retrieval optimization. Document retrieval is performed using cosine similarity in the embedding space. For each question, we retrieve the top $k = 5$ nearest documents, reflecting common practice in RAG systems with limited context windows.

The effectiveness of RAG critically depends on the characteristics of the underlying embedding space. Choices related to embedding dimensionality, training, and similarity computation are therefore expected to influence both retrieval performance and the quality of the generated answers. In the following sections, we will empirically assess these effects, with particular attention to systematic retrieval patterns that emerge from high-dimensional embeddings and their downstream impact on answer generation.

C. RAG Answer Generation

We next describe the RAG answer generation setup, which uses the top- k documents retrieved based on embedding similarity. These retrieved documents are provided as contextual input to a generative language model, denoted as *generator LLM*. For this purpose, we used Llama-3.2-3B-Instruct³ as generator LLM, which is a compact large language model based on the LLaMA architecture, optimized here for efficient local inference via Ollama. It supports an exceptionally long context length of up to 131,072 tokens and uses a 3,072-dimensional embedding space, enabling it to handle very large retrieved context windows effectively. The model is quantized using Q4_K_M, which significantly reduces memory usage while preserving strong generation quality. It supports text completion and tool use, and with an inference temperature of 0.7, it balances coherence and variability to produce concise, controlled answers when constrained by retrieved context [23].

The generator receives a prompt that includes the question and the retrieved documents, instructing it to generate an answer grounded in the provided context, as illustrated in Figure 1.

D. Answer Evaluation

We evaluate the generated answers using an independent language model acting as a judge, referred to as the *judge LLM*. Qwen2.5-7B-Instruct⁴ was employed as an LLM-as-a-judge to assess answer quality in a controlled and deterministic manner [24]. Built on the Qwen2 architecture and deployed via Ollama, the model supports a 32,768-token context window and uses a 3,584-dimensional embedding space, allowing it to reliably compare questions, ground truth answers, and model-generated responses within a single prompt. The model was quantized with Q4_K_M to improve inference efficiency

while maintaining evaluation fidelity. Using a temperature of 0 ensured fully deterministic judgments, which is essential for reproducible scoring. Guided by a strict evaluation prompt, this model served as an impartial evaluator rather than a generative assistant. We refer to the Appendix for the particular prompt used to evaluate the generated answers.

The judge LLM assigns a score on a discrete scale from 1 (poor) to 5 (excellent) based on the semantic correctness and completeness of the generated answer with respect to the ground-truth answer. This LLM-based evaluation enables scalable assessment of answer quality beyond exact match metrics.

We note that the evaluation scores depend on the behavior and calibration of the evaluator model. However, since the evaluator is held fixed across all experimental conditions, relative differences in scores can be attributed to changes in retrieval and generation behavior rather than evaluator variability.

IV. HUBNESS AND RANKING BIASES IN RAG

This section analyzes the emergence of hubness in embedding-based retrieval for RAG systems and introduces rank-sensitive variants of hubness, which are used in the following section to investigate their relationship with retrieval biases, document ranking behavior, and downstream answer generation. It also explains the diagnostic re-ranking that we utilize later in the evaluation.

A. Hubness

Embedding-based retrieval operates in high-dimensional representation spaces, which are known to exhibit geometric effects that do not arise in low-dimensional settings. One such effect is *hubness*, a phenomenon in which a small number of data points appear disproportionately often as nearest neighbors of many other points. As dimensionality increases, distances tend to concentrate, and certain points become systematically closer to many queries, even when they are not semantically relevant. This behavior has been documented across a wide range of nearest-neighbor retrieval tasks in high-dimensional spaces and is known to introduce systematic biases in retrieval-based systems [3]–[5], [11]. These effects are also observed under cosine similarity, which has been shown to exacerbate hub formation [25].

In the context of RAG, hubness is particularly relevant, as frequently retrieved documents repeatedly condition answer generation. To quantify this effect, based on the definition in [3], we define the *hubness score* of a document d , denoted $H(d)$, as the number of times d appears among the k -nearest neighbors of all queries. Since our focus is on the potentially harmful impact of hubs, we exclude from $H(d)$ those cases where d corresponds to the correct document for the query. This exclusion ensures that high hubness scores reflect documents that are frequently retrieved despite being irrelevant to the underlying question, rather than documents that are legitimately retrieved due to correctness.

³<https://huggingface.co/meta-llama/Llama-3.2-3B-Instruct>

⁴<https://huggingface.co/Qwen/Qwen2.5-7B-Instruct>

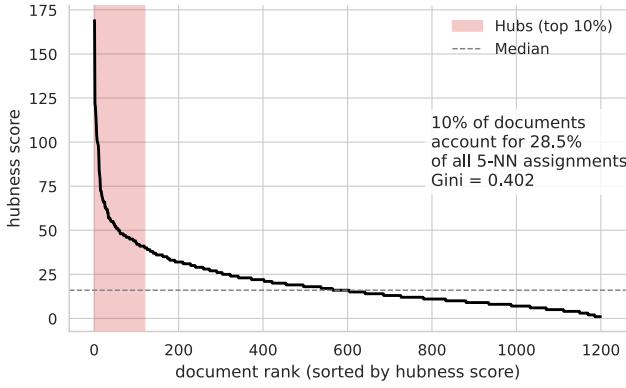


Fig. 2. Ranked distribution of hubness scores $H(d)$, defined as the number of times a document appears among the $k = 5$ nearest neighbors across all queries, excluding cases where the document corresponds to the correct answer. The distribution exhibits a sharp initial drop followed by a long tail, indicating that a small subset of documents is retrieved disproportionately often. The shaded region highlights the top 10% of documents, whose hubness scores substantially exceed the median level observed across the corpus (dashed line), motivating their designation as hubs.

Analyzing the distribution of $H(d)$ therefore provides insight into how often a small subset of documents systematically enters the retrieval sets used by RAG, potentially displacing more relevant context. Figure 2 shows the ranked hubness distribution under this definition. The distribution exhibits a sharp initial drop followed by a long tail, indicating that a limited fraction of documents is retrieved far more frequently than the rest. This skewness is further quantified by a Gini coefficient of 0.40, which reflects a substantial—though not extreme—inequality in how nearest-neighbor assignments are distributed across documents. In particular, the top 10% of documents account for a substantial share of all nearest-neighbor assignments, markedly exceeding the median hubness level observed across the corpus. Based on these observations, we designate the top 10% of documents as *hubs*.

B. Rank-Sensitive Hubness

Existing work on hubness has predominantly focused on *binary relevance* information, that is, whether a retrieved document is correct or incorrect with respect to a given query. Under this perspective, hubness measures such as $H(d)$ capture how frequently a document incorrectly appears among nearest-neighbor retrieval results, without accounting for its position relative to the correct document. While this view is sufficient for analyzing retrieval bias in isolation, it is incomplete in the context of RAG. Generator models typically condition on retrieved documents in ranked order or implicitly prioritize higher-ranked entries, such that incorrect documents appearing before the correct one can disproportionately influence the generated response [1], [26].

To capture this effect, we introduce a rank-sensitive variant of hubness, termed *precedence hubness* and denoted by $P(d)$. For a document d , $P(d)$ counts the number of queries for

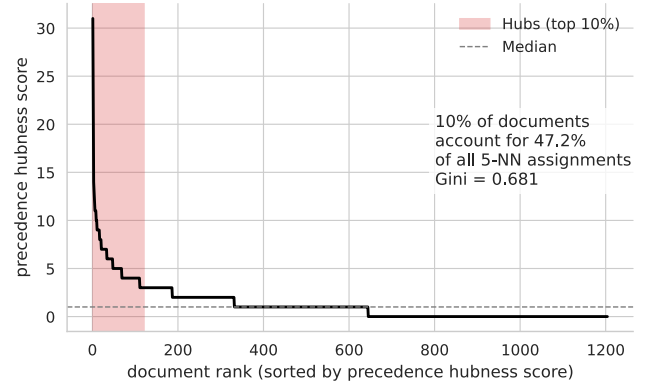


Fig. 3. Ranked distribution of precedence hubness scores, defined as the number of times a document appears at a higher rank than the correct document within the top- $k = 5$ retrieved results, excluding queries for which the correct document is not retrieved. The distribution is highly concentrated, with the top 10% of documents accounting for 47% of all precedence events. The shaded region highlights the designated hubs, illustrating that relative misranking is dominated by a small subset of documents.

which d is ranked ahead of the correct document within the top- k retrieved list. Unlike $H(d)$, which reflects the frequency with which a document is retrieved, $P(d)$ captures its tendency to appear ahead of the correct document in the ranking. Thus, this formulation allows us to distinguish documents that appear merely frequently from those that systematically outrank the correct document (i.e., relevant context), providing a more precise characterization of retrieval-induced biases that affect downstream answer generation. To isolate ranking effects, we exclude queries in which the correct document does not appear among the top- k retrieved results, as relative ordering cannot be assessed in such cases; empirically, we observe that this occurs only infrequently in our setting.

Analogously to $H(d)$, we analyze the distribution of precedence hubness, $P(d)$. Figure 3 shows that the ranked distribution is even more concentrated than that of standard hubness, with a small subset of documents repeatedly preceding the correct document across many queries. In particular, the top 10% of documents account for 47% of all precedence events, indicating that relative misranking is dominated by a limited fraction of documents. This pronounced concentration is further quantified by a Gini coefficient of 0.68, which is substantially higher than that observed for $H(d)$. Taken together, these results indicate that while many documents may occasionally precede the correct context, persistent rank dominance is driven by a much smaller set of documents, which are therefore likely to exert a disproportionate influence on downstream answer generation.

It is informative to examine whether documents that act as strong precedence hubs also tend to be hubs. Figure 4 shows the relationship between precedence hubness $P(d)$ and standard hubness $H(d)$. A clear positive association is observed (Pearson $r = 0.66$), indicating that documents which appear frequently among retrieved neighbors also tend, on average, to precede the correct document more often. This

correlation reflects a shared structural origin: documents that occupy dense or central regions of the embedding space are both retrieved across many queries and, when retrieved, are more likely to appear at higher ranks.

At the same time, the overlap between the two top-10% sets is only partial, with their intersection comprising 6.7% of all documents. This incomplete overlap highlights that the two measures capture related but distinct aspects of retrieval behavior. While $H(d)$ reflects overall retrieval frequency, $P(d)$ isolates systematic rank dominance with respect to the correct document. Consequently, some documents that are frequent neighbors do not consistently outrank the correct context, whereas other documents with more moderate overall hubness nonetheless dominate early ranks. This distinction motivates our emphasis on $P(d)$ in the subsequent analysis, as it more directly captures ranking effects that are critical for understanding downstream answer generation quality.

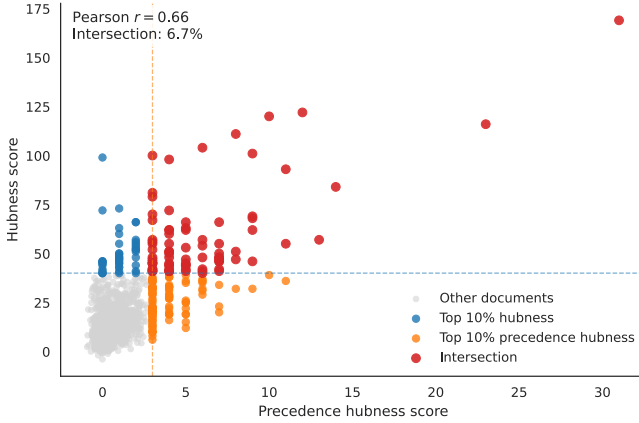


Fig. 4. Relationship between standard hubness $H(d)$ and precedence hubness $P(d)$. Each point corresponds to a document, with dashed lines indicating the thresholds defining the top 10% under each measure. A clear positive correlation is observed (Pearson $r = 0.66$), indicating that frequently retrieved documents also tend to precede the correct document more often. At the same time, the limited overlap between the two top-10% sets highlights that $H(d)$ and $P(d)$ capture distinct aspects of retrieval behavior: overall retrieval frequency versus relative rank dominance.

C. A Diagnostic Hub-Aware Re-Ranking Intervention

The analysis above establishes that a small subset of documents exhibits disproportionately high precedence hubness, systematically appearing ahead of the correct document in retrieval rankings. To better isolate the causal role of such documents in retrieval-induced answer degradation, we introduce a simple *hub-aware re-ranking intervention*. Importantly, this procedure is not intended as a deployable retrieval method nor as a competitive re-ranking algorithm. Instead, it serves as a controlled diagnostic tool that allows us to probe how demoting documents with strong precedence hub characteristics affects retrieval order and downstream answer generation.

The intervention relies on lightweight document-level statistics computed offline over the evaluation corpus. Specifically, for each document d , we define a *correctness score* $C(d)$ as

the number of queries for which d is the correct document containing the ground-truth answer. This quantity captures how often a document has acted as relevant evidence in the retrieval process. In contrast, the precedence hubness score $P(d)$, defined in the previous section, captures how often the same document precedes the correct document when both appear among the top- k retrieved results. Thus, a combination of both ranking signals $P(d)$ and $C(d)$ can provide a coarse indicator of whether a document tends to support or distort retrieval rankings across queries.

Based on these quantities, we associate each document d with the score

$$S(d) = \frac{C(d)}{C(d) + P(d)}, \quad (1)$$

which can be interpreted as a crude reliability ratio. Documents that frequently serve as correct retrieval targets but rarely outrank the correct document receive higher scores, whereas documents that often precede the correct context despite low correctness receive lower scores. When both $C(d)$ and $P(d)$ are zero, we set $S(d) = 0$.

For each query q , we first retrieve the top- k documents using cosine similarity, with $k = 5$ as in the standard RAG setup. The diagnostic re-ranking is then applied locally to this list through a conservative pairwise swapping procedure. Specifically, for each pair of consecutive documents (d_i, d_{i+1}) in the ranked list, their positions are swapped if both of the following conditions hold:

- $S(d_{i+1}) > 2 \cdot S(d_i)$, indicating a substantially higher reliability score; and
- the cosine similarities of d_i and d_{i+1} to the query differ by less than 10%, ensuring that the swap does not override a strong similarity-based preference.

This conservative design intentionally limits the extent to which similarity-based retrieval is altered. Documents are promoted only when there is strong evidence—based on corpus-level statistics—that their presence ahead of other documents is more likely to reflect hub-induced ranking dominance than genuine relevance. The specific numerical thresholds used in this procedure are not optimized and should not be interpreted as principled hyperparameter choices. Rather, they are chosen to enforce minimal and interpretable ranking perturbations that preserve the overall structure of the original retrieval results.

We emphasize that this intervention uses corpus-level statistics and therefore cannot be applied in a strictly online or zero-shot setting. Its purpose is solely to support the empirical analysis in the following section by enabling a controlled comparison between original rankings and rankings in which documents exhibiting strong precedence hubness are selectively demoted. The observed changes in retrieval order and answer quality should therefore be interpreted as evidence for the *existence and magnitude* of hub-induced effects, rather than as validation of a specific re-ranking strategy. The design of general, online, and learning-based hubness mitigation methods for RAG systems is left for future work.

V. EXPERIMENTAL EVALUATION

In this section, we evaluate the impact of precedence hubness on both retrieval effectiveness and downstream answer quality in the examined RAG setting. Our analysis quantifies how hub-induced ranking distortions affect retrieval behavior and how these distortions propagate to answer generation quality. First, in Section V-A, assess the overall performance of the retrieval utilized for RAG by measuring how frequently the correct document is being included among the k retrieved documents for each query. We then, in Section V-B, examine the role of the ranking position that the correct document obtains among the k retrieved documents on the quality of the generation. These results support an overall evaluation of the performance of the RAG system and of the role of the ranking. We then move on to examine, in Section V-C the impact of hubs on the retrieval (i.e., ranking of the correct document) and consequently on the quality of the generated text.

A. Retrieval Performance Overview

As an initial validation of the experimental setup, we evaluate the embedding-based retriever on the SQUAD dataset using standard ranking metrics. Table I reports the mean reciprocal rank (MRR) and Recall@ k for $k = 1$ to 5.

TABLE I
RETRIEVAL PERFORMANCE OF THE EMBEDDING MODEL ON THE SQUAD DATASET.

Metric	Score
MRR	0.87
Recall@1	0.80
Recall@2	0.90
Recall@3	0.93
Recall@4	0.95
Recall@5	0.96

Overall, the retriever achieves strong performance, with an MRR of 0.87 and a Recall@1 of 0.80, indicating that the correct document is ranked first for the majority of queries. Recall increases steadily with k , reaching 0.96 at $k = 5$, which suggests that in most cases the correct document appears within the top few retrieved candidates. This support the aforementioned generally strong RAG performance. At the same time, however, a non-negligible fraction of queries require more than one retrieval step to surface the correct document, motivating a closer examination of ranking effects.

B. Effect of Retrieval Rank on Answer Quality

We next examine how the relative rank of the correct document affects answer generation quality, a factor also highlighted in prior work [1], [26]. To structure the analysis, we restrict attention to queries for which the correct document appears among the top five retrieved results—excluding the 4% of queries for which it is not retrieved—and distinguish between two retrieval regimes:

- **Preferential ranking:** the correct document is ranked within the top three positions.
- **Non-preferential ranking:** the correct document is ranked fourth or fifth.

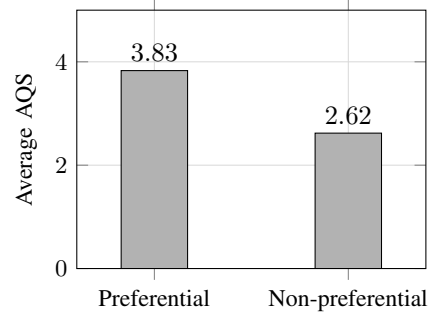


Fig. 5. Comparison of average answer quality score (AQS) between queries with preferential and non-preferential ranking of the correct document.

Under the aforementioned restriction, queries with preferential ranking achieve substantially higher average answer quality scores (AQS) than queries with non-preferential ranking, as illustrated in Figure 5. A Mann-Whitney U test confirms that this difference in means is statistically significant at the 0.01 level. These results demonstrate that answer generation quality is highly sensitive to whether the correct document appears early in the retrieval list.

C. Impact of Precedence Hubness

To assess the role of precedence hubness, we measure how frequently top precedence hubs (top 10% according to $P(d)$) appear in the top-5 retrieval lists under the two ranking regimes. As shown in Table II, at least one precedence hub (i.e., among the top-10%) appears in 74% of preferentially ranked cases, compared to 90% of non-preferentially ranked cases. This difference indicates that precedence hubs play an important role in pushing the correct document down the retrieval list, which causes reduction in answer quality.

TABLE II
PRESENCE OF TOP PRECEDENCE HUBS IN TOP-5 RETRIEVAL LISTS UNDER DIFFERENT RANKING REGIMES.

Ranking regime	Percentage of queries with at least one precedence hub
Preferential ranking	74%
Non-preferential ranking	90%

To more precisely isolate the impact of precedence hubs, we focus on the subset of non-preferentially ranked queries for which removing all top precedence hubs from the retrieval list would move the correct document into the top three positions, thereby converting them into preferentially ranked cases. We refer to these queries as *hub-attributable*. For this subset, the degraded ranking—and the resulting generated answers—can be attributed to the presence of precedence hubs. The remaining non-preferential cases are influenced by

additional retrieval errors and therefore cannot be attributed solely to hubness.

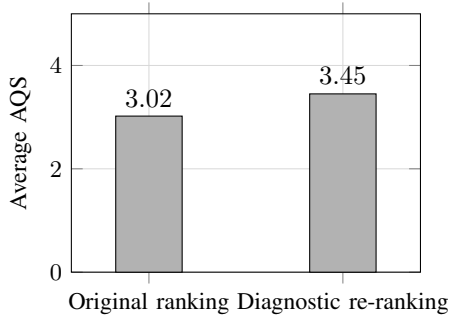


Fig. 6. Average answer quality score (AQS) for *hub-attributable* queries before and after applying the diagnostic re-ranking baseline.

Under this criterion, approximately 35% of non-preferentially ranked queries can be directly attributed to precedence hub effects. While hub-related phenomena may influence a broader portion of non-preferential cases, the subset of *hub-attributable* queries provides a clean and interpretable basis for isolating the direct impact of precedence hubs. For these queries, the average answer quality score (AQS) is 3.02, as shown in Figure 6. This value exceeds the average AQS observed across all non-preferentially ranked queries (2.62) because the latter also include cases in which the correct document is pushed further down the ranking by retrieval errors beyond precedence hub effects, leading to additional degradation in answer quality.

We next apply the diagnostic hub-aware re-ranking described in Section IV-C and regenerate answers for the *hub-attributable* queries using the adjusted rankings. This leads to an increased average AQS of 3.45 (see Figure 6). The improvement is statistically significant according to a Mann-Whitney U test at the 0.05 level. This gain provides an empirical estimate of the answer quality degradation attributable to precedence hubs and underscores potential benefits of hub-aware re-ranking strategies, which we leave for future work.

Finally, to support a qualitative interpretation of the observed differences in AQS, Table III presents illustrative examples of answers for some *hub-attributable* queries before and after applying the diagnostic re-ranking. These examples shed light on the types of retrieval-induced errors underlying the quantitative results. In particular, the observed improvements in AQS are primarily associated with a reduction in factual errors (as illustrated by the first two queries) and with fewer cases in which no concrete answer is produced (as illustrated by the last two queries).

D. Supplementary Results Across Models and Datasets

To assess the robustness of our findings, we conducted additional experiments with multiple embedding models, including BGE-LARGE⁵, E5-BASE⁶, and XLM-RoBERTa⁷.

While BGE-M3 achieved the highest retrieval performance in terms of MRR and is therefore used for the main analysis, we observed comparable hubness-related effects across all examined embedding models, indicating that the phenomenon is not model-specific.

In addition to SQuAD, we evaluated our approach on the LiveRAG benchmark dataset⁸, which is based on the FineWeb-10BT corpus—a large-scale web-derived collection of approximately 10 billion tokens. Due to space constraints, we report only a summary of the results. Consistent with the findings on SQuAD, the diagnostic re-ranking improves retrieval effectiveness, increasing Recall@3 from 95.6% to 97.5%. This result supports the generality of our conclusions across datasets, as improving the likelihood of retrieving the correct document within the top-3 positions has been shown to positively impact the quality of the generated text.

We emphasize that these additional experiments are intended as indicative validation rather than a comprehensive benchmark, and leave a broader evaluation across datasets, retrievers, and generator models to future work.

VI. DISCUSSION

The empirical findings presented above should be interpreted in their proper context. Overall, the RAG pipeline exhibits strong performance for the majority of queries, in line with prior observations [1]. Hub-induced effects do not dominate system behavior, nor do we suggest that hubness is the sole source of retrieval or generation errors.

At the same time, our analysis highlights an important asymmetry in the impact of retrieval failures. While the fraction of queries affected by precedence hubs is relatively limited, these cases are characterized by systematically degraded rankings and noticeably lower answer quality. Such failures can have a disproportionate effect on user perception: even infrequent but salient errors—particularly those resulting in confident yet incorrect or misleading answers—may undermine trust and acceptance of the system, regardless of strong average performance. From this perspective, hub-induced errors warrant attention not because they are ubiquitous, but because their consequences are unevenly distributed.

This asymmetry becomes particularly relevant when considering realistic deployment settings. Although the dataset examined in this study is relatively modest in size, RAG systems in practice are typically applied to much larger corpora and serve a substantially higher volume of queries. In such settings, even a small fraction of hub-attributable failures can translate into a large absolute number of affected queries, exposing a significant number of users to incorrect or misleading outputs. Consequently, the practical impact of hub-induced errors may be more pronounced than suggested by relative frequencies alone.

It is also important to contextualize these findings with respect to the underlying models. Our experiments rely on open-source language models, which inevitably exhibit limitations

⁵<https://huggingface.co/BAAI/bge-large-en-v1.5/tree/main>

⁶<https://huggingface.co/intfloat/e5-base>

⁷<https://huggingface.co/sentence-transformers/paraphrase-MiniLM-L6-v2>

⁸<https://huggingface.co/datasets/LiveRAG/Benchmark>

TABLE III
QUALITATIVE EXAMPLES FOR SOME *hub-attributable* QUERIES.

Query	Answer (Original Ranking)	Answer (Diagnostic Re-ranking)	Ground Truth
<i>What southern city did the Huguenots settle near?</i>	<i>New York.</i>	<i>Charleston, South Carolina.</i>	<i>Charleston, South Carolina.</i>
<i>Which genus is considered the "aunt" of ctenophores?</i>	<i>The genus considered the "aunt" of ctenophores is Vendobionta.</i>	<i>The genus Stromatoveris is considered the "aunt" of ctenophores.</i>	<i>Stromatoveris</i>
<i>Where did scientists find their Y. pestis sample?</i>	<i>There is no mention of a Y. pestis sample being found at any location. The retrieved contexts discuss the study and investigation of Y. pestis in relation to the Black Death pandemic, but do not provide information on where scientists found their Y. pestis samples.</i>	<i>They found their Y. pestis sample from Black Death victims in the East Smithfield burial site in England.</i>	<i>from the tooth sockets in human skeletons</i>
<i>What was the name of the imperialistic policy in China?</i>	<i>No specific name is mentioned in the retrieved contexts.</i>	<i>The "Open Door Policy" was an example of an imperialistic policy in China during the Scramble for Africa and "The Great Game" in Persian lands.</i>	<i>Open Door Policy</i>

in reasoning, robustness, and calibration, and may therefore influence absolute answer quality. At the same time, the use of open-source models enables systematic investigation under limited computational resources and facilitates reproducibility and transparency of the experimental setup. Importantly, the hubness effects observed in this study arise from the structure of the embedding space rather than from properties of a specific generator. As such, they should be understood as an intrinsic characteristic of high-dimensional embedding-based retrieval and are therefore expected to manifest in systems built on commercial or proprietary models as well.

Finally, our empirical evaluation focuses on a question answering dataset, which allows for relatively objective assessment of correctness and answer quality. While this choice facilitates controlled analysis, hubness is not specific to QA tasks. The core findings regarding the concentration of retrieval influence, rank dominance, and their downstream effects are expected to generalize to other RAG scenarios, such as document summarization, report generation, or decision support, where retrieved context similarly conditions generation. Extending the analysis to broader document types and task settings therefore represents an important direction for future work.

Taken together, these observations suggest that hubness-aware analysis should not be viewed as a critique of RAG effectiveness, but rather as a complementary lens for understanding and mitigating a specific class of retrieval-induced failures. Addressing such effects, even partially, may yield meaningful improvements in robustness, interpretability, and user trust.

VII. CONCLUSION

In this paper, we presented an exploratory analysis of hubness effects in dense retrieval for RAG systems. Our study demonstrates that hubness, a well-known phenomenon in high-dimensional spaces, manifests in document embedding spaces used for RAG and can systematically distort retrieval rankings.

We showed that such distortions disproportionately affect early-ranked documents, which are known to have a decisive influence on answer generation quality. As a result, hub-induced ranking displacement leads to measurable degradation in generated answers for a non-negligible subset of queries.

Importantly, our findings do not suggest that RAG systems are fundamentally flawed. On the contrary, RAG performs effectively in the majority of cases. However, our analysis indicates that hubness provides a principled geometric explanation for part of the remaining gap between retrieval effectiveness and generation quality. Given the asymmetric impact that even a small number of suboptimal outputs can have on user trust and system adoption, understanding and addressing such failure modes is practically relevant.

The primary contribution of this work lies in identifying and empirically characterizing hubness as an overlooked diagnostic factor in RAG pipelines, rather than in proposing optimized mitigation techniques. While we demonstrated that even a simple hub-aware adjustment can yield consistent improvements, the development of robust and general hubness mitigation strategies for RAG remains an open research direction. Promising avenues for future work include systematic evaluation of existing hubness reduction methods in retrieval settings [4], extending recent analyses of adversarial and multimodal hubness to RAG pipelines [12], and investigating hubness-aware learning objectives in other embedding-based systems such as knowledge graph embeddings.

Future studies should also consider broader datasets, alternative retrievers, hybrid retrieval architectures, and human-centered evaluation to further assess the generality of these findings. Overall, we argue that hubness-aware analysis offers a valuable lens for understanding retrieval behavior in modern RAG systems and can inform more reliable and interpretable retrieval-augmented generation in practice. Additionally, we plan to analyze the stability of precedence hubs across different query subsets, in order to better disentangle corpus-driven effects from those induced by the query distribution.

ACKNOWLEDGMENTS

The authors acknowledge the use of AI tools in this work. Gemini (Google AI) assisted in improving code for figure generation, and ChatGPT (OpenAI) was used for language editing (grammar and clarity). All content was reviewed by the authors.

REFERENCES

- [1] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, and S. Riedel, "Retrieval-augmented generation for knowledge-intensive nlp tasks," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.
- [2] V. Karpukhin, B. Oğuz, S. Min, L. Wu, S. Edunov, D. Chen, and W.-t. Yih, "Dense passage retrieval for open-domain question answering," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020.
- [3] M. Radovanović, A. Nanopoulos, and M. Ivanović, "Hubs in space: Popular nearest neighbors in high-dimensional data," *Journal of Machine Learning Research*, vol. 11, pp. 2487–2531, 2010.
- [4] R. Feldbauer and A. Flexer, "A comprehensive empirical comparison of hubness reduction in high-dimensional spaces," *Knowledge and Information Systems*, vol. 59, no. 1, pp. 137–166, 2019.
- [5] K. Buza, A. Nanopoulos, and G. Nagy, "Nearest neighbor regression in the presence of bad hubs," *Knowledge-Based Systems*, vol. 86, pp. 250–260, 2015.
- [6] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "Squad: 100,000+ questions for machine comprehension of text," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2016.
- [7] M. Antal and K. Buza, "Evaluating open-source llms in rag systems: A benchmark on diploma thesis abstracts using ragas," *Acta Universitatis Sapientiae, Informatica*, vol. 17, no. 1, pp. 1–20, 2025.
- [8] W. Fan, Y. Ding, L. Ning, S. Wang, H. Li, D. Yin, T. Chua, and Q. Li, "A survey on rag meeting llms: Towards retrieval-augmented large language models," in *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2024.
- [9] S. Gupta, R. Ranjan, and S. N. Singh, "A comprehensive survey of retrieval-augmented generation (rag): Evolution, current landscape and future directions," *arXiv preprint arXiv:2410.12837*, 2024.
- [10] W. X. Zhao, J. Liu, R. Ren, and J.-R. Wen, "Dense text retrieval based on pretrained language models: A survey," *ACM Trans. Inf. Syst.*, vol. 42, no. 4, pp. 89:1–89:60, 2022.
- [11] B. M. Nielsen and L. K. Hansen, "Hubness reduction improves sentencebert semantic spaces," in *Proceedings of the 5th Northern Lights Deep Learning Conference*, ser. Proceedings of Machine Learning Research, vol. 233, 2024, pp. 181–204.
- [12] T. Zhang, F. Suyu, R. Jha, C. Zhang, and V. Shmatikov, "Adversarial hubness in multi-modal retrieval," *arXiv preprint arXiv:2412.14113*, 2024.
- [13] Z. Lin, Z. Wang, T. Qian, P. Mu, S. Chan, and C. Bai, "Neighborretr: Balancing hub centrality in cross-modal retrieval," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.
- [14] F. Liu, R. Ye, X. Wang, and S. Li, "Hal: Improved text-image matching by mitigating visual semantic hubs," in *AAAI Conference on Artificial Intelligence*, 2020.
- [15] R. Nogueira, Z. Jiang, R. Pradeep, and J. Lin, "Document ranking with a pretrained sequence-to-sequence model," in *EMNLP (Findings)*, ser. Findings of ACL. Association for Computational Linguistics, 2020, pp. 708–718.
- [16] Z. Qin, R. Jagerman, K. Hui, H. Zhuang, J. Wu, L. Yan, J. Shen, T. Liu, J. Liu, D. Metzler, X. Wang, and M. Bendersky, "Large language models are effective text rankers with pairwise ranking prompting," in *NAACL-HLT (Findings)*, ser. Findings of ACL. Association for Computational Linguistics, 2024, pp. 1504–1518.
- [17] M. S. Rashid, J. A. Meem, Y. Dong, and V. Hristidis, "Ecorank: Budget-constrained text re-ranking using large language models," in *ACL (Findings)*, ser. Findings of ACL. Association for Computational Linguistics, 2024, pp. 13 049–13 063.
- [18] W. Sun, L. Yan, X. Ma, S. Wang, P. Ren, Z. Chen, D. Yin, and Z. Ren, "Is chatgpt good at search? investigating large language models as re-ranking agents," in *EMNLP*. Association for Computational Linguistics, 2023, pp. 14 918–14 937.
- [19] S. Yoon, E. Choi, J. Kim, H. Yun, Y. Kim, and S. Hwang, "Listt5: Listwise reranking with fusion-in-decoder improves zero-shot retrieval," in *ACL (1)*. Association for Computational Linguistics, 2024, pp. 2287–2308.
- [20] A. Abdallah, B. Piriyani, J. Mozafari, M. Ali, and A. Jatowt, "How good are llm-based rerankers? an empirical analysis of state-of-the-art reranking models," in *EMNLP (Findings)*. Association for Computational Linguistics, 2025, pp. 5693–5709.
- [21] P. Rajpurkar, R. Jia, and P. Liang, "Know what you don't know: Unanswerable questions for squad," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*. Association for Computational Linguistics, 2018.
- [22] L. Wang, N. Yang, X. Huang, B. Jiao, L. Yang, D. Jiang, R. Majumder, and F. Wei, "Text embeddings by weakly-supervised contrastive pre-training," *CoRR*, vol. abs/2212.03533, 2022.
- [23] L. Team, "The llama 3 herd of models," *CoRR*, vol. abs/2407.21783, 2024.
- [24] A. Yang, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Li, D. Liu, F. Huang, H. Wei, H. Lin, J. Yang, J. Tu, J. Zhang, J. Yang, J. Yang, J. Zhou, J. Lin, K. Dang, K. Lu, and K. Bao, "Qwen2.5 technical report," *arXiv preprint arXiv:2412.15115*, 2024.
- [25] M. Radovanović, A. Nanopoulos, and M. Ivanović, "On the existence of obstinate results in vector space models," in *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2010, pp. 186–193.
- [26] G. Izacard and E. Grave, "Leveraging passage retrieval with generative models for open domain question answering," in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, 2021.

APPENDIX

The judge LLM was used with the following prompt template:

You are an impartial evaluator for question-answering tasks.

Your job is to evaluate whether the LLM answer correctly answers the question, based on the provided ground truth answer.

Rules:

- Focus on factual correctness and completeness.
- Ignore differences in wording or style.
- Do not reward unsupported extra information.
- Do not penalize correct paraphrasing.

Scoring rubric:

- 5 = Fully correct and equivalent to the ground truth
- 4 = Mostly correct, very minor omission or imprecision
- 3 = Partially correct, missing key information
- 2 = Mostly incorrect, only small correct elements
- 1 = Completely incorrect or unrelated

Question:

{question}

Ground truth answer:

{answer}

LLM answer:

{llm_answer}

Output exactly two lines:

SCORE: <integer from 1 to 5>

EXPLANATION: <brief explanation>

An LLM-Assisted Framework for Improving the Robustness of Machine Learning Models to Label Noise

Ziming Fang
*Department of Autonomous Systems
University of Stuttgart
Stuttgart, Germany*

Marcel Dix
*Industrial AI
ABB AG Corporate Research Center
Mannheim, Germany*

Yuliang Ma
*Institute of Industrial Automation
and Software Engineering
University of Stuttgart
Stuttgart, Germany*

Andrey Morozov
*Institute of Industrial Automation and Software Engineering
University of Stuttgart
Stuttgart, Germany*

Abstract—The accuracy and trustworthiness of supervised machine learning in industrial settings are frequently undermined by label noise arising from time pressure, evolving taxonomies, and annotator disagreement. Empirical evidence shows that noisy supervision is common rather than exceptional, and that robustness cannot be inferred from clean-label evaluations alone. This paper presents an LLM-assisted framework for analyzing and improving the robustness of machine learning models to label noise. Our framework automates two bottlenecks faced by practitioners: (i) selection of applicable robustness best practices from the literature, and (ii) systematic evaluation of the resulting model variants under structured label-noise perturbations that reflect operational conditions. Given a baseline model and problem description, the LLM proposes concrete, executable modifications and experiment scripts; a deterministic harness then applies standardized metrics across noise types and rates to produce reproducible robustness reports. We show that our framework reduces manual engineering effort, surfaces method-scenario mismatches that are obscured by single-point evaluations, and yields actionable guidance for selecting robustness techniques aligned with industrial constraints and governance needs. The framework and results demonstrate a practical path from robustness intent to deployment-grade evidence in supervised machine learning with imperfect labels.

Index Terms—label noise, ML model robustness, robustness testing, test automation, large language models, industrial AI

I. INTRODUCTION

Industrial environments are increasingly instrumented and interconnected, making data-driven decision-making central to safe, reliable, and efficient operations. In practice, however, the effectiveness of *supervised* machine-learning pipelines depends on label quality, which is an assumption that is rarely satisfied on the factory floor. Labels assembled from incident logs, shift notes, or expert judgments often contain inconsistencies or outright errors due to limited context, time pressure, and subjective interpretation. As a result, model development and evaluation that rely on “golden labels” can paint an overly optimistic picture of performance, masking a hidden source of risk for downstream operations. Recent studies call for

systematically characterizing and mitigating label noise in industrial ML to ensure trustworthy deployment and governance of model behavior in production settings [1], [2]. Classical deterministic rules and model-based monitoring remain attractive for their simplicity and interpretability, yet they struggle as processes, sensors, and operating regimes grow more complex and non-stationary. Supervised ML promises to capture nonlinear, multivariate dependencies in high-dimensional data, but when labels are noisy or incomplete, models can become brittle: false negatives conceal incipient problems while false positives trigger costly interventions and erode user trust. Consequently, robustness to label noise is not a peripheral concern; it is central to the safe, economical use of supervised ML in industrial contexts. *Fault detection*, a relevant AI use case in the industrial domain, is a representative example where unreliable labels about known anomalies degrade the robustness of trained detectors, but the underlying challenge generalizes to any supervised classifier. Despite broad recognition of the prevalence and risks of label noise, practitioners face practical hurdles: annotation is performed by few domain experts under workload constraints; classes and taxonomies evolve over time; and evaluation practices are fragmented across datasets, metrics, and noise assumptions, complicating technology transfer. Moreover, implementing and comparing robustness techniques is time-consuming, requiring engineers to hand-code candidate methods and interpret trade-offs across noise regimes. These realities motivate automated, reproducible robustness-analysis procedures tailored to industrial needs, where deployment decisions must balance safety, maintenance cost, and operational reliability under imperfect supervision. To address these gaps, we propose an automated robustness-analysis framework that for a given baseline model (1) leverages large language models (LLMs) to generate robust code variants based on robust modeling best practices from the literature, and (2) benchmarks the robustness of the model variants with the help of automated label noise infusion. The

approach reduces manual, expertise-intensive iterations, supports reproducible experiments across datasets and tasks, and yields actionable guidance for selecting robustness techniques compatible with industrial constraints. The remainder of this paper is organized as follows: Section II reviews related work concerning label noise in supervised ML and the emerging role of Large Language Models in automating ML workflows. Section III details the proposed solution, describing the framework architecture, the LLM-driven input/output design, and the interaction between the robustness and analysis modules. Section IV presents the experimental evaluation conducted on the widely-used Tennessee Eastman Process (TEP) dataset [3], focusing on three prompting strategies for generating robust model variants. Section V concludes the paper and outlines directions for future work, including budget-aware search and governance artifacts.

II. RELATED WORK

Label noise is a pervasive and often underappreciated challenge across supervised machine learning: when training labels deviate from the true underlying classes, model estimates of generalization can be biased and downstream decisions misled. In industrial AI, where labels are assembled from expert judgments, incident logs, or derived heuristics, the problem is especially pronounced: labels reflect time pressure, evolving taxonomies, partial context, and annotator disagreement rather than an immutable “golden truth.” Recent industrial studies underscore both the prevalence and the risk, reporting that roughly 8–38% of real operational datasets are affected by label noise and arguing that robustness to imperfect supervision must be treated as a first-class requirement rather than a post-hoc adjustment [1], [2], [4]. These observations align with the broader LNL (learning with noisy labels) literature, which documents how noise degrades deep classifiers across domains and motivates methods that explicitly account for noise structure and rate during training and evaluation [4]–[6]. Surveys such as Song *et al.* synthesize a large body of LNL techniques into several families and, crucially, highlight that there is no universal best approach: effectiveness depends on the noise *type* (symmetric, class-dependent/pair-flip, instance-dependent), the *task* and architecture, and the *constraints* of the deployment context [4]. Classical taxonomies and empirical analyses further show that robustness depends not merely on the overall noise *rate* but on the *shape* of the noise distribution, i.e. uniform-flip assumptions often overstate robustness in realistic settings where mislabeling is structured and feature-dependent [5], [7]. Industrial evidence mirrors these findings: labels produced under heterogeneous operating regimes and shifting definitions rarely conform to uniform corruption models, which is why robustness characterization must move beyond single-point estimates toward structured perturbation scenarios and standardized reporting [1], [2]. Across technique families, several best practices recur. *Sample selection and reweighting* methods exploit the small-loss principle and inter-model disagreement to prioritize likely-clean examples (e.g., Co-teaching, MentorNet), or jointly partition and

refine labels as in DivideMix [8]–[10]. *Robust losses and training dynamics* attenuate the influence of corrupted labels via MAE or generalized cross-entropy (GCE), prediction-guided bootstrapping, label smoothing, and curriculum learning [11]–[14]. *Explicit noise modeling* corrects the loss using transition matrices or adds adaptation layers to learn mappings from clean to noisy labels [15], [16]. Complementary directions include architectural/regularization choices (e.g., batch normalization, dropout, early stopping), consistency regularization, and representation learning (self-/semi-supervised) to reduce reliance on perfect labels [4], [6], [17]. There is also nuanced evidence that, under certain conditions, over-parameterized networks can remain competitive in the presence of significant label noise, though such results are highly context-dependent and do not obviate the need for careful robustness evaluation [18].

LLMs as automation enablers for robustness workflows.: Large Language Models (LLMs) have evolved from general-purpose text generators into programmable, tool-using components that can be invoked through APIs and integrated into ML engineering pipelines. Beyond instruction following [19], code-focused variants can synthesize non-trivial program fragments and end-to-end scripts [20], while prompting strategies such as chain-of-thought improve planning and decomposition for multi-step tasks [21]. For automation, LLMs can interleave reasoning with external actions (e.g., calling tools, running code, querying data) [22] and orchestrate heterogeneous ML components [23], enabling them to propose, implement, and iteratively refine robustness experiments rather than merely describe them. Complementary lines of work apply LLMs to pipeline assistance, AutoML-style configuration and code synthesis with testing feedback, demonstrating practical gains in engineering throughput [24], [25], as well as label-quality triage where LLM-based analyses help surface annotation issues for human review [26]. We position our contribution in this automation space, but with a focus on *robustness-oriented model changes* and *standardized, reproducible testing* under structured label-noise regimes.

Positioning our contribution.: Building on the above capabilities and the need for systematic, noise-aware evaluation highlighted in industrial settings [1], [2], we leverage an LLM not only as a literature-aware recommender of label-noise robustness practices, but also as an *automation companion* that proposes concrete model/training modifications and executable evaluation code. The LLM retrieves and prioritizes candidate techniques (e.g., sample selection, robust losses, correction layers, regularization) from the LNL taxonomy [4], instantiates them for the given baseline model and dataset constraints, and then interfaces with a deterministic robustness harness that applies structured label-noise perturbations and standardized metrics. The resulting measurements close the loop for data-driven selection: instead of relying on ad-hoc memory of best practices, the pipeline systematically *tests* how LLM-suggested changes affect robustness under realistic noise models, with reproducible reports suitable for governance and deployment decisions.

III. PROPOSED SOLUTION

This section presents a reusable, automation-ready framework for evaluating and improving the robustness of supervised machine learning models under label noise. Our solution extends a framework that has been presented in [1], [2] for measuring the robustness of machine learning models to label noise. The original framework was designed solely for model testen. It takes a given model provided by the user, and then its fundamental concept was to make use of perturbation to gradually introduce an increasing amount of noise to the ground truth labels, and observes the impact of the noise to the model’s performance. If the reported performance is not satisfactory to the users, they themselves would then have to manually implement possible model changes to the baseline model, that could improve the model’s robustness, and give the model to the framework for a new test run. Depending on the success of the user’s changes, this may result in a longer manual trial-and-error process until an acceptable robustness is found for the given baseline model.

The basic idea of our proposed framework extension presented in this paper is to automatize this model adaptation process with the help of LLM-assisted code generation. For a given baseline model provided by the user, the framework automatically implements various possible model-adaptions that may improve the model’s robustness, creating several model variants. For each variant the framework makes use of label perturbation to test the robustness of the model, as it was already done in the original framework presented in [1], [2]. Finally, the framework presents the user a result of the analysis as a comparison which model adaptation worked best to improve robustness. A stop criterion, such as a maximum number of model variants to test, can control the overall analysis duration and depth.

The design follows three principles: (i) leveraging LLMs to generate robust model variants, for a given baseline model, based on established robust learning best practices from the literature, and (ii) automatically evaluating the generated model variants with help of the framework by injecting realistic label-noise regimes with the help of label noise perturbations. Figure 1 summarizes the architecture.

A. System Overview and Architecture

The core functionality of the framework can be divided into three major parts:

- **LLM-Assisted Improvement Subsystem:** Proposes executable, robustness-oriented model/training variants from a given baseline using literature-backed best practices under schema constraints.
- **Label Perturbation Subsystem:** Generates structured label-noise scenarios (symmetric, class-dependent/pair-flip, instance-dependent) with minimal, configurable interfaces and logged seeds/configs.
- **Robustness Analysis Subsystem:** Trains/evaluates baseline and LLM variants under these scenarios with a deterministic loop, unified metrics, and compact robustness reports.

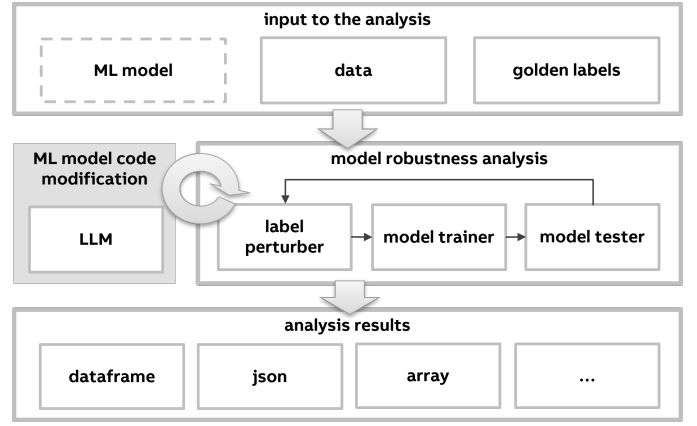


Fig. 1: Core robustness analysis framework with LLM integration. The LLM proposes model variants; a deterministic harness executes standardized evaluations under structured label noise (adapted from [1], [2]).

This separation makes responsibilities explicit, prevents LLM-generated code from altering noise instrumentation or metrics, and eases audits. **LLM-assisted automation.** The LLM acts as an *automation companion* at two touchpoints. (1) **model enhancement:** It proposes concrete, executable modifications to a supplied baseline (e.g., switching to robust losses, adding loss-correction layers, introducing sample-selection strategies, or regularization changes); and (2) **noise scenario control.** The framework provides structured label-noise scenarios aligned with industrial error modes and exposes them through a minimal interface for analysis. Outputs are constrained by schemas and validated before execution. **Workflow at a glance.** The end-to-end loop is:

- 1) Ingest dataset, baseline model, and operational constraints.
- 2) Instantiate a noise generator and select noise configurations.
- 3) Obtain LLM-proposed model/training variants (Sec. III-B).
- 4) For each (model, noise) pair, run standardized training/evaluation.
- 5) Aggregate results into *robustness reports* over noise ratios and types.
- 6) Optionally feed measurements back to the LLM to guide the next round.

B. LLM Workflow for Generating Robust Model Variants

The LLM maps a textual problem description (task, baseline model architecture/cooe, and constraints) into *runnable* code variants, implementing robust learning best practices from the literature: To assure there required outputs, the LLM prompts are structured as follows:

- **Role header:** Fixes domain context (“You are an expert on industrial AI focusing on ML model robustness against label noise...”) in robust learning from noisy labels”)
- **Task card:** Requests a bounded set of changes (e.g., “Generate N variants that improve robustness under

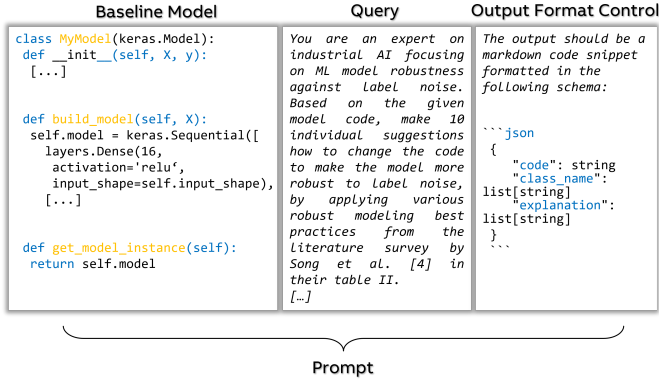


Fig. 2: Example for LLM prompt components to generate robust model variants for a given baseline model

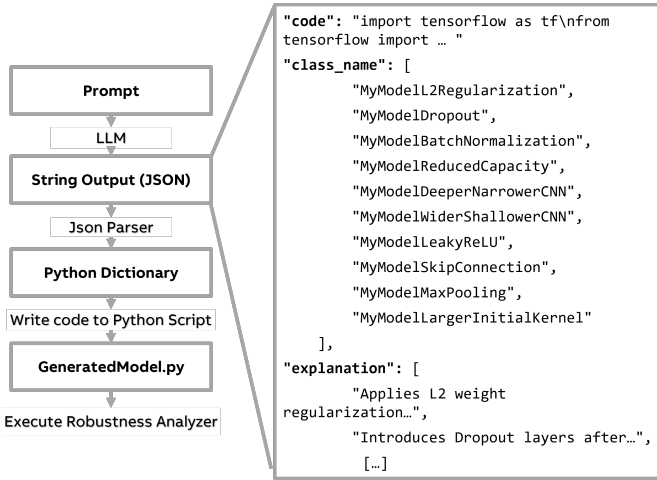


Fig. 3: LLM prompting workflow: the baseline model code, user query, and output schema guiding the LLM to generate robust model variants.

- class-dependent and instance-dependent noise; keep public method signatures; respect compute/time budgets”).
- *Hard constraints:* Output schema and naming rules for code modules; any forbidden operations; reproducibility requirements.
- *Optional few-shot exemplars:* Small, validated snippets that illustrate acceptable edits and templates, used to stabilize outputs.

An example is given in Fig. 2 The LLM returns structured JSON blocks or code stubs, enabling deterministic parsing and execution. Figure 3 illustrates how the LLM is queried and how its structured outputs are consumed in by the solution framework.

The system enforces schema checks, unit tests, static inspection, and a controlled import path. Proposals that fail validation are rejected; the LLM is re-queried with feedback. This ensures that automation does not compromise reproducibility or safety.

C. Integration From LLM Output to Executable Modules

LLM responses are written to versioned source files (e.g., `model_variant_XYZ.py`). A loader then imports the modules at runtime and registers callable model constructors as first-class, auditable pipeline components. All model variants expose uniform public methods (e.g., `train(X, y; ...)` and `predict(X)`). The analysis harness handles re-initialization from a fixed seed, capped epochs with early stopping, and evaluation without per-model glue code.

D. Label Perturber Module

This module implements perturbers that can create realistic label noise patterns highlighted in literature [4], including:

- 1) *Symmetric (uniform) noise:* Labels flip uniformly to any other class.
- 2) *Class-dependent/pair-flip noise:* Mislabeling flows between confusable classes.
- 3) *Instance-dependent noise:* Flip probability varies with sample difficulty or context.

Each family is parameterized by a noise ratio $\rho \in [0, 1]$ and, where applicable, a confusion structure or difficulty map. Hybrids (e.g., asymmetric + instance-dependent) are supported once validated. *Realism*, *diversity*, and *reproducibility* (fixed seeds, logged configs) support fair, repeatable comparisons. For each noise family, the harness evaluates models over a schedule of ratios (e.g., $\rho \in \{0\%, 5\%, 10\%, \dots\}$), yielding performance-versus-noise curves and compact area-based summaries (Sec. III-F).

E. Robustness Analysis Module

For each (model variant, noise config): (i) perturb training labels (test labels remain clean unless explicitly testing noisy evaluation), (ii) train with fixed re-initialization and a standard optimizer, (iii) evaluate on the clean test set, (iv) log metrics and artifacts, (v) repeat across ρ values and multiple seeds. Runs are fully logged: dataset version and split policy, model code hash, hyperparameters, noise type/ratio, random seeds, and metrics, enabling *exact* reruns and straightforward ablations.

F. Metrics and Reporting

We report key metrics (e.g., accuracy, F1) across ρ to visualize robustness degradation as the label noise increases. Curves are averaged over seeds and may be summarized at selected ρ points (e.g., 0%, 10%, 20%, 30%). To summarize robustness across noise regimes with a single scalar, we use the *area under the performance-vs-noise curve* (AUC), as illustrated in Fig. 4. Intuitively, a model is more robust if it maintains high performance as the label-noise ratio ρ increases, producing a larger shaded area. Because clean accuracy can differ across variants, we optionally report a *scaled AUC* (dividing by each model’s clean score) to emphasize stability rather than peak accuracy. In Section IV, we rely primarily on AUC-based bar charts to compare models.

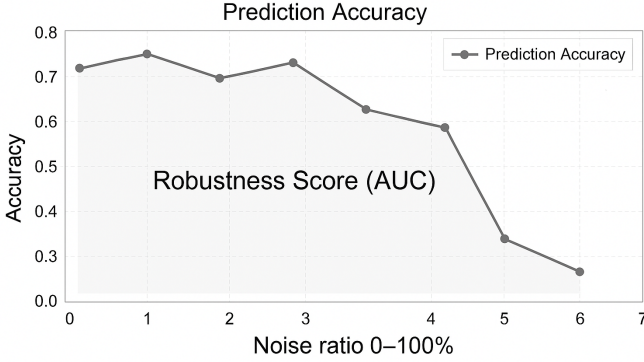


Fig. 4: Concept of robustness as AUC under a performance–vs–noise curve. Higher AUC indicates a model maintains useful performance across larger noise levels. Scaled AUC normalizes by clean performance to emphasize stability.

G. Operational Requirements and Constraints

Prompts, code, and datasets are versioned; seeds fixed; artifacts persisted with metadata (who/when/which config). LLM-generated code must pass schema validation, static checks, and unit tests; execution occurs in a sandbox with restricted access. Human approval gates can be retained for high-stakes contexts. These safeguards are particularly relevant in safety-critical industrial applications, where unintended model behavior may lead to production disruptions or safety hazards. Parallel sweeps across noise ratios and variants are supported. Trials are bounded by capped epochs and early stopping; partial reports stream as runs complete.

H. Feedback and Iteration

The framework optionally feeds evaluation signals back to the LLM. In a “*Query with Reference*” mode, the LLM receives summary metrics and AUCs from prior runs and is asked to generate refined variants that preserve gains while addressing weaknesses. A lightweight “*Survival of the Fittest*” iteration selectively combines promising techniques (e.g., robust loss + early stopping) under a constrained search to avoid combinatorial explosion. To contextualize these capabilities and delimit intended usage, we next outline the scope and positioning of the framework within robustness-oriented model development.

I. Scope and Intended Use

It is important to note that the framework is not an AutoML system and does not replace expert design; it is a *robustness exploration companion*. It automates recall of literature-backed practices and runs noise-aware experiments to generate deployment-grade evidence on how specific modifications affect robustness for a given model and dataset. In summary, the framework unifies (a) LLM-driven proposals of robustness techniques tailored to a given supervised baseline, (b) structured and realistic label-noise generation, and (c) standardized, reproducible robustness measurement. In Section IV,

we show how AUC-based summaries expose method–scenario mismatches and identify resilient model variants under realistic noise. From an industrial perspective, this enables practitioners to translate robustness evaluation results directly into risk-informed model selection and lifecycle management decisions.

IV. EVALUATION

This section evaluates the framework on a representative industrial dataset and analyzes how LLM-assisted proposals impact robustness under structured label noise. We first describe the experimental setup and then present results for three querying strategies defined in Sec. IV-B: *Open Query*, *Query with Reference*, and *Survival of the Fittest*. To keep the exposition reader-friendly, we report robustness using AUC (cf. Fig. 4) and present results exclusively as *bar charts*; we do not include per-noise line plots.

A. Experimental Setup and Data

We use the *Tennessee Eastman Process (TEP)* dataset [3], which is a widely used multivariate process benchmark featuring multiple operating modes and fault conditions. Time-series signals are windowed into supervised examples; labels designate operating conditions/classes.

Continuous variables are scaled to $[0, 1]$. A train/test split preserves temporal patterns and avoids leakage. Fixed-length windows (W) and overlap are used as customary for sequence classification. All experiments use fixed random seeds for exact reruns. The baseline is a compact 1D CNN classifier for multivariate sequences. *Microsoft Copilot* is used as the LLM assistant to propose executable *model/training modifications* following the workflow in Sec. III. All LLM-suggested variants pass schema checks and unit tests before evaluation in the deterministic analysis loop.

We evaluate robustness against the noise families defined in Sec. III-D (symmetric, class-dependent/pair-flip, instance-dependent). For each family, we vary the noise ratio ρ on the training labels (the test set remains clean unless specified otherwise), train from scratch with fixed initial weights, and record test performance. Finally, we summarize robustness as *AUC* over the performance–vs–noise curve. Results are shown as two-column bar charts comparing baseline and LLM-generated variants across noise families; values are averaged over multiple seeds.

B. Results: LLM-Driven Querying Strategies

We instantiate the LLM prompting flow (Fig. 3) to explore three practical strategies that balance exploration with engineering efficiency. In each strategy, Copilot generates code-conformant proposals that the framework compiles, trains, and evaluates under structured noise.

1) *Strategy 1: Open Query*: Here, the idea is to encourage diverse proposals by asking the LLM to freely suggest robustness-oriented changes to the baseline, architectural regularizers, robust losses, loss-correction layers, or sample-selection routines, without prior performance feedback.

Example: “Given the following baseline time-series classifier [code is then given separately with this prompt, cf. Fig. 2], generate ten executable model/training variants that are likely to improve robustness to label noise. Keep public method signatures stable, respect the runtime budget, and return code as modules that pass the provided unit tests.”

Open-ended query strategy yields a broad mix of viable variants. Several proposals, e.g., adding *batch normalization*, enabling *early stopping*, or switching to a *noise-tolerant loss*, consistently increase AUC over the baseline, especially in symmetric and class-dependent noise. Dispersion remains noticeable: some variants underperform in instance-dependent settings due to over-regularization or capacity constraints. As a discovery tool, Open Query is effective for generating candidate improvements, but it benefits from further guidance.

2) *Strategy 2: Query with Reference:* This is a more focused approach where the LLM is provided with reference architectures or robust modeling best practices from the literature, to guide its design proposals. A well-working reference we found for our experiments was to point the survey paper by Song et al. [4], particularly to their table II in this survey, which provides a thorough overview of various robust learning techniques and best practices. Our approach for a query with reference balances LLM-creativity with domain-specific constraints, often leading to higher-performing yet stable architecture.

Example: “You are an expert on industrial AI focusing on ML model robustness against label noise. Based on the given model code, make 10 individual suggestions how to change the code to make the model more robust to label noise, by applying various robust modeling best practices from the literature survey by Song et al. [4] in their table II.”

With references, proposal quality improves and the variance decreases. Most refined variants outperform the baseline across all families; combinations such as *batch normalization + label smoothing* or *early stopping + focal/GCE* achieve higher AUC and more uniform behavior across families. Compared with Open Query, the fraction of underperformers drops markedly, showing that feedback effectively steers Copilot toward more robust remedies.

3) *Strategy 3: Survival of the Fittest:* Here, the idea is to iteratively combine ingredients from the strongest individual variants identified so far, while constraining the LLM to avoid combinatorial explosion. The search is constrained by providing the LLM with *performance feedback* from prior runs. The working hypothesis is that combinations of individually strong techniques tend to remain strong, provided their interactions are not antagonistic.

Example: “Starting from the best AUC variant so far, generate five new, executable models that combine its components with one additional robustness technique from the candidate list (e.g., *early stopping*, *label smoothing*, *loss correction*, *sample selection*). Avoid heavyweight changes or incompatible pairs; return code modules and a short rationale per variant.”

Selective combinations such as *data-centric augmentation*, and *early stopping* or *batch normalization* and *focal/GCE* often achieve the highest AUCs, particularly in low-to-moderate noise ranges relevant to deployment. However, not all pairings are synergistic (e.g., some curricula with strong regularization reduce AUC in instance-dependent noise), but the constrained search keeps the failure rate low. Gains taper after one iteration, suggesting a practical stopping point.

Three themes emerge from the AUC summaries. First, breadth-first exploration (Open Query) is useful for discovery, but performance feedback (Query with Reference) improves proposal quality and consistency across noise families. Second, carefully combining strong ingredients (Survival of the Fittest) yields the highest robustness without combinatorial search, especially in realistic noise ranges. Third, AUC/scaled-AUC bar charts provide a compact, decision-focused lens: practitioners can quickly identify variants that trade clean accuracy for markedly improved stability, or the reverse, and select the model that best matches their operational constraints and risk profile.

V. DISCUSSION AND CONCLUSIONS

This paper examined how supervised machine learning in industrial settings can be made more resilient to the issue of *label noise*, by combining an LLM-assisted proposal mechanism with a standardized robustness evaluation harness. Starting from the observation that label noise is common rather than exceptional in the industrial domain, we designed a framework that turns literature-grounded guidance into executable model and training modifications and then measures their effect under realistic, structured noise.

The architecture constrains the LLM to act as an automation companion that produces code-conformant proposals which are validated before use. Robustness is reported in a compact and decision-oriented manner through performance-versus-noise curves and their AUC summaries, enabling practitioners to judge trade-offs quickly and reproducibly while keeping prompts, code, seeds, and configurations under version control.

With the help of the Tennessee Eastman Process (TEP) dataset our analysis was able to present that breadth-first exploration with an *Open Query* can surface a diverse set of useful ideas, such as *batch normalization*, *early stopping*, and *noise-tolerant losses*, that can improve model robustness at low to moderate noise levels, albeit with noticeable variance across generated model variants. Providing references to robust learning best practices from the literature through our *Query with Reference* strategy yields more consistent gains and reduces the number of underperforming variants by steering proposals toward literature-backed techniques. A lightweight *Survival of the Fittest* iteration, which introduces performance feedback from prior runs and selectively combines the strongest ingredients found so far, was able to achieve improvements to model robustness without resorting to exhaustive search. At the same time, improvements tend to plateau after one round, which provides a practical and defensible stopping point. The

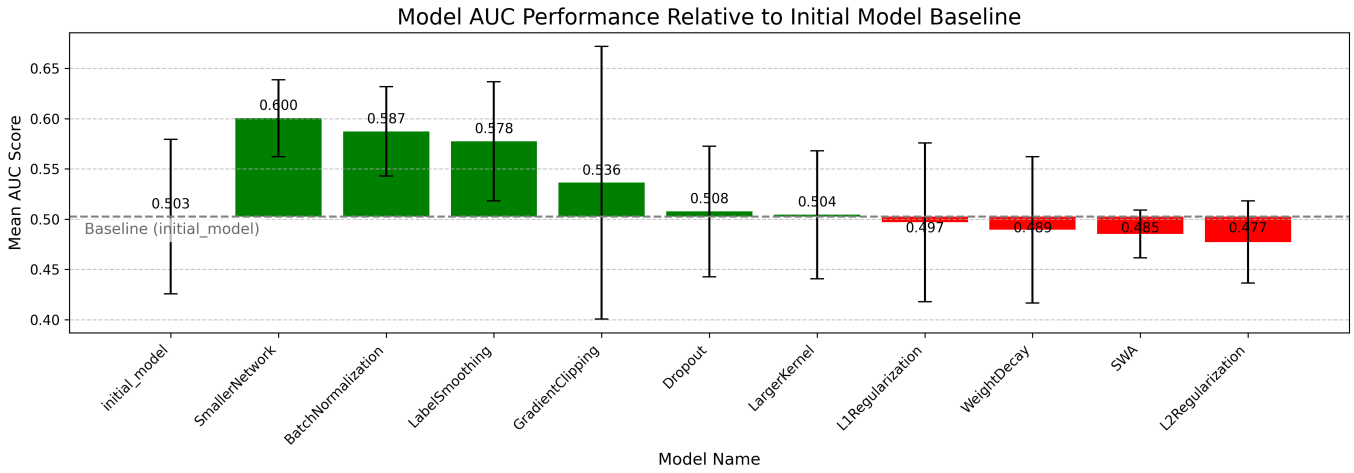


Fig. 5: Open Query: AUC (and optionally scaled-AUC) bars for LLM-generated variants vs. baseline across noise families. Multiple variants exceed the baseline, with variability across families indicating where techniques generalize (or not).

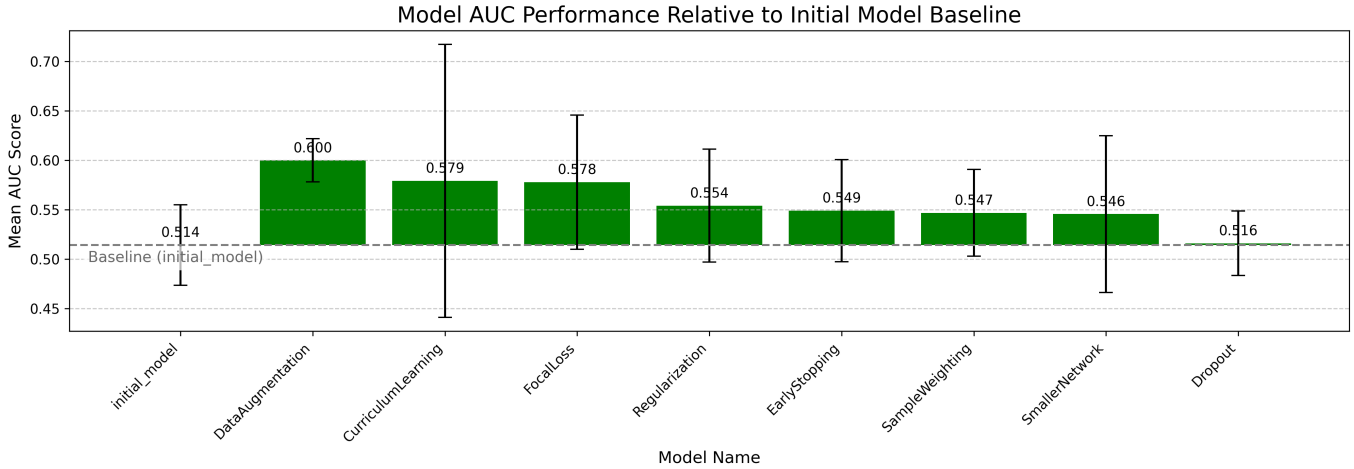


Fig. 6: Query with Reference: LLM-proposals, guided by references from the label noise literature, deliver higher AUCs with reduced dispersion across noise families, indicating better overall robustness compared to unconstrained exploration.

results suggest that robustness enhancement benefits from a progression of exploration, reference-guided refinement, and feedback-driven combination, and that standardized robustness reporting makes the resulting trade-offs transparent for governance and deployment. In particular, variants optimized for realistic low-to-moderate noise regimes, those most frequently encountered in operations, offer the best balance between stability and engineering cost, whereas aggressive regularization or complex schedules become less predictable as noise grows instance-dependent.

A current limitation of our study is being demonstrated on only a single, albeit widely used, multivariate industrial dataset (TEP). Our further research therefore includes testing additional datasets, e.g. from other industrial domains, to evaluate and strengthen the generalization of our framework. Additionally, our future research will include exploring the integration of automated hyperparameter optimization (AutoML) to overcome current manual tuning limitations and un-

lock further robustness gains. Developing adaptive, retrieval-augmented prompting strategies that dynamically learn from prior experimental outcomes could further enhance the quality of model adaptation proposal created by the LLM, and reduce computational costs for testing them in the framework. Finally, advancing toward fully autonomous, multi-agent pipelines with enhanced visualization and user interfaces could facilitate a broader adoption in industrial settings and streamline governance-compliant deployment workflows.

VI. ACKNOWLEDGMENTS

This paper presents results from the underlying master's thesis, which was conducted entirely by the main author and without the use of AI. For the preparation of this paper, AI tools were used to assist in condensing the results, and for grammar correction and wording polishing.

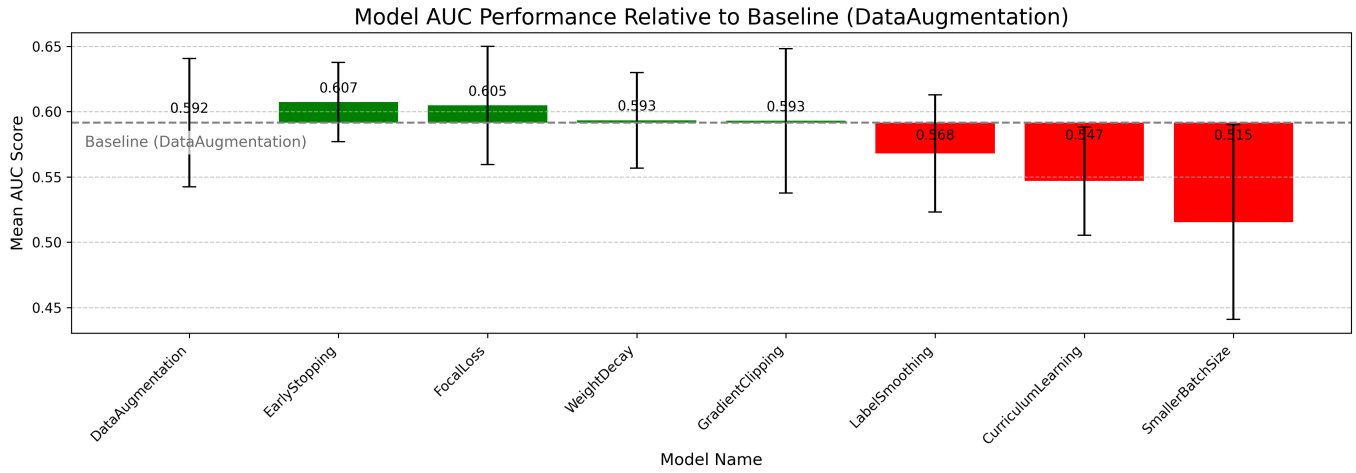


Fig. 7: Survival of the Fittest: constrained combinations reach the top AUCs without exhaustive search. Diminishing returns appear after one round, which is operationally useful for bounding runtime.

REFERENCES

- [1] M. Dix, G. Manca, K. C. Okafor, R. Borrisson, K. Kirchheim, D. Sharma, K. Chandrika, D. Maduskar, and F. Ortmeier, "Measuring the robustness of ml models against data quality issues in industrial time series data," in *2023 IEEE 21st International Conference on Industrial Informatics (INDIN)*. IEEE, 2023, pp. 1–8.
- [2] M. Dix, G. Manca, and A. Fay, "The truth about labels: Unveiling the hidden risk to supervised machine learning models," in *2025 IEEE 30th International Conference on Emerging Technologies and Factory Automation (ETFA)*. IEEE, 2025, pp. 1–8.
- [3] J. Downs and E. Vogel, "A plant-wide industrial process control problem," *Comp. Chem. Eng.*, vol. 17, no. 3, pp. 245–255, 1993. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/0098135493800181>
- [4] H. Song, M. Kim, D. Park, Y. Shin, and J.-G. Lee, "Learning from noisy labels with deep neural networks: A survey," *IEEE transactions on neural networks and learning systems*, vol. 34, no. 11, pp. 8135–8153, 2022.
- [5] B. Frénay, A. Kabán *et al.*, "A comprehensive introduction to label noise," in *ESANN*, 2014.
- [6] G. Algan and I. Ulusoy, "Image classification with deep learning in the presence of noisy labels: A survey," *Knowledge-Based Systems*, vol. 215, p. 106771, 2021.
- [7] D. Oyen, M. Kucer, N. Hengartner, and H. S. Singh, "Robustness to label noise depends on the shape of the noise distribution," *Advances in Neural Information Processing Systems*, vol. 35, pp. 35 645–35 656, 2022.
- [8] B. Han, Q. Yao, X. Yu, G. Niu, M. Xu, W. Hu, I. Tsang, and M. Sugiyama, "Co-teaching: Robust training of deep neural networks with extremely noisy labels," *Advances in neural information processing systems*, vol. 31, 2018.
- [9] L. Jiang, Z. Zhou, T. Leung, L.-J. Li, and L. Fei-Fei, "Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels," in *International conference on machine learning*. PMLR, 2018, pp. 2304–2313.
- [10] J. Li, R. Socher, and S. C. Hoi, "Dividemix: Learning with noisy labels as semi-supervised learning," *arXiv preprint arXiv:2002.07394*, 2020.
- [11] Z. Zhang and M. Sabuncu, "Generalized cross entropy loss for training deep neural networks with noisy labels," *Advances in neural information processing systems*, vol. 31, 2018.
- [12] S. Reed, H. Lee, D. Anguelov, C. Szegedy, D. Erhan, and A. Rabinovich, "Training deep neural networks on noisy labels with bootstrapping," *arXiv preprint arXiv:1412.6596*, 2014.
- [13] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.
- [14] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *Proceedings of the 26th annual international conference on machine learning*, 2009, pp. 41–48.
- [15] G. Patrini, A. Rozza, A. Krishna Menon, R. Nock, and L. Qu, "Making deep neural networks robust to label noise: A loss correction approach," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1944–1952.
- [16] J. Goldberger and E. Ben-Reuven, "Training deep neural-networks using a noise adaptation layer," in *International conference on learning representations*, 2017.
- [17] E. Engleson and H. Azizpour, "Consistency regularization can improve robustness to label noise," *arXiv preprint arXiv:2110.01242*, 2021.
- [18] D. Rolnick, A. Veit, S. Belongie, and N. Shavit, "Deep learning is robust to massive label noise," *arXiv preprint arXiv:1705.10694*, 2017.
- [19] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.
- [20] M. Chen, "Evaluating large language models trained on code," *arXiv preprint arXiv:2107.03374*, 2021.
- [21] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," *Advances in neural information processing systems*, vol. 35, pp. 24 824–24 837, 2022.
- [22] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafraan, K. R. Narasimhan, and Y. Cao, "React: Synergizing reasoning and acting in language models," in *The eleventh international conference on learning representations*, 2022.
- [23] Y. Shen, K. Song, X. Tan, D. Li, W. Lu, and Y. Zhuang, "Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face," *Advances in Neural Information Processing Systems*, vol. 36, pp. 38 154–38 180, 2023.
- [24] S. Zhang, C. Gong, L. Wu, X. Liu, and M. Zhou, "Automl-gpt: Automatic machine learning with gpt," *arXiv preprint arXiv:2305.02499*, 2023.
- [25] G. Dolcetti, V. Arceri, E. Iotti, S. Maffei, A. Cortesi, and E. Zaffanella, "Helping llms improve code generation using feedback from testing and static analysis," *arXiv preprint arXiv:2412.14841*, 2024.
- [26] O. Nahum, N. Calderon, O. Keller, I. Szepietor, and R. Reichart, "Are llms better than reported? detecting label errors and mitigating their effect on model performance," in *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 2025, pp. 26 770–26 797.

From Ethical Principles to Design Practice: Implementing Trustworthy AI in the AUT-KI Project

Rauen, Daniela
Ferdinand-Steinbeis-Institut

Heilbronn, Germany
daniela.rauen@ferdinand-steinbeis-
institut.de

Abstract— This paper examines how the principles of trustworthy artificial intelligence can be operationalized in the development of inclusive assistive technologies. Drawing on the European Union’s normative framework for trustworthy AI, which emphasizes lawfulness, ethics, and robustness, the study analyzes the AUT-KI project as a practical case. AUT-KI aims to develop a digital assistive tool that supports workplace communication for people on the autism spectrum through a participatory and user-centered design process. By reflecting on the project in relation to the EU’s four ethical principles—respect for human autonomy, prevention of harm, fairness, and explicability—and the seven key requirements for trustworthy AI, the paper demonstrates how abstract ethical guidelines can inform concrete development decisions. The findings show that participatory design, context-sensitive risk assessment, transparent governance, and continuous evaluation are essential for building trustworthy AI in socially sensitive contexts. At the same time, tensions remain between empowerment and dependency, inclusion and stigmatization, and technical feasibility and ethical responsibility. The AUT-KI case illustrates that trustworthy AI is not achieved through compliance alone but through ongoing, reflective, and interdisciplinary practice that integrates ethical values into the entire development lifecycle.

Keywords—Trustworthy AI, Assistive Technologies, Participatory Design, Autism Spectrum, AI Ethics, Workplace Inclusion

I. INTRODUCTION

The introduction of artificial intelligence (AI) is no longer exclusively a technical challenge. In addition to questions of performance and efficiency, ethical, social, and societal aspects are increasingly coming into focus [1, 2]. In particular, AI must be perceived as trustworthy by the people who interact with it. Trust is not created solely through technical robustness, but also through transparency, fairness, data protection, and the preservation of human autonomy [3].

These requirements are particularly important in the context of assistive technologies for people with disabilities or limitations. Such systems often interfere with sensitive areas of life and work and address groups of people who are structurally exposed to an increased risk of discrimination, stigmatization, or external control. This makes it all the more important that AI applications in this field are not only functionally effective but also designed in an ethically responsible manner [4, 5].

Against this backdrop, the European Union has created the “Framework for Trustworthy AI” [6], a normative orientation framework that formulates key requirements for lawful,

ethical, and robust AI. The framework provides an important basis for systematically promoting trust in AI systems. However, the question remains as to how these abstract principles and requirements can actually be implemented in the practical development of specific AI applications.

This is precisely where this paper comes in. The aim is to examine how the aspects of the EU framework for trustworthy AI can be taken into account and operationalized in practice. This is done using the AUT-KI project as an example, which pursues the development of a digital assistive technology to support communication in the work context for people on the autism spectrum. Based on this project, we analyze the extent to which and the manner in which trustworthy AI principles are applied in an inclusive, participatory development process and what insights can be derived from this for the design of trustworthy AI in a social context.

This article therefore addresses the question: “How can the principles of trustworthy AI be put into practice in the development of inclusive assistive technologies?”

II. THEORETICAL FRAMEWORK: TRUSTWORTHY AI IN THE EUROPEAN UNION

With the “Ethics Guidelines for Trustworthy AI” from the High-Level Expert Group on Artificial Intelligence [6], the European Union has presented a normative framework for the development and implementation of artificial intelligence. The aim is to design AI systems in such a way that they are not only technically capable, but also legally compliant, ethically sound, and socially responsible. Trustworthy AI is understood as the interplay of three dimensions: lawfulness (lawful AI), ethics (ethical AI), and robustness (robust AI). All three dimensions must be cumulatively fulfilled in order to ensure sustainable trust in AI [6].

At the heart of the framework are four fundamental ethical principles that serve as normative guidelines for the development and evaluation of AI systems: (A) Respect for human autonomy, (B) Prevention of harm, (C) Fairness, and (D) Explicability [6]. These principles form the ethical basis from which the more specific requirements for trustworthy AI are derived.

A. Respect for Human Autonomy

The principle of respect for human autonomy requires that AI systems support rather than undermine individuals’ self-determination and freedom of choice. AI should empower people to make informed decisions rather than manipulating or disempowering them through algorithmic processes. In particular, this implies that human oversight must be ensured

and that AI systems must not be designed in a way that creates dependencies or automates decision-making processes in a non-transparent manner [6, 7].

In the European understanding, autonomy is closely linked to dignity, freedom, and democratic participation. AI must therefore not lead to individuals being reduced to data points or structurally disadvantaged. Rather, it should complement and strengthen human capabilities. [8]

B. Prevention of Harm

The principle of harm avoidance requires that AI systems do not cause physical, psychological, social, or economic harm. This includes both direct risks – such as wrong decisions or security flaws – and indirect or long-term effects, such as stigmatization, social exclusion, or structural disadvantage.

In the context of trustworthy AI, this means that systems must be designed to be technically robust and secure. At the same time, potential side effects must be identified and minimized at an early stage. Careful risk assessment is particularly necessary for vulnerable groups, such as people with disabilities, as asymmetrical power relations and increased protection needs may exist in these cases [6, 9].

C. Fairness

The principle of fairness aims at justice, non-discrimination, and equal treatment. AI systems must not reproduce or reinforce existing social inequalities. Rather, they should take diversity into account and be designed to be inclusive.

Fairness encompasses both the avoidance of algorithmic bias and the consideration of different perspectives in the development process. This implies participatory approaches in which affected groups are actively involved in design and decision-making processes. In the European context, fairness is also understood as a contribution to social cohesion and societal well-being [6, 10].

D. Explicability

The principle of explicability complements the three principles mentioned above and acts as a prerequisite for them. Only when the functioning, decision-making logic, and objectives of an AI system are comprehensible can autonomy, damage prevention, and fairness be verified and guaranteed.

Explicability encompasses both transparency in the technical sense – such as the traceability of data processing and models – and communicative comprehensibility for users. This includes clear information about the purpose, limitations, and risks of the system. Furthermore, accountability is closely linked to explicability, as responsibilities can only be assigned if processes are transparent [6, 11].

In summary, the EU framework for trustworthy AI provides a normative reference framework that combines technological innovation with ethical responsibility. The four ethical principles – autonomy, harm avoidance, fairness, and explainability – serve as fundamental guidelines for the design of human-centered AI systems. They ensure that AI is not evaluated solely on the basis of efficiency or performance standards, but is developed and deployed in accordance with European fundamental values.

Against this backdrop, the central question arises as to how these principles can be operationalized in concrete development processes and translated into practical design decisions. This question forms the starting point for the subsequent analysis of the AUT-KI project. In the following, the AUT-KI project will first be presented and then reflected upon based on the seven key requirements of the EU Framework for Trustworthy AI.

III. METHODOLOGY

This paper follows a theory-driven, framework-based analytical approach to examine how principles of trustworthy AI can be operationalized in practice. The analysis is grounded in the EU “Ethics Guidelines for Trustworthy AI” [6], which provide both a normative foundation and a set of seven Key Requirements that serve as an analytical framework.

The selection of the EU framework is motivated by several considerations. First, it offers a comparatively comprehensive and operationalized structure, as it translates broad ethical principles into concrete and assessable requirements. In contrast to other international approaches, such as those proposed by the OECD [12] or the United Nations [13], which primarily formulate high-level principles, the EU framework provides a more detailed set of criteria that can be directly applied to concrete development processes. Second, the framework is particularly relevant in the European context in which the AUT-KI project is situated, both in terms of regulatory developments and normative expectations regarding human-centered and socially responsible AI. Third, its explicit emphasis on fundamental rights, inclusion, and societal well-being makes it especially suitable for analyzing assistive technologies in sensitive social contexts.

The AUT-KI project is used as an illustrative case. The methodological procedure consists of a deductive application of the seven Key Requirements to the project. In a first step, the requirements – (1) Human Agency and Oversight, (2) Technical Robustness and Safety, (3) Privacy and Data Governance, (4) Transparency, (5) Diversity, Non-Discrimination and Fairness, (6) Societal and Environmental Well-Being, and (7) Accountability – are defined as pre-established analytical categories derived from the theoretical framework.

In a second step, relevant project materials, including the conceptual design, methodological approach, and documented development practices of AUT-KI, are systematically examined in relation to each category. The analysis focuses on identifying how and to what extent the project addresses the respective requirements, as well as where tensions, limitations, or design challenges emerge.

The evaluation is qualitative and interpretative in nature. Rather than measuring compliance in a quantitative sense, the analysis aims to provide a structured, context-sensitive assessment of how normative principles are translated into concrete design decisions. This includes reflecting on both explicit design choices and implicit assumptions within the project.

Finally, the findings are synthesized to derive cross-cutting insights regarding the practical operationalization of trustworthy AI in inclusive assistive technologies. The framework thus functions both as a structuring device for the analysis and as a normative benchmark for evaluation.

IV. THE PROJEKT AUT-KI

The AUT-KI project addresses a key social challenge: the equal participation of people on the autism spectrum in the primary labor market, especially during the sensitive transition phase from school to work. Although many people on the autism spectrum have strong technical skills, analytical abilities, special perceptual strengths, or high reliability, their entry into the workforce and long-term employment often fail not because of a lack of technical qualifications, but because of communication and social barriers in everyday working life [14, 15, 16, 17].

The primary labor market is largely characterized by implicit rules, unspoken expectations, and informal forms of communication. It is precisely these aspects that pose considerable hurdles for many people on the autism spectrum. Misunderstandings, uncertainties, or misinterpretations can lead to stress, overload, and withdrawal – resulting in job losses or permanent exclusion from regular employment. This not only means that individual opportunities are lost, but also that considerable social and economic potential is lost [18, 19].

Against this backdrop, AUT-KI aims to make communication in the workplace more understandable, transparent, and inclusive for people with autism and their colleagues. The focus is on developing a prototype for a digital assistive technology that can be used as a low-threshold tool in everyday working life. The application is intended to help reduce communication misunderstandings, clarify expectations, and create security in everyday work situations [20].

The project is being carried out by a research team of the Ferdinand-Steinbeis-Institut in cooperation with TU Dortmund. As a practice partner the social institution CJD (Christliches Jugenddorfwerk Deutschlands) provides ten experts as well as access to the target group.

A key unique selling point of the AUT-KI project is its consistently participatory and open research approach. While many existing studies, concepts, or support services for people on the autism spectrum are based on pre-formulated assumptions about what is considered helpful or relieving, AUT-KI deliberately takes a different approach: people on the autism spectrum are taken seriously as experts in their own life and work experiences from the outset and are actively involved [21].

The project deliberately refrains from specifying in advance what kind of support might be “good” or “right.” Instead, needs, challenges, and solutions are developed directly from the experiences of those affected. This openness is crucial for developing tailor-made, accepted, and actually used solutions – and for avoiding the danger of well-intentioned but impractical support offers [22].

Methodologically, this is implemented through a multi-stage, qualitative research design. An initial group discussion with people with relevant expertise serves to establish a common understanding of key issues. Building on this, problem-centered expert interviews are conducted with employees with and without autism in order to reveal different perspectives on communication requirements, misunderstandings, and support needs in the work context. In addition, gamified cultural probes are used to enable participants to document experiences, situations, and needs in

a creative and low-threshold manner that is close to everyday life. The insights gained in this way are directly incorporated into the conception and design of the prototype [23].

V. AUT-KI FROM THE PERSPECTIVE OF TRUSTWORTHY AI

The four ethical principles – respect for human autonomy, prevention of harm, fairness, and explicability – form the normative basis of the European understanding of trustworthy AI. They define the values that should guide the development, implementation, and evaluation of AI systems (see II. Theoretical Framework). However, as abstract guiding principles, they need to be further specified in order to actually guide everyday development work.

Against this background, the EU framework formulates seven so-called Key Requirements for Trustworthy AI [6], which are to be understood as a practical operationalization of the ethical principles. These requirements translate the normative principles into verifiable and configurable criteria for the development process. They include:

1. Human Agency and Oversight
2. Technical Robustness and Safety
3. Privacy and Data Governance
4. Transparency
5. Diversity, Non-Discrimination and Fairness
6. Societal and Environmental Well-Being
7. Accountability

The seven requirements specify how autonomy can be ensured, damage avoided, fairness guaranteed, and explainability implemented. For example, “Human Agency and Oversight” operationalizes the principle of respect for human autonomy, while “Technical Robustness and Safety” and “Privacy and Data Governance” contribute to damage prevention. “Diversity, Non-Discrimination, and Fairness” specifies the principle of fairness, and ‘Transparency’ is closely related to the requirement for explainability. “Accountability,” in turn, creates the institutional and organizational conditions necessary for ethical responsibility to be effectively exercised.

The key requirements thus act as a bridge between normative ethics and practical technology design. They make it possible to systematically review specific development processes to determine the extent to which they meet the requirements of trustworthy AI.

In the further course of this paper, the AUT-KI project will be reflected upon based on these seven practical requirements. The aim is to analyze the extent to which and the manner in which the principles of Trustworthy AI are taken into account in the concrete development process and to identify the challenges and scope for design that arise in the process.

A. Human Agency and Oversight

This requirement aims to ensure human self-determination. AI systems should support and empower people, not replace or patronize them. Decisions with relevant implications must always remain under human control (“human-in-the-loop” or “human-on-the-loop”). In addition,

users should be able to make informed decisions about the use of the system [6].

AUT-KI is designed as an assistive technology that supports communication processes but does not make autonomous decisions about work performance or evaluation. The application is intended to enable employees on the autism spectrum to better understand communication situations and act more independently. It is crucial that the system does not impose normative guidelines on how people should behave “correctly,” but rather offers situational guidance.

Participatory development – involving people with autism – further strengthens the preservation of autonomy, as those affected are actively involved in the design and are not merely considered a target group.

B. Technical Robustness and Safety

Technical robustness encompasses the reliability, safety, and fault tolerance of a system. AI must not exhibit uncontrollable or dangerous malfunctions. Risks must be identified, tested, and minimized at an early stage. Resilience to misuse or misinterpretation is also relevant [6].

In the context of AUT-KI, robustness particularly concerns the reliability of communicative cues or support functions. Misinterpretations or imprecise recommendations could reinforce misunderstandings rather than reduce them. Therefore, a careful, empirically based development process – for example, through expert interviews and cultural probes – is central to mapping real communication situations.

In addition, the system should communicate clearly defined limits to its functionality in order to avoid false expectations or overconfidence (“automation bias”).

C. Privacy and Data Governance

This requirement calls for responsible handling of data. This includes data minimization, purpose limitation, secure storage, and transparency regarding data processing. Particularly sensitive data—such as health-related or personal information—is subject to increased protection requirements [6].

Since AUT-KI is situated in the context of autism and workplace communication, the project deals with potentially sensitive personal data. A data protection-compliant design is therefore essential.

Ideally, as little personal data as possible should be processed, and users should retain control over their information. It is particularly important that no data is used for performance monitoring or labor law assessments.

D. Transparency

Transparency requires traceability of the functioning and objectives of an AI system. Users must be able to understand what the system does, on what basis it operates, and what its limitations are. Transparency is a prerequisite for trust and accountability [6].

For AUT-KI, transparency means that both people on the autism spectrum and colleagues can understand how the tool generates support. The application must not be a “black box” whose suggestions appear to be objective truth.

A comprehensible explanation of the functional logic – for example, in simple language or visual form – is

particularly important in order to avoid misunderstandings and to classify the system as a supportive tool.

E. Diversity, Non-Discrimination and Fairness

This requirement stipulates that AI systems must not produce discriminatory effects or reinforce existing inequalities. Diversity must be taken into account and inclusion actively promoted. Bias in data or models must be avoided [6].

AUT-KI explicitly aims to promote professional inclusion and addresses structural communication barriers in the work context. The project is therefore directly related to the principle of fairness.

However, it is important that the system does not implicitly signal that people with autism must adapt exclusively to neurotypical communication norms. A fair design takes mutual understanding into account and also sensitizes colleagues and managers.

Involving those affected in the development process helps to avoid one-sided perspectives.

F. Societal and Environmental Well-Being

AI should serve the good of society and promote sustainable, inclusive structures. In addition to ecological aspects, this includes social impacts such as participation, justice, and social cohesion [6].

The project makes a direct contribution to the social participation of people on the autism spectrum in the primary labor market. By reducing communication barriers, long-term job stability, self-efficacy, and social integration can be strengthened.

AUT-KI therefore does not primarily pursue increased efficiency, but rather social inclusion—a goal that is closely linked to the European understanding of AI as a technology oriented toward the common good.

G. Accountability

Accountability requires clear responsibilities for the development, use, and impact of AI systems. Mechanisms for review, complaint, and correction must be in place. Responsibility must not be delegated to the technology [6].

In the AUT-KI project, accountability is particularly evident in the scientifically based development process and the documented involvement of various stakeholders.

For subsequent implementation, it is crucial that there are clear rules governing who is responsible for possible malfunctions, data protection violations, or negative effects—for example, developers, employers, or operators of the application.

The analysis shows that AUT-KI addresses numerous requirements of the EU framework for trustworthy AI in its objectives and methodological orientation. At the same time, it becomes clear that actual trustworthiness does not result solely from the intention of the project, but from its concrete design, implementation, and continuous reflection along the seven key requirements.

AUT-KI thus offers a suitable case study for systematically investigating the practical operationalization of European AI ethics in the context of inclusive assistive technologies.

VI. DISCUSSION

The analysis of the AUT-KI project against the background of the EU Framework for Trustworthy AI illustrates both the potential and the complexity of operationalizing ethical principles in concrete development processes. The project demonstrates that the requirements of trustworthy AI are not merely abstract normative ideals, but can meaningfully guide the design of inclusive assistive technologies. At the same time, the case study reveals tensions, trade-offs, and open questions that are characteristic of AI systems deployed in socially sensitive contexts.

A. From Normative Principles to Design Practice

A central insight of this study is that the EU framework provides a valuable orientation structure, but its practical implementation requires continuous translation work. Concepts such as autonomy, fairness, or explicability must be interpreted within the specific socio-technical context in which an AI system operates. In the case of AUT-KI, this context is shaped by workplace dynamics, neurodiversity, power asymmetries between employers and employees, and the sensitive nature of communication processes.

The participatory and open research design of AUT-KI can be understood as a key mechanism for translating normative principles into practice. By involving people on the autism spectrum as experts in their own lived experience, the project operationalizes respect for human autonomy and fairness in a concrete manner. Participation here is not an add-on, but a structural element of ethical design. This suggests that participatory approaches are not only methodologically beneficial, but ethically necessary when developing AI for vulnerable or marginalized groups.

However, participation alone does not automatically guarantee trustworthiness. It must be embedded in transparent decision-making structures and accompanied by clear documentation of how stakeholder input influences technical and organizational design choices. Otherwise, there is a risk of symbolic inclusion without substantive impact.

B. Autonomy and New Forms of Dependencies

While AUT-KI aims to strengthen the autonomy of employees on the autism spectrum, the discussion must also consider potential unintended effects. Assistive technologies can empower users by providing clarity and reducing uncertainty. At the same time, they may create new dependencies or subtly reinforce normative expectations about “appropriate” workplace behavior.

A critical question is therefore whether the system primarily supports self-determined action or whether it implicitly promotes adaptation to dominant neurotypical communication norms. Even well-intentioned support tools can contribute to a new form of dependencies if they define what counts as successful communication without sufficient reflexivity.

To address this tension, trustworthy AI in inclusive contexts should not only focus on supporting individuals but also on transforming environments. In the case of AUT-KI, this would mean designing the system not merely as a compensatory tool for autistic employees, but as an instrument for mutual understanding that also sensitizes colleagues and supervisors. In this way, autonomy is

strengthened not through adaptation alone, but through structural change in communication cultures.

C. Harm Prevention Beyond Technical Safety

The requirement of technical robustness and safety is particularly complex in communication-based assistive systems. Harm in this context is less likely to be physical and more likely to be psychological or social. Misleading recommendations, oversimplified interpretations of social situations, or incorrect contextual analyses could exacerbate misunderstandings and potentially damage professional relationships.

The AUT-KI project addresses these risks through qualitative, empirically grounded development methods. Nevertheless, the discussion highlights that harm prevention must extend beyond technical validation. Continuous evaluation in real-world settings, feedback loops, and mechanisms for correction are essential to identify subtle negative effects that may not be visible in controlled testing environments.

Moreover, the social setting of the workplace introduces additional risks. If such a tool were to be used in a way that indirectly monitors or evaluates employees, it could undermine trust and autonomy. Clear governance rules and strict separation from performance assessment mechanisms are therefore indispensable.

D. Data Protection and Power Asymmetries

Privacy and data governance are of particular importance in a project dealing with autism and workplace communication. The processing of sensitive personal data in a hierarchical employment context creates structural power asymmetries. Even if data processing is legally compliant, users may feel pressured to use the system or to share data in order to demonstrate willingness to adapt.

The discussion therefore underlines that formal data policy compliance is a necessary but not sufficient condition for trustworthiness. Voluntariness, informed consent, and meaningful control over personal data are central. Technical solutions such as local data processing, data minimization, and user-controlled deletion mechanisms can contribute to this. At the same time, organizational safeguards – such as clear agreements with employers – are equally important.

Trustworthy AI in employment contexts must always be aware that technological design decisions interact with institutional power structures.

E. Explicability as Communicative Practice

The principle of explicability is often reduced to technical interpretability of models. However, in the context of AUT-KI, explicability must primarily be understood as communicative comprehensibility. Users need to understand what the system can and cannot do, how suggestions are generated, and what degree of uncertainty is involved.

This is particularly relevant to avoid automation bias – the tendency to over-rely on algorithmic outputs. If users perceive the system as an objective authority, they may override their own judgment. Therefore, explicability should include not only explanations of system logic but also explicit communication of limitations and potential error sources.

The case of AUT-KI suggests that explainability must be adapted to the cognitive and communicative needs of

different user groups. Simple language, visualizations, and context-sensitive explanations may be more effective than technically detailed model descriptions.

F. Fairness, Inclusion, and the Broader Social Perspective

AUT-KI is explicitly oriented toward promoting inclusion in the primary labor market. In this sense, it aligns strongly with the EU's understanding of AI as a contributor to social cohesion and well-being. The project addresses structural communication barriers rather than individual deficits alone.

Nevertheless, fairness must be considered dynamically. An assistive technology that benefits one group may inadvertently create new distinctions or stigmatization if its use marks individuals as "different." Designers must therefore carefully consider how the tool is introduced and framed within organizations.

In addition, fairness requires intersectional awareness. People on the autism spectrum are not a homogeneous group. Differences in gender, cultural background, education level, or co-occurring conditions may influence needs and experiences. A truly fair and inclusive system must remain flexible and adaptable to diverse realities.

G. Accountability as Ongoing Responsibility

Finally, the requirement of accountability highlights that trustworthiness is not achieved once and for all at the development stage. It requires long-term structures of responsibility. Clear allocation of roles between developers, operators, employers, and possibly public institutions is necessary to ensure that complaints, malfunctions, or unintended harms can be addressed effectively.

For research-based prototypes such as AUT-KI, the transition from pilot phase to real-world deployment represents a critical moment. Ethical reflection must accompany scaling processes, commercialization decisions, and potential integration into existing workplace systems.

Accountability also includes a willingness to revise or even withdraw a system if significant negative effects become apparent. In this sense, trustworthy AI is a dynamic process rather than a static certification status.

H. Implications for the Operationalization of Trustworthy AI

Overall, the AUT-KI case study illustrates that the EU Framework for Trustworthy AI can serve as a practical evaluation matrix for inclusive AI development. The seven key requirements provide a structured lens through which design decisions can be systematically reflected upon.

At the same time, the discussion shows that operationalization depends heavily on context-sensitive interpretation, participatory processes, and continuous reflexivity. Trustworthiness emerges not only from technical properties, but from the interaction between technology, users, institutions, and social norms.

For the development of inclusive assistive technologies, three overarching implications can be derived:

- Ethical principles must be embedded structurally in the development process, particularly through genuine participation and interdisciplinary collaboration.

- Trustworthiness must be understood relationally, taking into account power structures and social dynamics in the application context.
- Evaluation must be continuous and adaptive, recognizing that social impacts evolve over time.

In conclusion, the AUT-KI project demonstrates that the principles of trustworthy AI can be meaningfully integrated into the development of inclusive assistive technologies. However, it also makes clear that ethical AI is not the result of compliance with a checklist, but of an ongoing, reflective practice that combines technological innovation with social responsibility.

VII. CONCLUSION AND OUTLOOK

This article addressed the research question: How can the principles of trustworthy AI be put into practice in the development of inclusive assistive technologies? Using the AUT-KI project as a case study, the analysis has shown that the operationalization of trustworthy AI is neither an abstract theoretical exercise nor a purely technical task. Rather, it is a structured, context-sensitive, and reflexive design process that translates normative principles into concrete socio-technical practices.

The EU Framework for Trustworthy AI provides a coherent and valuable orientation structure. Its four ethical principles together with the seven key requirements, offer a systematic matrix for guiding development processes. In the case of AUT-KI, these principles were not only referenced normatively but embedded in the methodological and organizational design of the project. Participatory research formats, qualitative needs assessments, attention to data protection, and the explicit focus on inclusion demonstrate how ethical requirements can shape concrete design decisions.

The findings suggest that trustworthy AI in inclusive assistive contexts can be operationalized through three central mechanisms:

- Participatory and inclusive development, in which affected individuals are actively involved as co-designers rather than passive recipients.
- Context-sensitive risk reflection, addressing not only technical malfunctions but also social, psychological, and institutional harms.
- Transparent governance and accountability structures, ensuring clarity of responsibilities, data protection, and long-term oversight.

At the same time, the analysis makes clear that trustworthiness is not guaranteed solely by good intentions or formal compliance. It must be continuously negotiated and reassessed in light of real-world use, power asymmetries, and evolving social dynamics. In particular, inclusive assistive technologies must carefully balance empowerment and the

risk of subtle paternalism, as well as individual support and structural change within organizations.

Looking ahead, several implications for research and practice emerge.

First, future research should further investigate how participatory approaches can be institutionalized in AI development beyond individual projects. Comparative studies across different assistive technologies and social contexts could help identify best practices for embedding the EU framework into everyday development routines.

Second, longitudinal evaluations are needed to assess the long-term social impact of inclusive AI systems. Trust, autonomy, and fairness are dynamic constructs that may evolve as technologies become integrated into organizational structures. Continuous monitoring and adaptive governance models will therefore be essential.

Third, there is a need to bridge ethical guidelines and regulatory developments, particularly in light of the evolving European AI regulatory landscape. The interplay between the Ethics Guidelines for Trustworthy AI and binding legal instruments such as the EU AI Act will shape future design requirements for assistive technologies. Research should explore how ethical reflection and regulatory compliance can complement each other rather than being treated as separate domains.

Finally, the broader societal perspective must remain central. Assistive AI systems such as AUT-KI should not only aim to compensate for individual challenges but also contribute to transforming environments toward greater inclusivity. Trustworthy AI, in this sense, is not only about reliable technology but about fostering social structures that respect diversity and promote equal participation.

In conclusion, the operationalization of trustworthy AI in inclusive assistive technologies is possible when ethical principles are integrated as guiding structures throughout the entire development lifecycle. AUT-KI illustrates that combining participatory research, contextual sensitivity, and structured ethical reflection can translate European AI values into tangible design practice. Trustworthy AI thus emerges not as a fixed state, but as an ongoing process of responsible innovation aligned with human dignity, inclusion, and social well-being.

ACKNOWLEDGMENTS / AI DISCLOSURE

Digital tools (AI) were used to support the writing of this paper, particularly during the early and revision phases of the writing process. These included applications that helped to reflect on conceptual approaches, structure potential directions for the content, and refine the linguistic formulation. Appropriate tools were also utilized in reviewing and categorizing relevant literature, as well as in translating individual passages into appropriate academic English.

However, I bear full responsibility for the content, as well as for the final selection, evaluation, and elaboration of the arguments presented.

REFERENCES

- [1] S. Alpers and B.-J. Krings, "Ethische, rechtliche und soziale Aspekte (ELSA) der Gestaltung von KI-Systemen und Integration der verschiedenen Akteure in den Prozess [Ethical, Legal and Social Aspects (ELSA) of the Design of AI Systems and the Integration of Various Stakeholders into the Process]," in *Proc. Frühjahrskongress 2023: Nachhaltig Arbeiten und Lernen – Analyse und Gestaltung lernförderlicher und nachhaltiger Arbeitssysteme und Arbeits- und Lernprozesse* [Spring Congress 2023: Sustainable Work and Learning – Analysis and Design of Learning-Conducive and Sustainable Work Systems and Work and Learning Processes], B.2.3, Hannover, Germany: Gesellschaft für Arbeitswissenschaft (GfA), 2023.
- [2] J. Whittlestone, R. Nyrop, A. Alexandrova, K. Dihal and S. Cave, *Ethical and Societal Implications of Algorithms, Data, and Artificial Intelligence: A Roadmap for Research*, London, UK: Nuffield Foundation, 2019. Available: [Ethical-and-Societal-Implications-of-Data-and-AI-report-Nuffield-Foundation.pdf](https://www.nuffieldfoundation.org/data-and-ai-report-nuffield-foundation)
- [3] O. Fritz, C. Weber, A. König and J. Wolf, *Ethische Aspekte der Künstlichen Intelligenz* [Ethical Aspects of Artificial Intelligence], KCT Schriftenreihe der FOM, No. 1, Essen, Germany: MA Akademie Verlags- und Druck-Gesellschaft mbH, 2019. Available: <https://www.econstor.eu/bitstream/10419/224465/1/173246250X.pdf>
- [4] T. Wangmo, M. Lipps, R. W. Kressig and M. Ienca, "Ethical concerns with the use of intelligent assistive technology: findings from a qualitative study with professional stakeholders," *BMC Med. Ethics*, vol. 20, Art. no. 98, Dec. 2019. Available: <https://pmc.ncbi.nlm.nih.gov/articles/PMC6924051/>
- [5] A. M. Cook, "Ethical issues related to the use/non-use of assistive technologies," *Developmental Disabilities Bulletin*, vol. 37, no. 1–2, pp. 127–152, 2009. Available: <https://files.eric.ed.gov/fulltext/EJ920692.pdf>
- [6] High-Level Expert Group on Artificial Intelligence (AI HLEG), *Ethics Guidelines for Trustworthy AI*, European Commission, Brussels, Belgium, Apr. 2019. Available: https://www.europarl.europa.eu/cmsdata/196377/AI%20HLEG_Ethics%20Guidelines%20for%20Trustworthy%20AI.pdf
- [7] L. Floridi, J. Cowls, M. Beltrametti, R. Chatila, P. Chazerand, V. Dignum, C. Luetge, R. Madelin, U. Pagallo, F. Rossi, B. Schafer, P. Valcke and E. Vayena, "AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations," *Minds and Machines*, vol. 28, no. 4, pp. 689–707, Nov. 2018. Available: <https://link.springer.com/article/10.1007/s11023-018-9482-5>
- [8] Deutscher Ethikrat, *Mensch und Maschine – Herausforderungen durch Künstliche Intelligenz: Stellungnahme* [Human and Machine – Challenges of Artificial Intelligence: Statement], Berlin, Germany, 20 Mar. 2023. Available: <https://www.ethikrat.org/presse/pressekonferenzen/veroeffentlichung-der-stellungnahme-mensch-und-maschine-herausforderungen-durch-kuenstliche-intelligenz/>
- [9] European Disability Forum (EDF), "EDF Position Paper on EU Artificial Intelligence (AI) Regulation," Nov. 2021. Available: <https://www.edf-feph.org/content/uploads/2021/11/EDF-Position-Paper-on-EU-Artificial-Intelligence-AI-Regulation-1.pdf>
- [10] A. Manzeschke, *IMPACT: Faire Algorithmen und die Fairness von Erklärungen: Informatische, rechtliche und ethische Perspektiven* [Impact: Fair Algorithms and the Fairness of Explanations: Informatics, Legal, and Ethical Perspectives], Policy Paper, Universität Duisburg-Essen, 2022. Available: https://duepublico2.uni-due.de/servlets/MCRFileNodeServlet/duepublico_derivate_00076040/Impact_PolicyPaper_2022_Fairness.pdf
- [11] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bénéttot, S. Tabik, A. Barbado, S. García, S. Gil-Lopez, D. Molina, R. Benjamins and R. Chatila, "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," **Information Fusion**, vol. 58, pp. 82–115, 2020. Available: <https://www.sciencedirect.com/science/article/abs/pii/S1566253519308103?via%3Dihub>
- [12] OECD, *OECD AI Principles*, Organisation for Economic Co-operation and Development 2026. Available: <https://www.oecd.org/en/topics/ai-principles.html>
- [13] United Nations System Chief Executives Board for Coordination (CEB), *Framework for a Model Policy on the Responsible Use of AI in the UN System*, 2024. Available:

- <https://unsceb.org/sites/default/files/2024-11/Framework%20for%20a%20Model%20Policy%20on%20the%20%20Responsible%20Use%20of%20AI%20in%20UN%20System.pdf>
- [14] F. Frank, M. Jablotschkin, T. Arthen, A. Riedel, T. Fangmeier, L. P. Hölzel and L. Tebartz van Elst, "Education and employment status of adults with autism spectrum disorders in Germany – a cross-sectional survey," *BMC Psychiatry*, vol. 18, no. 1, Art. no. 75, 2018, doi: 10.1186/s12888-018-1645-7.
 - [15] M. Dalferth, T. Ederer, H. Kastorff, K. Reich and H. Schleper, "Autismus und Arbeit. Risiko- und Förderfaktoren einer nachhaltigen Beschäftigung [Autism and Work: Risk and Support Factors for Sustainable Employment]," *Die berufliche Rehabilitation*, no. 4, pp. 30–47, 2021.
 - [16] U. Sünkel, "Autismus-Spektrum-Störung und die Arbeitswelt [Autism Spectrum Disorder and the World of Work]," in *Autismus-Spektrum-Störungen im Erwachsenenalter. Neuauflage des Standardwerkes 'Das Asperger-Syndrom im Erwachsenenalter' [Autism Spectrum Disorders in Adulthood: New Edition of the Standard Work 'Asperger's Syndrome in Adulthood']*, 3rd ed., L. Tebartz van Elst, Ed., Berlin, Germany: Medizinisch Wissenschaftliche Verlagsgesellschaft, pp. 331–346, 2021.
 - [17] AUT-1A, *Autismus – Unternehmen – Arbeit [Autism – Company – Work]*, 2025. Available: https://www.autismusundarbeit.de/files/ugd/4f521c_dc4c445acdac47d683a12e556b3cfc14.pdf
 - [18] J. Röhner and A. Schütz, *Psychologie der Kommunikation [Psychology of Communication]*, 3rd rev. and updated ed., Wiesbaden, Germany: Springer, 2020.
 - [19] K. Beck, *Soziologie der Online-Kommunikation [Sociology of Online Communication]*, Wiesbaden, Germany: Springer VS (Essentials), 2014.
 - [20] D. Rauen, H. Lasi and F. Mörike, "Assistive Technology for the Inclusion of People With Autism at Work," in *AAATE 2025 Short Papers Collection*, Nicosia, Cyprus: European University Cyprus; Association for the Advancement of Assistive Technology in Europe (AAATE), Sep. 2025. Available: https://aaate2025.eu/wp-content/uploads/sites/33/2025/09/Short_Papers_Collection_Volume_Final.pdf
 - [21] S. Burtscher and K. Gerling, "Neurodivergence and work in human-computer interaction: Mapping the research landscape," in *Proc. 3rd Annual Meeting of the Symposium on Human-Computer Interaction for Work (CHIWORK '24)*, New York, NY, USA: Association for Computing Machinery, 2024, Art. no. 5, pp. 1–14.
 - [22] F. Mörike and I. Kiossis, "Workarounds as enabling factor in the use of assistive technology amongst white-collar workers with visual impairment," *J. Enabling Technol.*, early access, 2024, doi: 10.1108/JET-02-2024-002.
 - [23] S. Çerçi, M. Cecchinato and J. Vines, "How Design Researchers Interpret Probes: Understanding the Critical Intentions of a Designerly Approach to Research," in *Proc. CHI Conf. Human Factors in Computing Systems (CHI '21)*, Yokohama, Japan, May 2021, New York, NY, USA: Association for Computing Machinery, 2021, Art. no., 15 pp., doi: 10.1145/3411764.3445328.

Beyond Bandwidth: Profiling Software-Level Communication Bottlenecks in Distributed Large Language Model Training

Maximilian Pazer
DHBW Stuttgart
Hewlett Packard Enterprise
Stuttgart, Germany
maximilian.pazer@hpe.com

Hardik Soni
HPE Labs
Heidelberg, Germany
hardik.soni@hpe.com

Max Zeitler
DHBW Stuttgart
Hewlett Packard Enterprise
Stuttgart, Germany
zeitler@hpe.com

Abstract—The continued growth in the size and complexity of large language models has shifted the primary bottleneck of distributed training from computation to communication. As inter-GPU communication becomes increasingly dominant, training efficiency is constrained not only by hardware bandwidth but also by software-level collective mechanisms. This paper investigates the internal communication behavior of Llama 3.1 70B during model-parallel training and examines whether communication inefficiencies arise from hardware saturation or software-induced synchronization. Using a design science research methodology, we develop an analysis artifact capable of extracting fine-grained insights from NCCL Profiler v4 traces. Our experiments on 32 and 64 GPUs reveal that communication inefficiencies predominantly originate from synchronization barriers within collective operations rather than from insufficient interconnect bandwidth utilization. At 64 GPUs, the interconnect operates at only 15% of its 400 Gb/s capacity, yet All-Reduce and Reduce-Scatter operations exhibit $7\times$ and $22\times$ time increases respectively. Wait-state analysis shows All-Reduce spends 49% of its time in synchronization stalls, split between receive-wait (50%) and send-wait (48%). These findings provide empirical evidence that model-parallel scaling inefficiencies are primarily software-driven.

Index Terms—distributed training, large language models, NCCL, collective communication, GPU synchronization, performance profiling

I. INTRODUCTION

The rapid growth of Large Language Models (LLMs) has led to increasing computational demands, pushing training workloads to thousands of GPUs and requiring sophisticated distributed training methodologies [1]. A widely adopted methodology in distributed training frameworks is the implementation of inter-GPU communication through NVIDIA’s Collective Communication Library (NCCL) [2]. Training Llama 3.1 405B required 16,000 H100 GPUs operating for over 54 days [3], underscoring the scale at which communication efficiency becomes critical.

Although hardware capabilities have advanced, research shows that distributed training is increasingly limited by communication rather than compute [1]. Scaling efficiency degrades as more GPUs participate in model and tensor parallel workloads [4], [5]. Studies have shown that GPUs spend

up to 63% of execution time idle during large-scale training, primarily due to communication [6]. As models continue to scale, the efficiency of inter-GPU communication becomes a dominant performance factor.

Research investigating inefficiency at scale highlights that networking hardware remains substantially underutilized during LLM training [7], [8]. Prior work explicitly calls for further investigation of interconnect behavior at larger scales [8]. Other studies suggest that these inefficiencies originate not from hardware saturation but from software-level collective communication mechanisms [7]. However, existing analyses often characterize communication behavior at a coarse level, leaving the underlying mechanisms largely unexplored [9], [10]. To date, there is no prior work examining the NCCL communication layer during large-scale LLM training at a fine-grained level [11].

This observability gap constitutes the central problem addressed in this paper. When training state-of-the-art models such as Llama 3.1 70B using 4D parallelism (tensor, pipeline, context, and data parallelism), it remains unknown whether performance degradation at scale is caused by interconnect saturation or by software-induced synchronization stalls within the communication layer. We address the following research question:

How does the communication behavior of Llama 3.1 70B training scale from 32 to 64 GPUs, and to what extent are communication inefficiencies caused by hardware limitations rather than software-induced mechanisms within NCCL?

This leads to three sub-questions:

- **RQ1:** To what extent does interconnect bandwidth utilization approach its theoretical peak during model-parallel training when scaling from 32 to 64 GPUs?
- **RQ2:** How do NCCL collective communication primitives (AllReduce, AllGather, ReduceScatter) and point-to-point (P2P) operations contribute to overall communication overhead?
- **RQ3:** Which software-level synchronization behaviors, such as NCCL wait states and collective operation stalls, contribute to communication overhead?

Contributions. This paper makes three contributions:

- We provide an end-to-end scaling study of Llama 3.1 70B model-parallel training from 32 to 64 GPUs, combining interconnect telemetry and fine-grained communication traces to quantify how communication overhead changes with scale.
- We develop a reproducible analysis artifact that reconstructs NCCL Profiler v4 events and attributes overhead to both communication primitives and NCCL-internal wait states, addressing the observability gap for NCCL behavior at scale.
- We empirically demonstrate that the dominant scaling degradation at 64 GPUs is driven by software-induced synchronization (wait states) rather than by interconnect bandwidth saturation, with All-Reduce spending 49% of its time in stalls despite the interconnect operating at only 15% capacity.

II. BACKGROUND

A. Distributed Training Parallelism

Training large-scale LLMs requires distributing workloads across multiple GPUs using several parallelism strategies. *Data Parallelism* (DP) replicates the model across GPUs, with each processing different data subsets and synchronizing gradients via All-Reduce [12]. While DP scales well for models that fit in single-GPU memory, larger models require partitioning.

Tensor Parallelism (TP) shards matrix operations within layers across GPUs, requiring frequent All-Reduce, All-Gather, and Reduce-Scatter operations within each transformer layer [4], [13]. For a model with L layers using TP degree t , each forward and backward pass requires $O(L \cdot t)$ collective operations, making TP communication-intensive but enabling training of models that exceed single-GPU memory.

Pipeline Parallelism (PP) partitions layers into stages distributed across GPUs, using point-to-point communication between adjacent stages [14]. PP reduces the frequency of communication compared to TP but introduces pipeline bubbles where some GPUs remain idle while waiting for activations or gradients.

Context Parallelism (CP) distributes input sequences across GPUs to reduce activation memory [3]. For long sequences common in modern LLMs (8K–128K tokens), CP partitions the sequence dimension, requiring All-to-All communication patterns for attention computation.

Modern training frameworks combine these strategies as *4D parallelism*, where each GPU participates in multiple parallelism groups simultaneously. The Llama 3.1 70B configuration used in this work employs TP=4, PP=4, CP=2, with DP scaling from 1 to 2 as GPU count increases from 32 to 64.

B. Collective Communication Operations

Collective operations form the backbone of distributed training communication. *All-Reduce* combines data from all processes using a reduction operator (typically summation for

gradients) and distributes the result back to all processes [2]. It is the primary communication primitive in tensor parallelism, occurring twice per transformer layer—once during the forward pass for attention output and once during the backward pass for gradients [13]. Critically, All-Reduce introduces synchronization barriers where all devices must wait for the slowest participant before proceeding [15].

All-Gather ensures each process shares its data with all others, assembling distributed tensors into complete copies on each GPU. This operation is essential for reassembling activations split during the forward pass [16]. *Reduce-Scatter* performs element-wise reduction and scatters results so each process receives a unique partition [2]. It is used in the backward pass for gradient aggregation, where each GPU needs only its shard of the gradient tensor [16].

NCCL implements these operations with optimizations for dense multi-GPU clusters [2], using ring and tree algorithms that pipeline data transfers across GPUs. Benchmarks show NCCL outperforms other communication libraries on NVIDIA hardware [17].

C. Communication Observability and NCCL Wait States

While aggregate profiling (e.g., step time, communication vs. computation splits) can indicate that communication dominates scaling behavior, it often cannot explain *why* a particular collective scales poorly. Traditional profilers like NVIDIA Nsight Systems capture NCCL API call timing but cannot trace internal memory transfers within NCCL kernels because NCCL uses direct load/store operations rather than traced CUDA APIs.

NCCL Profiler v4, introduced in NCCL 2.27, addresses this gap by exposing NET-level events that capture when NCCL communication threads are blocked [18]. These events include:

- **RecvWait:** Time spent waiting to receive data from peer GPUs, indicating inter-rank data dependencies.
- **SendGpuWait:** Time spent waiting for the local GPU to be ready to send data, indicating local resource contention.
- **SendPeerWait:** Time spent waiting for peer readiness to receive data.

In this paper, we interpret these events as *software-induced synchronization* within collectives. This enables answering not only which primitives dominate overhead (RQ2), but also which stall mechanisms dominate those primitives (RQ3), providing a mechanical explanation for scaling inefficiency.

III. METHODOLOGY

We employ Design Science Research Methodology (DSRM) [19] combined with a controlled laboratory experiment. DSRM provides a structured framework for creating and evaluating IT artifacts; we instantiate it through three solution objectives: (O1) restore NCCL communication events from profiler traces, (O2) aggregate and quantify communication primitives, and (O3) extract fine-grained wait-state information to identify synchronization bottlenecks.

A. Experimental Setup

We conducted experiments training Llama 3.1 70B on an HPC cluster managed by SLURM, using NVIDIA H200 GPUs (141 GB HBM3 each) connected via 400 Gb/s InfiniBand (8 NICs per node). Two configurations were evaluated: 32 GPUs (4 nodes $\times$ 8 GPUs) and 64 GPUs (8 nodes $\times$ 8 GPUs). Intra-node communication uses NVLink, providing high-bandwidth, low-latency transfers within NUMA domains. Inter-node communication traverses the InfiniBand fabric via PCIe-attached NICs.

We used an NVIDIA NeMo reference recipe for Llama 3.1 training [20], which provides a standardized workload enabling controlled scaling experiments. The software stack includes NCCL 2.27.3, CUDA 12.9.1, PyTorch 2.8.0, NeMo 2.5.0, and Megatron-Core 0.14.0. We follow the reference configuration for model architecture (80 layers, hidden dimension 8192, 64 attention heads) and keep precision fixed at BF16, which is standard for LLM training as it reduces communication volume while maintaining numerical stability [3].

To isolate the scaling effect on communication, we keep tensor-parallel (TP=4) and pipeline-parallel (PP=4) dimensions constant while scaling data parallelism from DP=1 (32 GPUs) to DP=2 (64 GPUs). Context parallelism remains fixed at CP=2. This weak-scaling design doubles the global batch size proportionally with GPU count, maintaining constant per-GPU workload.

Each experiment runs for 40 iterations. We treat the initial iterations as warm-up (covering CUDA graph capture, kernel compilation, and communication initialization) and activate system-level profiling only after iteration 20. This ensures measurements reflect steady-state training behavior rather than initialization overhead.

B. Experimental Design: Variables and Hypotheses

The independent variable is *system scale*, operationalized as GPU count (32 vs. 64). The choice of these specific scale points enables analysis of the transition from single-node-dominated communication (32 GPUs on 4 nodes) to multi-node-dominated communication (64 GPUs on 8 nodes), where inter-node collective traffic increases.

Dependent variables align with our research questions. For RQ1, we measure interconnect bandwidth utilization via InfiniBand hardware counters and the communication-vs.-computation time split via Nsight Systems. For RQ2, we measure per-primitive durations (AllReduce, AllGather, ReduceScatter, P2P) from NCCL Profiler traces. For RQ3, we measure NCCL NET-level wait states (RecvWait, SendGPUWait, SendPeerWait) and compute wait-to-active ratios.

To ensure internal validity, we hold confounding variables constant: model architecture, sequence length (8192 tokens), precision (BF16), parallelism configuration (TP/PP/CP), random seeds, and cluster partition. Run-to-run variability due to system noise (OS jitter, network contention) is mitigated by profiling multiple iterations and reporting aggregated statistics.

We test three hypotheses derived from prior observations of underutilized networks and synchronization-driven scaling limits:

- **H1:** Communication overhead increases from 32 to 64 GPUs while interconnect utilization remains below the theoretical peak.
- **H2:** Collective communication (especially All-Reduce and Reduce-Scatter) grows disproportionately compared to P2P when scaling model-parallel training.
- **H3:** NCCL Profiler wait states reveal software-induced synchronization stalls as a dominant contributor to collective overhead.

TABLE I
TRAINING AND PARALLELISM CONFIGURATION

Parameter	32 GPUs	64 GPUs
Model	Llama 3.1 70B	
Precision	BF16	
Sequence length	8192	
Tensor parallelism (TP)	4	4
Pipeline parallelism (PP)	4	4
Context parallelism (CP)	2	2
Data parallelism (DP)	1	2
Global batch size	64	128

C. Measurement and Data Collection

To connect high-level scaling behavior with low-level root causes, we combine three complementary telemetry sources, summarized in Table II.

InfiniBand Counters. We sample hardware performance counters on each InfiniBand NIC to obtain time series of transmit (TX) and receive (RX) bandwidth. Counters are read at 100 ms intervals throughout training. We aggregate across links and nodes to capture effective scale-out utilization and determine whether observed slowdown could plausibly be attributed to hardware saturation.

NVIDIA Nsight Systems. We decompose GPU execution time into compute and communication components using Nsight Systems profiling. This captures CUDA kernel timings and NCCL API calls, providing a bridge between step-level scaling behavior (RQ1) and communication-layer analyses (RQ2/RQ3). Nsight profiles are collected from one representative node (8 GPUs) per experiment.

NCCL Profiler v4. We enable NCCL Profiler v4 [18] to obtain fine-grained, per-operation traces and NET-level wait states. The profiler integrates via a shared library that hooks into NCCL’s internal callbacks. Unlike Nsight Systems, which captures API-level timing, NCCL Profiler exposes operation-internal behavior including channel utilization and synchronization stalls. Traces are collected from all 32 or 64 ranks, generating multi-gigabyte JSON files that the artifact processes offline.

Experimental Workflow. The experimental workflow proceeds in four stages: (1) environment initialization (Python

3.11, dependencies, container setup via Enroot), (2) configuration loading and environment variable setup, (3) training execution via SLURM job submission with integrated profiling, and (4) data collection, archival, and validation. No manual intervention is required after job submission. To ensure measurement reliability, both experiments execute on identical cluster partitions under equivalent system load conditions. Random seeds and job scheduling parameters are fixed to ensure reproducibility.

TABLE II
TELEMETRY SOURCES AND THEIR ROLE IN THE ANALYSIS

Source	Signal and use
InfiniBand counters	RX/TX throughput over time; used to quantify interconnect utilization and verify whether hardware saturation explains overhead (RQ1).
Nsight Systems	Step-time decomposition into compute vs. communication; used to quantify communication-driven scaling overhead and establish the communication fraction (RQ1).
NCCL Profiler v4	Per-operation timings and NET-level waits (RecvWait, SendGpuWait, SendPeerWait); used to attribute overhead to specific primitives (RQ2) and synchronization mechanisms (RQ3).

D. Profiling Artifact

The artifact processes NCCL Profiler v4 output to reconstruct collective and P2P operations by pairing begin/end records using composite keys (rank, thread, operation). This design ensures correct reconstruction even under heavy interleaving of concurrent collective operations across multiple CUDA streams.

Communication events are aggregated via SQL queries, and NET-level wait states (RecvWait, SendGpuWait, SendPeerWait) are extracted and linked to parent collectives. The artifact is designed around three functional requirements derived from the research questions: (R1) restoring operation boundaries from interleaved traces, (R2) aggregating overhead by primitive type, and (R3) attributing overhead to specific wait-state mechanisms.

The implementation consists of three layers. The *Data Ingestion Layer* loads NCCL Profiler traces in JSON (Chrome Trace Event) format and transforms them into structured event representations. During ingestion, communication operations are reconstructed by pairing their begin and end records using Algorithm 1. The *Data Extraction Layer* stores pre-processed events in an SQLite database, exposing data through a declarative query interface for filtering, aggregation, and joining. The *Analysis Layer* consumes extracted data and provides statistical summaries and visualizations.

Mapping wait states to collectives is performed by joining NET events to the enclosing collective interval on the same rank and thread. Because NCCL operations communicate

Algorithm 1 Reconstructing NCCL operations from Profiler v4 events

```

1: Input: trace events  $E$  with fields (rank, thread, name, ts, phase)
2: Output: reconstructed operations  $O$  with durations and attributes
3: Initialize hash map  $M$  from composite key  $k$  to start event
4: for all event  $e \in E$  in timestamp order do
5:    $k \leftarrow (e.rank, e.thread, e.name)$ 
6:   if  $e.phase$  is “begin” then
7:      $M[k] \leftarrow e$ 
8:   else if  $e.phase$  is “end” and  $k \in M$  then
9:      $s \leftarrow M[k]$ 
10:    Emit operation  $o$  with  $o.start = s.ts$ ,  $o.end = e.ts$ ,  $o.dur = e.ts - s.ts$ 
11:    Delete  $M[k]$ 
12:   end if
13: end for
14: Store  $O$  in relational table; join with NET events for wait attribution.

```

across multiple parallel channels that execute simultaneously, raw wait times must be normalized. We divide aggregate wait time by the average number of channels per operation instance (16.00 channels for Reduce-Scatter, 4.74 channels for All-Reduce) to obtain effective per-operation wait time. The wait-to-active ratio $W/A = t_{wait}/t_{active}$ provides a measure of time spent stalling relative to productive data movement.

TABLE III
ARTIFACT CAPABILITIES DERIVED FROM RESEARCH QUESTIONS

Requirement	Artifact capability
R1 (RQ2)	Reconstruct and aggregate AllReduce, AllGather, ReduceScatter, and P2P events by pairing begin/end records with composite keys.
R2 (RQ3)	Extract NET-level wait events and map them to parent collectives via interval joins to quantify synchronization stalls.
R3 (Scale)	Process multi-rank traces (up to 1M+ events) without manual pairing or per-rank inspection.

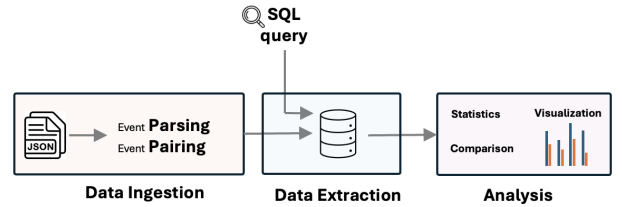


Fig. 1. Artifact architecture: Data Ingestion reconstructs operations from JSON traces, Data Extraction stores events in SQLite with query interface, Analysis Layer produces statistics and visualizations.

Across both runs, the artifact reconstructs all communication operations without unpaired events, yielding 95,512 valid operation pairs for 32 GPUs and 1,000,178 pairs for 64 GPUs. The absence of unpaired events validates the reconstruction algorithm’s correctness under high-concurrency conditions.

IV. RESULTS

A. RQ1: Interconnect Utilization

The average iteration time increases slightly from 14.49 s at 32 GPUs to 14.59 s at 64 GPUs (0.7% increase). While the per-iteration difference appears small, the total GPU time across profiled iterations shows clearer trends. Decomposition using Nsight Systems (Table IV) reveals that of the 8.76 s increase in total GPU time when scaling to 64 GPUs, 6.92 s (79%) stems from communication while only 1.85 s (21%) originates from computation. Thus, the scaling inefficiency is predominantly communication-driven.

TABLE IV
NSIGHT SYSTEMS STEP-TIME DECOMPOSITION (AGGREGATED ACROSS PROFILED ITERATIONS 20–40)

Metric	32 GPUs	64 GPUs	Delta
Total GPU time (s)	471.88	480.64	+8.76
Communication time (s)	150.85	157.77	+6.92
Computation time (s)	321.03	322.88	+1.85
Communication fraction	31.97%	32.82%	+0.85pp

To determine whether this communication overhead originates from hardware limitations, we analyze interconnect bandwidth utilization. Figure 2 shows receive-side bandwidth over time for both configurations. The 32-GPU configuration exhibits a brief initialization peak of approximately 15 Gb/s, followed by stable utilization around 6 Gb/s during training steps. The 64-GPU configuration shows periodic bursts reaching 50–60 Gb/s as training progresses, but remains far below the theoretical 400 Gb/s capacity.

At peak, bandwidth utilization is only 15% of available capacity. Transmit-side bandwidth follows a similar pattern with bursts up to 60 Gb/s at 64 GPUs. The bursty utilization at 64 GPUs suggests that communication occurs in distinct phases aligned with model-parallel collectives—tensor-parallel reductions during forward/backward passes create periodic traffic spikes rather than sustained load. If the interconnect were the primary limiter, we would expect sustained plateaus near peak bandwidth with increasing queueing delay at scale. Instead, the interconnect remains underutilized while communication time increases, indicating the bottleneck is not hardware saturation. This validates hypothesis H1 and motivates investigating software-layer causes.

B. RQ2: Communication Primitives

Table V presents the communication primitive breakdown extracted via the artifact. When scaling from 32 to 64 GPUs, total communication time increases by 77.8% (from 187,147 ms to 332,838 ms). However, this growth is not uniformly distributed.

Collective communication increases $3.0\times$ (from 44,340 ms to 133,131 ms), while P2P increases only $1.4\times$ (from 142,807 ms to 199,708 ms). Consequently, the collective share rises from 23.7% at 32 GPUs to 40.0% at 64 GPUs. Of

TABLE V
COMMUNICATION PRIMITIVES ANALYSIS (AGGREGATED ACROSS ALL RANKS)

Operation	32 GPUs		64 GPUs		Scale
	Tot (ms)	Avg (μ s)	Tot (ms)	Avg (μ s)	
P2P	142,807	68.7	199,708	48.4	$1.40\times$
Collectives	44,340	22.0	133,131	32.5	$3.00\times$
AllGather	35,390	53.95	38,969	32.89	$1.10\times$
AllReduce	6,262	5.80	44,182	20.31	$7.06\times$
ReduceScatter	2,292	10.61	49,849	74.18	$21.75\times$

the 145,691 ms of *additional* communication overhead at 64 GPUs, collectives account for 88,791 ms (60.9%).

The scaling inefficiency is highly operation-specific. All-Reduce increases $7.1\times$ and Reduce-Scatter increases $21.7\times$, while All-Gather shows modest growth of only $1.1\times$. At 64 GPUs, All-Reduce and Reduce-Scatter together account for 70.6% of all collective overhead (94,031 ms of 133,131 ms).

The average duration per event provides additional insight. P2P average duration *decreases* from 68.7μ s to 48.4μ s at scale, suggesting efficient handling of pipeline parallelism traffic. In contrast, Reduce-Scatter average duration increases $7\times$ from 10.61μ s to 74.18μ s, indicating amplified synchronization cost as the number of participating ranks grows.

This asymmetry is consistent with the role of these collectives in tensor parallelism: All-Reduce and Reduce-Scatter synchronize gradients and partial results across TP groups, introducing strict barriers. All-Gather, which reassembles split activations, benefits from more regular data exchange patterns. The efficient scaling of P2P and All-Gather serves as an internal control, demonstrating the inefficiency is operation-specific rather than a general hardware saturation effect. This validates hypothesis H2.

Event Count Analysis. Beyond total time, examining event counts reveals how operation frequency changes with scale. P2P event count doubles from 2,080 to 4,128 (proportional to scale), but collective event count also doubles from 2,016 to 4,096. The key difference is in duration growth: while P2P events become faster at scale (average duration decreases 30%), collective events become slower. This suggests that P2P operations (used for pipeline-parallel activation passing) benefit from increased parallelism, while collective operations (used for tensor-parallel synchronization) suffer from increased coordination overhead as more ranks participate.

C. RQ3: Software Stall Analysis

To understand *why* All-Reduce and Reduce-Scatter scale poorly, we analyze their internal wait-state composition. Table VI presents the raw wait-state breakdown at 64 GPUs, and Table VII presents normalized metrics.

Because NCCL operations execute across multiple parallel channels simultaneously, raw wait times must be normalized by channel count. The wait-to-active (W/A) ratio $W/A = t_{wait}^{norm}/t_{active}$ measures time spent stalling relative to productive work.



Fig. 2. RX bandwidth utilization during training. The 64-GPU configuration shows periodic bursts up to 60 Gb/s (15% of 400 Gb/s capacity), while 32 GPUs maintains steady 6 Gb/s.

TABLE VI
RAW WAIT-STATE COMPOSITION AT 64 GPUS

Collective	RecvW (ms)	SndGpuW (ms)	SndPeerW (ms)	Total (ms)
ReduceScatter	318,412	0	0	318,412
AllReduce	117,763	113,145	4,720	235,628

TABLE VII
NORMALIZED WAIT-STATE ANALYSIS AT 64 GPUS

Collective	Active (ms)	Wait (ms)	W/A	RecvW (%)	SndGW (%)
ReduceScatter	49,849	8,344	0.17	100	0
AllReduce	44,182	21,801	0.49	50	48

All-Reduce exhibits the most severe inefficiency with a W/A ratio of 0.49, indicating that for every two units of active data transfer, nearly one additional unit is spent in software wait states. The stall mechanism is bidirectional, split between RecvWait (50%) and SendGpuWait (48%). This represents a dual-sided barrier where GPUs wait for both peer readiness (RecvWait—waiting to receive data from remote ranks) and local GPU readiness (SendGpuWait—waiting for the local GPU to complete prior work before launching the send). This behavior aligns with All-Reduce’s theoretical synchronization requirements: all ranks must complete their local reductions before the result can be distributed.

Reduce-Scatter shows a specialized profile with 100% RecvWait and 0% SendGpuWait. This unidirectional receive-side barrier indicates that ranks are exclusively blocked waiting for peer data, not local resources. The pattern is consistent

with Reduce-Scatter’s function in gradient aggregation: each rank must receive partial gradient contributions from all peers before it can complete its output shard.

These findings validate hypothesis H3: the communication overhead originates from measurable, software-induced synchronization stalls rather than hardware saturation. Critically, these stalls occur despite the interconnect operating at only 15% capacity, confirming that the synchronization and scheduling mechanisms within NCCL’s collective algorithms—specifically the requirement that all GPUs reach the same execution point—are the limiting factor.

Comparing 32 vs 64 GPU Wait Behavior. While our detailed wait-state analysis focuses on the 64-GPU configuration where overhead is most pronounced, comparison with 32-GPU behavior is instructive. At 32 GPUs, collective operations account for only 23.7% of communication time, and the tensor-parallel groups (TP=4) operate entirely within single nodes via NVLink. At 64 GPUs, the introduction of DP=2 creates cross-node collective traffic, and the collective share rises to 40%. The dramatic increase in Reduce-Scatter duration (21.7 $\times$) and its pure RecvWait profile suggest that cross-node gradient synchronization introduces substantial waiting for remote data arrival. This architectural transition from NVLink-dominated to InfiniBand-involving collectives appears to amplify synchronization overhead disproportionately.

V. DISCUSSION

Our findings validate all three hypotheses and provide new insights into the nature of communication bottlenecks in distributed LLM training.

Hardware vs. Software Bottlenecks. The combination of RQ1–RQ3 results establishes that communication inefficiencies in model-parallel Llama 3.1 70B training are software-induced rather than hardware-limited. At 64 GPUs, communication time increases by 4.6% while interconnect utilization peaks at only 15% of capacity. This paradox—increased communication overhead alongside underutilized hardware—is resolved by the wait-state analysis showing that collective operations spend substantial time in synchronization stalls. The All-Reduce W/A ratio of 0.49 means that for every second of active data transfer, an additional half-second is spent waiting for peer or GPU readiness. This pattern is consistent with prior observations of 63% GPU idle time during large-scale training [6], but our analysis attributes this idle time to specific NCCL-internal mechanisms.

The underutilization we observe (15% of bandwidth capacity) represents significant inefficiency at scale. For perspective, if training efficiency scaled linearly with hardware investment, the 64-GPU configuration should complete iterations faster than the 32-GPU configuration. Instead, iteration time increases slightly, and the interconnect remains mostly idle. This suggests that simply adding more GPUs or faster interconnects will not resolve the bottleneck—the software coordination overhead must be addressed.

Operation-Specific Scaling Behavior. The $3\times$ increase in collective time versus $1.4\times$ for P2P shifts communication composition from 24% collectives at 32 GPUs to 40% at 64 GPUs. Importantly, this shift is driven by two specific operations: All-Reduce ($7.1\times$) and Reduce-Scatter ($21.7\times$), while All-Gather maintains near-constant overhead ($1.1\times$). This operation-specific behavior suggests that optimization efforts should target synchronization-heavy collectives rather than general communication infrastructure.

The different wait-state profiles of All-Reduce (bidirectional: 50% RecvWait, 48% SendGpuWait) and Reduce-Scatter (unidirectional: 100% RecvWait) indicate distinct bottleneck mechanisms. All-Reduce’s bidirectional stalls suggest both inter-rank coordination and local GPU scheduling contribute to overhead—ranks must wait for peers to be ready while also waiting for their own GPU to complete prior work. Reduce-Scatter’s receive-only stalls indicate that ranks wait exclusively on peer data dependencies, not local resource availability. These mechanistic differences may require different optimization strategies: All-Reduce might benefit from better local GPU scheduling and pipelining, while Reduce-Scatter might benefit from relaxed synchronization or speculative execution.

Implications for System Design. These findings have practical implications for both training framework developers and system architects. First, increasing interconnect bandwidth alone will not resolve scaling inefficiencies when the bottleneck is software synchronization—the interconnect is already operating at 15% capacity. Investing in 800 Gb/s or faster interconnects would yield diminishing returns unless the software-level synchronization overhead is addressed.

Second, algorithmic innovations that reduce synchroniza-

tion points or enable overlap between communication and computation may yield larger gains than hardware upgrades. Techniques such as gradient compression, asynchronous SGD, or overlapped communication-computation scheduling directly target the identified bottleneck. Our wait-state decomposition suggests that reducing RecvWait (the dominant component in both All-Reduce and Reduce-Scatter) should be the primary focus.

Third, profiling at the NCCL-internal level (as enabled by our artifact) is necessary to diagnose root causes; aggregate profiling at the step-time or kernel level hides the operation-specific nature of the bottleneck. Without the wait-state breakdown, one might incorrectly attribute overhead to general “communication” rather than specific synchronization mechanisms within particular collective operations.

Fourth, the transition from 32 to 64 GPUs crosses an architectural boundary: at 32 GPUs (4 nodes), most tensor-parallel communication can use fast intra-node NVLink, while at 64 GPUs, the DP=2 configuration introduces cross-node collective traffic via slower InfiniBand. This transition amplifies wait times disproportionately, suggesting that parallelism configurations should be topology-aware and minimize cross-node collective traffic where possible.

Potential optimization directions include: (1) asynchronous collective algorithms that reduce strict barrier requirements, (2) improved overlap between computation and communication phases, (3) topology-aware collective scheduling that reduces wait times in hierarchical networks, and (4) compression techniques that reduce message sizes and thus synchronization frequency. Because our study focuses on dense, synchronous transformer training, the dominant communication bottlenecks may differ in systems with other architectural or coordination choices. This should be considered when assessing the generalizability of our findings.

Artifact Utility. The analysis artifact fulfilled its design objectives derived from the Design Science Research methodology. It successfully reconstructed 95,512 operation pairs at 32 GPUs and over 1 million pairs at 64 GPUs without data loss, validating the reconstruction algorithm’s correctness under high-concurrency conditions. The SQL-based query interface enabled systematic extraction of per-primitive statistics and wait-state attribution, supporting reproducible analysis that can be applied to future experiments or different workloads.

By automating trace processing, the artifact eliminated the need for manual per-rank inspection, which would be infeasible at the scale of traces generated (multi-gigabyte JSON files per run containing interleaved events from dozens of concurrent threads). The artifact’s three-layer architecture (ingestion, extraction, analysis) provides a template for similar NCCL profiling workflows. The ingestion layer handles the complexity of pairing begin/end events under heavy interleaving; the extraction layer provides a stable relational schema for analysis; and the analysis layer supports both manual SQL queries and automated statistical summaries.

One limitation is that the current implementation processes traces offline after training completes; real-time profiling in-

tegration would enable adaptive optimization during training, potentially adjusting collective algorithms or batch sizes based on observed wait patterns. Future versions could also incorporate machine learning models to predict wait-state behavior and recommend configuration changes.

Generalizability Considerations. While our results are specific to Llama 3.1 70B on H200 GPUs, several findings likely generalize to other LLM training scenarios. The pattern of underutilized interconnects combined with software-dominated overhead has been observed in other studies [6], [8], suggesting this is a common phenomenon rather than an artifact of our specific configuration. The disproportionate scaling of All-Reduce and Reduce-Scatter reflects their fundamental synchronization requirements (all ranks must reach a barrier before proceeding) rather than implementation-specific behavior—any collective algorithm implementing these semantics will face similar challenges at scale.

However, the specific W/A ratios and wait-state distributions may vary with different NCCL versions (which may implement different collective algorithms), hardware topologies (e.g., fat-tree vs. dragonfly networks), or parallelism configurations (different TP/PP/CP/DP ratios). Studies on different architectures would help establish the boundaries of these findings and identify which aspects are fundamental versus configuration-dependent.

VI. THREATS TO VALIDITY

Internal validity. We hold model configuration, precision, sequence length, and parallelism strategy constant across runs; however, distributed training exhibits run-to-run variability due to system noise (OS jitter, network contention, thermal throttling). We mitigate warm-up effects by enabling system-level profiling after iteration 20 and report statistics aggregated across multiple iterations. Fixed random seeds and cluster partition reduce scheduling variability.

Construct validity. Bandwidth utilization is derived from InfiniBand hardware counters and captures aggregate link throughput, not per-flow congestion or application-level effective bandwidth. NCCL wait states measure NCCL-internal blocking; they are a proxy for synchronization overhead but do not uniquely identify root causes (e.g., load imbalance vs. algorithmic choice vs. network congestion). The W/A ratio assumes that wait time and active time are separable, which may not hold for pipelined implementations.

External validity. Results are specific to Llama 3.1 70B, NeMo/Megatron-style 4D parallelism, NVIDIA H200 GPUs, 400 Gb/s InfiniBand, and NCCL 2.27.3. Other models with different layer counts, parallelism configurations with different TP/PP ratios, alternative interconnects (e.g., Slingshot, RoCE), or different NCCL versions may exhibit different scaling behavior. The two-point scaling study (32 and 64 GPUs) limits generalization to larger scales; behavior at 128+ GPUs remains unevaluated. Additionally, we focus on the training phase; inference workloads with different communication patterns may show different characteristics.

Reproducibility. The experimental setup uses publicly available components: NVIDIA NeMo reference recipes from NGC Catalog, standard NCCL Profiler v4 instrumentation, and InfiniBand hardware counters accessible via standard Linux interfaces. The artifact’s SQL-based analysis approach is reproducible given the trace data. However, the specific H200 cluster configuration may not be universally available, limiting exact replication.

VII. RELATED WORK

Communication Bottlenecks in Distributed Training. Prior work has characterized communication as a major bottleneck in distributed training [6], [7]. Shah et al. [6] reported that GPUs spend up to 63% of execution time idle during large-scale model training. Dai et al. [8] demonstrated interconnect underutilization during LLM training and called for investigation at larger scales. Our work extends these findings by attributing idle periods to specific NCCL-internal wait states (RecvWait, SendGpuWait) and providing fine-grained breakdown of collective operation behavior.

NCCL and Collective Communication. NCCL is the dominant communication library for NVIDIA GPU clusters [2], [21]. Prior profiling studies [10], [11] have characterized NCCL performance but treated it largely as a black box, measuring end-to-end operation latency without visibility into internal mechanisms. Soyurk et al. [11] identified the need for fine-grained NCCL observability but lacked access to internal wait-state metrics. NCCL Profiler v4 [18], introduced in NCCL 2.27, provides the NET-level instrumentation that enables our analysis. Our work is among the first to leverage this capability for systematic study of collective behavior under realistic LLM workloads.

Parallelism Strategies. Tensor parallelism [4], [13], pipeline parallelism [14], and data parallelism [12] have been extensively studied for enabling training of large models. Korthikanti et al. [16] proposed sequence parallelism and selective activation recomputation to reduce memory footprint and communication in tensor parallelism. Our work complements these algorithmic contributions by providing empirical evidence of which communication primitives contribute most to overhead under 4D parallelism, identifying All-Reduce and Reduce-Scatter as the dominant bottlenecks. This suggests that future optimization efforts should prioritize these specific operations.

Comparison with Prior Work. Compared to prior studies that observe communication overhead at the step-time or kernel level, we focus on reconstructing and relating NCCL-internal wait states to parent collectives under a realistic, model-parallel LLM workload. This bridges the gap between high-level training performance observations and low-level, operation-specific synchronization behavior that ultimately determines scaling efficiency.

Profiling Methodologies. Our dual-profiling approach (Nsight Systems for system-level decomposition, NCCL Profiler for communication-layer detail) provides complementary perspectives. Nsight Systems cannot trace NCCL internal

transfers (which use direct load/store rather than traced CUDA APIs), while NCCL Profiler lacks the system-wide context. Combining them enables end-to-end attribution from step time to specific wait-state mechanisms. Similar multi-layer profiling strategies may be necessary for other distributed systems where performance depends on interactions across multiple software layers.

VIII. CONCLUSION

This paper investigated the communication behavior of distributed LLM training, examining whether communication overhead originates from hardware bandwidth saturation or software-level synchronization. Through a controlled scaling study of Llama 3.1 70B training from 32 to 64 GPUs, we demonstrate that communication inefficiencies originate primarily from software-induced synchronization barriers within NCCL collective operations rather than from hardware bandwidth limitations.

Our key findings are: (1) At 64 GPUs, the interconnect operates at only 15% of its 400 Gb/s capacity while communication time increases, confirming the bottleneck is not hardware saturation. (2) Collective operations scale disproportionately ($3\times$) compared to P2P ($1.4\times$), with All-Reduce and Reduce-Scatter exhibiting $7\times$ and $22\times$ increases respectively and accounting for 71% of collective overhead. (3) Wait-state analysis reveals All-Reduce spends 49% of time in synchronization stalls (bidirectional: 50% RecvWait, 48% SendGpuWait), while Reduce-Scatter exhibits unidirectional receive-only barriers (100% RecvWait).

The developed profiling artifact addresses the observability gap for NCCL behavior by processing NCCL Profiler v4 data to reconstruct operations and attribute overhead to specific wait-state mechanisms. By enabling systematic detection of software-induced inefficiencies, this approach contributes to more effective diagnosis and optimization of distributed training workloads.

For our own work, the most important next step is to move from diagnosis to intervention by using the artifact to evaluate concrete optimizations for the two dominant bottlenecks, All-Reduce and Reduce-Scatter, in the 64-GPU configuration. Future work should then test larger scales, additional models, different system settings, and inference workloads to assess how robust and generalizable the observed wait-state patterns are. Such comparisons would help distinguish effects that are specific to the present Llama 3.1 training setup from effects that persist across broader classes of large-scale AI systems. Future work should also relate these results to alternative system designs, such as pipeline-heavy configurations or sparsely activated models, where different communication patterns may dominate.

The artifact could be extended to incorporate Nsight Systems data for end-to-end observability spanning kernel execution and communication, enabling correlation between computation variability and wait-state duration. Real-time profiling integration would enable adaptive optimization during training, potentially adjusting collective algorithms or scheduling based

on observed wait patterns. Finally, extending the methodology to inference workloads, which have different communication patterns (primarily All-Gather for weight distribution), would broaden the applicability of these findings.

ACKNOWLEDGMENT

The authors acknowledge the use of AI-assisted tools during the preparation of this work. Perplexity and Consensus were used to support literature search and identify relevant prior work on distributed training, communication bottlenecks, and parallelism strategies. ChatGPT was used for language refinement, including improvements to clarity, grammar, structure, and academic tone. Claude Sonnet 4.5, used via GitHub Copilot Enterprise, was used to assist code debugging during the Llama 3.1 training workflow and to draft scripts for the bandwidth-utilization analysis and the corresponding figures. All AI-assisted outputs were reviewed, validated, and edited by the authors; the experimental design, data collection, analysis, interpretation of results, and final wording remain the responsibility of the authors.

REFERENCES

- [1] A. Sapio, M. Canini, C.-Y. Ho, J. Nelson, P. Kalnis, C. Kim, A. Krishnamurthy, M. Moshref, D. Ports, and P. Richtarik, "Scaling distributed machine learning with In-Network aggregation," in *18th USENIX Symposium on Networked Systems Design and Implementation (NSDI 21)*. USENIX Association, Apr. 2021, pp. 785–808. [Online]. Available: <https://www.usenix.org/conference/nsdi21/presentation/sapio>
- [2] A. Weingram, Y. Li, H. Qi, D. Ng, L. Dai, and X. Lu, "xCCL: A Survey of Industry-Led Collective Communication Libraries for Deep Learning," *Journal of Computer Science and Technology*, vol. 38, no. 1, pp. 166–195, Feb. 2023. [Online]. Available: <https://doi.org/10.1007/s11390-023-2894-6>
- [3] A. Grattafiori, A. Dubey, A. Jauhri *et al.*, "The Llama 3 Herd of Models," Nov. 2024, arXiv:2407.21783 [cs]. [Online]. Available: <http://arxiv.org/abs/2407.21783>
- [4] M. Shoenybi, M. Patwary, R. Puri, P. LeGresley, J. Casper, and B. Catanzaro, "Megatron-LM: Training Multi-Billion Parameter Language Models Using Model Parallelism," Mar. 2020, arXiv:1909.08053 [cs]. [Online]. Available: <http://arxiv.org/abs/1909.08053>
- [5] J. Duan, S. Zhang, Z. Wang, L. Jiang, W. Qu, Q. Hu, G. Wang, Q. Weng, H. Yan, X. Zhang, X. Qiu, D. Lin, Y. Wen, X. Jin, T. Zhang, and P. Sun, "Efficient Training of Large Language Models on Distributed Infrastructures: A Survey," Jul. 2024, arXiv:2407.20018 [cs]. [Online]. Available: <http://arxiv.org/abs/2407.20018>
- [6] A. Shah, V. Chidambaram, M. Cowan, S. Maleki, M. Musuvathi, T. Mytkowicz, J. Nelson, O. Saarikivi, and R. Singh, "TACCL: Guiding collective algorithm synthesis using communication sketches," in *20th USENIX Symposium on Networked Systems Design and Implementation (NSDI 23)*. Boston, MA: USENIX Association, Apr. 2023, pp. 593–612. [Online]. Available: <https://www.usenix.org/conference/nsdi23/presentation/shah>
- [7] Z. Zhang, C. Chang, H. Lin, Y. Wang, R. Arora, and X. Jin, "Is Network the Bottleneck of Distributed Training?" in *Proceedings of the Workshop on Network Meets AI & ML*. Virtual Event USA: ACM, Aug. 2020, pp. 8–13. [Online]. Available: <https://dl.acm.org/doi/10.1145/3405671.3405810>
- [8] L. Dai, H. Qi, W. Chen, and X. Lu, "High-speed data communication with advanced networks in large language model training," *IEEE Micro*, vol. 44, no. 2, pp. 31–40, 2024.
- [9] W. Li, X. Liu, Y. Li, Y. Jin, H. Tian, Z. Zhong, G. Liu, Y. Zhang, and K. Chen, "Understanding communication characteristics of distributed training," in *Proceedings of the 8th Asia-Pacific Workshop on Networking*, ser. APNet '24. New York, NY, USA: Association for Computing Machinery, 2024, pp. 1–8. [Online]. Available: <https://doi.org/10.1145/3663408.3663409>

- [10] A. A. Awan, A. Jain, C.-H. Chu, H. Subramoni, and D. K. Panda, "Communication profiling and characterization of deep learning workloads on clusters with high-performance interconnects," in *2019 IEEE Symposium on High-Performance Interconnects (HOTI)*, 2019, pp. 49–53.
- [11] M. A. Soytürk, P. Akhtar, E. Tezcan, and D. Unat, "Monitoring collective communication among gpus," in *Euro-Par 2021: Parallel Processing Workshops*. Cham: Springer International Publishing, 2022, pp. 41–52.
- [12] S. Rajbhandari, J. Rasley, O. Ruwase, and Y. He, "Zero: memory optimizations toward training trillion parameter models," in *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, ser. SC '20. IEEE Press, 2020.
- [13] D. Narayanan, M. Shoeybi, J. Casper, P. LeGresley, M. Patwary, V. Korthikanti, D. Vainbrand, P. Kashinkunti, J. Bernauer, B. Catanzaro, A. Phanishayee, and M. Zaharia, "Efficient large-scale language model training on gpu clusters using megatron-lm," in *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, ser. SC '21. New York, NY, USA: Association for Computing Machinery, 2021. [Online]. Available: <https://doi.org/10.1145/3458817.3476209>
- [14] Y. Huang, Y. Cheng, A. Bapna, O. Firat, M. X. Chen, D. Chen, H. Lee, J. Ngiam, Q. V. Le, Y. Wu, and Z. Chen, *GPipe: efficient training of giant neural networks using pipeline parallelism*. Red Hook, NY, USA: Curran Associates Inc., 2019.
- [15] J. Li, B. Nicolae, J. Wozniak, and G. Bosilca, "Understanding scalability and fine-grain parallelism of synchronous data parallel training," in *2019 IEEE/ACM Workshop on Machine Learning in High Performance Computing Environments (MLHPC)*, 2019, pp. 1–8.
- [16] V. Korthikanti, J. Casper, S. Lym, L. McAfee, M. Andersch, M. Shoeybi, and B. Catanzaro, "Reducing activation recomputation in large transformer models," in *Proceedings of Machine Learning and Systems*, vol. 5. MLSys, 2023, pp. 415–434. [Online]. Available: https://proceedings.mlsys.org/paper_files/paper/2023/file/80083951326cf5b35e5100260d64ed81-Paper-mlsys2023.pdf
- [17] A. Castelló, E. S. Quintana-Ortí, and J. Duato, "Accelerating distributed deep neural network training with pipelined MPI allreduce," *Cluster Computing*, vol. 24, no. 4, pp. 3797–3813, Dec. 2021. [Online]. Available: <https://doi.org/10.1007/s10586-021-03370-9>
- [18] NVIDIA, "Nccl profiler plugin v4," <https://github.com/NVIDIA/nccl/tree/v2.27.7-1/ext-profiler>, 2025, accessed: 2025-10-22.
- [19] K. Peffers, T. Tuunanen, M. Rothenberger, and S. Chatterjee, "A design science research methodology for information systems research," *Journal of Management Information Systems*, vol. 24, pp. 45–77, 01 2007.
- [20] "GPU-optimized AI, Machine Learning, & HPC Software | NVIDIA NGC." [Online]. Available: <https://catalog.ngc.nvidia.com/orgs/nvidia/teams/dgxc-benchmarking/resources/llama31-70b-dgxc-benchmarking-a?version=24.11.1>
- [21] NVIDIA, "Nccl," <https://github.com/NVIDIA/nccl/tree/v2.27.7-1>, 2025, accessed: 2025-10-24.

LLM-Based Knowledge Graph Construction for Curriculum Analysis

Automated Skill Extraction and ESCO Alignment

1st Florian Pallmann

Computer Science

DHBW Heidenheim

Heidenheim, Germany

<https://orcid.org/0009-0005-4898-4728>

2nd Franziska Schütz

Computer Science

DHBW Heidenheim

Heidenheim, Germany

franziska.schuetz@dhbw-heidenheim.de

Abstract—Manual competency extraction from module descriptions is feasible but does not scale to large curricula or multi-institutional settings such as EU4Dual, a European dual-study alliance. This paper presents an automated pipeline using Large Language Models (LLMs) for skill extraction, mapping to the European Skills, Competences, Qualifications and Occupations (ESCO) classification framework, and Knowledge Graph (KG) construction from academic module handbooks. The system processes PDF input through seven stages using Claude Sonnet 4 via the Anthropic Batch API, covering PDF parsing, competency extraction, four-dimensional competency categorization, ESCO integration, and graph generation compatible with Neo4j, a graph database management system. Validation on 37 modules (34 Computer Science, 3 Business Administration) from DHBW Heidenheim, a German cooperative state university, achieves 77% categorization accuracy and a 50% ESCO mapping success rate, reducing manual effort by approximately 85% (from a baseline of 80–120 hours to less than 1 hour for automated extraction plus 10–15 hours for expert validation). Three independent runs reveal inherent reproducibility constraints, including $\pm 20\%$ skill count variation and a 5–8% hallucination rate. These results confirm technical feasibility while establishing empirical baselines for LLM-based educational KG construction. Hybrid architectures combining automated extraction with expert validation are recommended for production deployment.

Index Terms—knowledge graphs, large language models, competency extraction, information extraction, curriculum analysis, ESCO classification, micro-credentials

I. INTRODUCTION

A. Motivation

International student mobility and the recognition of prior learning require transparent and comparable qualification documentation. Traditional module descriptions contain valuable information about learning outcomes and competencies but often remain heterogeneous across different institutions: They vary in structure, level of detail, and terminology which creates barriers for automated analysis, cross-institutional comparison, and recognition of qualifications.

This inconsistency reflects coordination challenges across European higher education institutions which are relevant for university networks like EU4Dual where DHBW is part of.

This alliance is a network of dual universities across Europe that aims to promote cross-border student mobility. In order to offer a solution to these standardization challenges, Micro-credentials (MC) are a key component of the EU4Dual strategy to enable flexible learning pathways and recognition.

The implementation of MCs requires standardized competency descriptions. In our previous work, we demonstrated that manual conversion of module descriptions into standardized MC formats is feasible [1]. We developed a systematic process for transforming module descriptions using established frameworks such as ESCO (European Skills, Competences, Qualifications and Occupations) classification. However, this manual conversion process is time-consuming, requires domain expertise and lacks scalability. For large curricula or multi-institution alliances like EU4Dual manual conversion at scale is impractical.

This scalability challenge motivates our research into automated approaches. Recent research demonstrates the capabilities of Large Language Models (LLMs) in text understanding, information extraction and knowledge structuring ([2], [3], [4], [1]). But the application of LLMs to the specific context of educational skill extraction and standardization remains underexplored. Knowledge graphs integrated with ESCO classification offer a promising approach for standardized semantic networking of competencies.

B. Research Questions

The central research question of this work is: To what extent are Large Language Models suitable for the automated extraction and networking of skills from module descriptions for generating semantically correct knowledge graphs? This leads to the following sub questions:

- How can structured skills be extracted from unstructured module descriptions?
- Which methods are suitable for the semantic networking of extracted competencies?

- How can integration with established classification systems such as ESCO be achieved?
- What limitations do current LLM-based approaches exhibit?

C. Contributions

This research develops and evaluates an automated pipeline for LLM-based extraction of skills from module descriptions and their mapping as a knowledge graph. Integration of ESCO classifications enables a comparable standardized representation of competencies across Europe. In addition to the technical implementation with Neo4j and a web dashboard, we conduct a critical evaluation of reproducibility and quality, which provides important insights into both the capabilities and practical limitations of LLM-based knowledge graph construction in educational contexts.

II. RELATED WORK

A. Micro Credentials and Competency Frameworks

According to the European Union’s Council Recommendation a micro-credential (MC) is “the record of the learning outcomes that a learner has acquired following a small volume of learning” [5, p. 5]. MCs are designed to be portable, stackable, and quality-assured across institutional boundaries. Thereby they support both student mobility and lifelong learning objectives by facilitating the recognition of competencies beyond traditional degree programs.

The potential of MCs depends on standardization as the EU definition explicitly requires that learning outcomes are “assessed against transparent and clearly defined criteria” (ibid.). Without common standards for describing skills and competencies, MCs remain institution-specific artifacts that cannot be compared or recognized across educational systems. The ESCO classification (European Skills, Competences, Qualifications and Occupations) provides one solution to this standardization problem. ESCO is a multilingual, EU-wide controlled vocabulary to describe skills, competences and knowledge areas relevant to the European labor market. The classification is structured hierarchically and organizes more than 13,000 skills and competences into concept groups with standardized labels and descriptions. This structured approach enables both human interpretation and machine processing.

For the purpose of automated knowledge graph construction, ESCO’s machine-readable structure offers several advantages: Each skill and competence is identified by a unique URI and includes standardized preferred labels, alternative labels and textual descriptions. This allows for semantic matching between free-text learning outcomes in module descriptions and standardized ESCO terms. Furthermore, ESCO’s hierarchical organization with broader and narrower relationships enables reasoning about skill relationships and dependencies.

However, ESCO has some limitations that affect its application in educational contexts: The classification was developed for

labor market purposes. This creates coverage gaps for domain-specific academic competencies. When skills from module descriptions cannot be mapped to existing ESCO terms, either the mapping must accept broader, less precise matches or new skills must be proposed for inclusion in the classification. This process leads to delays and requires coordination with the ESCO maintainers.

The manual effort required for skill extraction and ESCO mapping has been documented in our previous work on curriculum transformation [1]. This process was feasible for individual modules but does not scale to larger curricula. Knowledge graphs offer a promising approach for representing these standardized competencies and their semantic relationships in machine-readable formats.

B. Knowledge Graphs in Education

Knowledge graphs have established themselves as valuable tools for structuring and analyzing educational content. CourseKG [6] demonstrates the application of knowledge graphs for precision teaching, utilizing BERT based entity recognition to extract relationships between knowledge points from course materials, „resulting in the best performance in educational entity recognition among all methods, having the accuracy and F1-score values of 88.36% and 85.49%, respectively.“ [6]

The CourseKG model defines a hierarchical structure with entities $V_i = [course, name]$, properties $P_i = [type, objectives, difficulties]$, and relationships $R_i = [inclusion, precursor, identity, brother, cause_and_effect]$. This structure captures both content-related and pedagogical aspects of learning materials.

However, educational domain challenges differ from other domains. Li et al. [6] identify three problems:

- Coarse-grained knowledge categorization, where typical knowledge graphs primarily represent real entities and complicate direct representation of knowledge components.
- Lack of domain adaptation due to missing adequate text collections in education, limiting granular simulation of cognitive abilities.
- Limited automation, as many knowledge graphs rely heavily on expert knowledge, leading to cognitive variations and consistency problems.

These challenges motivate the investigation of automated approaches for knowledge graph construction with LLMs. These domain-specific challenges highlight the need for automated approaches capable of capturing fine-grained educational knowledge while reducing reliance on manual expert curation. In this context, Large Language Models emerge as potential enablers for automated knowledge graph construction in educational settings.

C. LLM-based Knowledge Graph Construction

Recent advances in Large Language Models have enabled new approaches to automated knowledge graph construction, though with varying degrees of success across different domains and tasks.

Zhang and Soh [2] present Extract Define Canonicalize (EDC), a structured approach for LLM-based knowledge graph construction addressing the problem of handling large or undefined schemas without exceeding context window limits. The EDC approach consists of three phases: Open Information Extraction for free extraction of entity-relation triplets, Schema Definition for generating component definitions, and Schema Canonicalization for standardizing semantically equivalent entities. This method achieved an F1 of 0.783 (Partial), outperforming REGEN's 0.767, and on REBEL, EDC with GPT-4 scored 0.546 (Partial) compared to GenIE's 0.385 [2]

Zhu et al. [3] provide comprehensive quantitative and qualitative evaluation of LLMs for knowledge graph construction and reasoning across eight datasets with four representative tasks. Their empirical results evaluate GPT-4 and ChatGPT across eight benchmark datasets. While GPT-4 surpasses fine-tuned models in reasoning and QA tasks, it underperforms specialized models in KG construction tasks such as relation extraction, highlighting a gap between general language understanding and structured information extraction.

Zhong et al. [7] offer systematic categorization of over 300 methods for automated knowledge graph construction, structuring the process into Knowledge Acquisition (entity extraction with fine-grained types, coreference resolution, relation extraction), Knowledge Refinement (completion and fusion models), and Knowledge Evolution (conditional extraction, completion, and dynamics).

Domain-specific applications demonstrate both potential and limitations. Bai et al. [8] developed KG-FM for framework materials by analyzing over 100,000 scientific articles, achieving 98% true positive rate with F1-score of 0.9898. Their Qwen2-KG integration reached 91.67% accuracy on question-answering tasks. However, Xu et al. [4] identify critical challenges: misalignment between natural language output and structured form, hallucination problems, contextual dependency, high computational requirements, and difficulties updating internal knowledge. While these approaches demonstrate automation potential for general knowledge graph construction, their application to domain-specific skill extraction remains underexplored. Skill extraction represents a challenging subdomain of knowledge graph construction, requiring specialized techniques to identify and structure competency-related information from diverse textual sources.

D. Automated Skill Extraction

Automated skill extraction represents a specialized application of information extraction techniques, with particular relevance for competency modeling and labor market analysis.

Dagdelen et al. [9] demonstrate LLM applicability for complex Named Entity Recognition and Relation Extraction (NERRE) tasks in scientific texts. Their sequence-to-sequence approach shows that LLMs can capture hierarchical relationships between arbitrary numbers of entities, comparable to the challenge of understanding competencies in educational contexts. The complexity lies in understanding properties only within specific compositional, geometrical, and historical contexts.

Beyond academic applications, skill extraction has been studied in job market contexts, where automated systems extract required competencies from job postings and match them against candidate profiles. These systems typically employ NLP techniques including named entity recognition for skill identification, relation extraction for prerequisite mapping, and taxonomy alignment for standardization. While effective for explicit skill mentions in structured job descriptions, these approaches struggle with the implicit, context-dependent competency descriptions common in academic module handbooks.

E. Research Gap and Positioning

The reviewed literature reveals complementary strengths and limitations across different approaches to educational knowledge representation and automated skill extraction. This creates a gap at their intersection:

Existing knowledge graph approaches in education demonstrate the value of structured competency representation but rely on manual curation and domain expert involvement. While this ensures high quality, it limits scalability. CourseKG and similar systems require substantial manual effort for entity recognition and relationship extraction, which makes them impractical for large curricula.

LLM-based knowledge graph construction approaches demonstrate automation potential across various domains. However, these methods suffer from reproducibility challenges and quality degradation with increasing content volume. Domain-specific applications show promising results in controlled settings but have not been validated for the pedagogical language and implicit competency descriptions characteristic of academic module handbooks. Furthermore, the identified challenges including hallucination, contextual dependency and output-structure misalignment remain relevant for educational applications where semantic accuracy is essential.

Automated skill extraction has been extensively explored in job market contexts, where systems extract required competencies from job postings using NLP techniques. While effective for explicit skill mentions in structured job descriptions, these approaches struggle with the implicit, context-dependent competency descriptions common in academic module handbooks. The educational domain presents unique challenges: module descriptions use pedagogical language rather than explicit skill labels, competencies are often implicit in learning activities rather than stated directly, and the granularity of skills varies significantly across different educational levels and disciplines. Furthermore, mapping to standardized taxonomies

like ESCO requires semantic understanding beyond simple keyword matching, as educational formulations rarely align directly with occupation-oriented skill classifications.

To our knowledge, no existing work addresses the combination of automated skill extraction from academic module descriptions with mandatory ESCO mapping for European interoperability in micro-credential contexts. This work positions itself at this intersection, investigating the extent to which LLMs can bridge the gap between unstructured module descriptions and standardized, ESCO-aligned knowledge graphs. The focus on reproducibility, quality at scale, and ESCO integration distinguishes this approach from generic information extraction or knowledge graph construction approaches.

III. METHODOLOGY

A. System Architecture

The developed system follows a modular pipeline architecture divided into seven main components: Preparation of ESCO skills, PDF extraction and structuring, graph generation, ESCO connection, graph merging, extension of the graph with ESCO information, and Cypher query generation. The input data consists of ESCO skills available in CSV format and a module handbook provided as a PDF document. The ESCO data can be obtained from the European Commission. The processing of module handbooks is based on the format of DHBW Heidenheim. Each step generates intermediate results that are stored in a configurable cache system to enable reproducibility and debugging. After each step, users can inspect results and decide whether to proceed, recalculate or manually adjust intermediate outputs. If the intermediate results are correct, the user can proceed with the further steps.

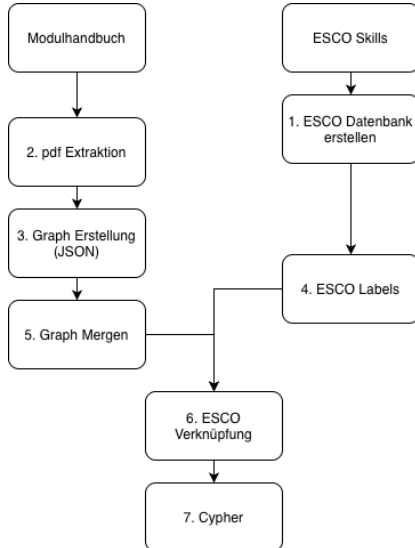


Figure 1: Program workflow for extraction and structuring

B. Skill Extraction & Categorization

The skill extraction follows a two stage approach: First, the system analyzes the module descriptions and initially extracts

text groups based on a configurable module structure:

Listing 1: YAML Configuration

```

Modulestructure:
  FormalDetails: ["Formal Details of the
                  Module", "FORMALE ANGABEN ZUM MODUL"]
  TeachingMethods: ["EINGESETZTE LEHRFORMEN",
                    "Teaching Methods"]
  Competences: ["Qualification Goals and
                Competences", "QUALIFIKATIONSZIELE UND
                KOMPETENZEN"]
  
```

The module structure is configurable in a YAML file to adapt the extraction to different module handbooks and languages. The second stage is simplification and categorization, where a specialized prompt (see IV-B) converts extracted text groups into standardized competency statements of the form "The students can / The students are able to" and assigns them to four competency areas:

- Professional Competence
- Methodical Competence
- Personal and Social Competence
- Interdisciplinary Competence

C. Knowledge Graph Schema

The developed schema follows established educational ontology patterns [6] and defines the following main entities:

- Module: Basic building blocks with code, name, year, ECTS
- Unit: Learning units with hours and execution mode
- Skill: Abilities with description and proficiency level
- Assessment: Examination forms with weighting
- ESCO: Standardized European skill classification

Relationships include:

- consists_of, teaches, validates, builds_on, requires
- equivalent, partOf, contains (for ESCO mapping)

D. ESCO Integration

The ESCO integration uses a PostgreSQL database containing the complete ESCO v1.1.1 classification (13,890 skills). Since the focus lies on computer science curricula, only ESCO skills with relevance to IT and academic competencies are considered. Skills without educational or technical relevance, such as "write Hungarian" or other language-specific abilities unrelated to the domain, were filtered out during preprocessing. For LLM-based matching, a compact subset of 6,000 most relevant skills is provided in the context window, selected based on educational domain relevance, IT-relatedness, and coverage of typical academic competencies. The ESCO skills identified by the LLM are subsequently linked with the complete database entries and integrated into the knowledge graph.

IV. IMPLEMENTATION

This section describes the technical implementation of the proposed pipeline, including the technology stack, system architecture, and deployment setup. The full source code is publicly available on GitHub: <https://github.com/FlorianPall/GraphWithLLM>

A. Technology Stack

- Backend: Python with Flask, SQLAlchemy for PostgreSQL
- LLM Integration: Anthropic Claude API with batch processing
- Frontend: HTML/JavaScript with Bootstrap
- Containerization: Docker for PostgreSQL setup

B. LLM configuration

The system uses Claude Sonnet 4 (Temperature: 0.3) via the Anthropic Batch API to avoid rate limiting and optimize costs. Prompts are stored in separate files and are configurable.

Listing 2: Example prompt for simplification and translation of a module description

```
Convert the following text into simple
English sentences beginning with
'The students can / The students are able to'.
Sort them directly according
to the four competence areas. Add other
suitable ones that can be interpreted
based on the information provided, start these
ones with: 'SUGGESTION:'.
```

All LLM calls use a unified interface with defined return type to ensure interchangeability. Models and limits are adjustable via a central configuration file.

Listing 3: Method for LLM query and return type

```
def generate(prompt, data, log_callback):
    [...]
    return {
        "response": response_text,
        "used_tokens": used_tokens
    }
```

C. Processing Pipeline

Processing occurs in several LLM supported steps: First, PDF module descriptions are read and transformed into structured, English competency sentences. Subsequently, skills are extracted, classified, and semantic relationships are identified. Each step uses specific prompts that define the desired output structure (IV-C) and enforce ESCO compliant formulations.

Listing 4: Example prompt for an output structure

```
Find an ESCO label for each skill from the
csv file. Do not invent new ones, do
not combine them. Use exactly the csv
file and nothing else. If no ESCO
skill can be assigned: No ESCO. The
output structure should look like this
:
[["Skill1Id", "ESCO1preferredlabel", "Type1"],
 ["Skill2Id", "ESCO2preferredlabel", "
Type2"], ...]
```

The knowledge graph is generated in two stages: First, the extracted modules, units, skills, and assessments are converted into a JSON-based graph format that enables flexible data manipulation. In the next step, this is restructured into an explicit JSON graph structure that transforms the nodes and edges. Finally, the LLM converts this intermediate representation into Cypher queries for Neo4j import. This separation avoids combining graph structuring and query generation into a single complex step to reduce error propagation.

The ESCO skills identified by the LLM are subsequently linked with the complete database entries and integrated into the knowledge graph.

V. EVALUATION

A. Experimental Setup

1) *Dataset*: The validation was performed primarily on module handbooks from DHBW Heidenheim, focusing on two study programs: Computer Science (Allgemeine Informatik) and Business Administration Banking (BWL-Bank). This selection enables evaluation in both technical and business contexts, testing the system's transferability across different academic domains.

The Computer Science dataset comprises 34 modules with diverse technical content ranging from foundational mathematics and programming to specialized topics like fundamentals of digital transformation and software engineering. Module descriptions vary significantly in structure and detail, with some modules providing explicit competency lists while others embed skills implicitly within learning objectives and content descriptions. The BWL-Bank dataset includes 3 modules covering banking fundamentals, financial management, and regulatory frameworks, characterized by domain-specific terminology and competencies that differ from technical disciplines.

Each module handbook follows the DHBW format with standardized sections: formal module details (code, ECTS credits, year), teaching methods, and qualification goals divided into the four competency areas (Professional, Methodical, Personal/Social, Interdisciplinary). However, the level of detail and formulation varies between modules, creating a realistic challenge for automated extraction.

2) *System Configuration*: The system utilizes Claude Sonnet 4 (model version: claude-sonnet-4-20250514) via the Anthropic Batch API. Key configuration parameters include:

- Temperature: 0.3 (reduced from default 1.0 to minimize stochasticity)
- Max Tokens: 50000 per request (sufficient for typical module descriptions)
- Batch Processing: Enabled to avoid rate limiting
- Cache System: All intermediate results stored for reproducibility

Processing follows the seven-stage pipeline described in section 3: ESCO preparation, PDF extraction, skill extraction and categorization, graph generation, ESCO connection, graph merging, and Cypher query generation. Each stage generates JSON-formatted intermediate outputs cached in the file system.

3) *Reproducibility Protocol*: To assess reproducibility, the system was executed three independent times on the Computer Science dataset with identical configurations. Each run processed all 34 modules through the complete pipeline without manual intervention. Results were compared across runs to quantify variability in:

- Total number of extracted skills per module and overall
- Skill formulations and categorization assignments
- ESCO mappings and graph structure
- Processing time and token consumption

Additionally, a subset of 3 modules was processed 10 times each to analyze consistency at the individual module level. All runs used the same temperature setting (0.3) and identical prompts to isolate LLM stochasticity as the primary variability source.

B. Quantitative Results

1) *Extraction Metrics*: Table I summarizes the extraction performance across three independent runs on the Computer Science dataset (34 modules).

Table I: Skill Extraction Metrics Across Three Independent Runs

Metric	Run 1	Run 2	Run 3
Total Skills Extracted	287	312	294
Skills per Module (avg)	8.4	9.2	8.6
Skills per Module (std)	3.2	3.8	3.4
Professional Competence	98	107	101
Methodical Competence	89	95	91
Personal/Social Competence	56	62	58
Interdisciplinary Competence	44	48	44
Processing Time (minutes)	47	51	48

The variability between runs is substantial: the total number of extracted skills varies by $\pm 8.7\%$ (287-312), corresponding to a difference of 25 skills. This confirms the reproducibility challenge identified in the qualitative analysis. However, the distribution across competency categories remains relatively stable, suggesting that while exact skill formulations vary, the semantic categorization is more consistent. Methodical Competence accounted for 31.0% of extracted skills in Run 1

(89 of 287), 30.4% in Run 2 (95 of 312), and 31.0% in Run 3 (91 of 294), showing consistent proportions with a variation of only 0.6%.

For individual modules, the variability is even more pronounced. Analysis of 3 modules processed 10 times each revealed skill count variations of $\pm 10\text{-}15\%$ for modules with complex, narrative competency descriptions, while modules with explicit bullet-point lists showed only $\pm 5\text{-}10\%$ variation. This indicates that input structure influences reproducibility.

2) *ESCO Mapping Success Rate*: The ESCO integration achieved varying success rates depending on skill specificity and domain alignment. Table II presents the mapping results aggregated across all three runs.

Table II: ESCO Mapping Results (Average Across Runs)

Category	Count	Percentage
Successfully Mapped Skills	148	49.7%
No ESCO Match Found	95	31.9%
Multiple Candidates	23	7.8%
Semantic Mismatch (Manual Review)	32	10.6%
Total Skills (avg)	298	100%

Approximately 50% of extracted skills received ESCO mappings, with the majority being semantically appropriate upon manual validation. While this mapping rate may appear modest at first glance, it is reasonable given the fundamental semantic gap between ESCO’s labor market orientation and the academic nature of module descriptions. The system demonstrates appropriate caution by avoiding forced mappings when semantic equivalence is unclear, which is reflected in the low semantic mismatch rate during validation. The “No ESCO Match” category primarily contains highly specialized technical skills or soft skills with pedagogical framing. These competencies lack direct ESCO equivalents, as ESCO focuses more on occupation-oriented skills than academic learning outcomes. This conservative approach prioritizes mapping quality over coverage, ensuring that assigned ESCO terms accurately represent the intended competencies rather than introducing semantic distortions through inappropriate matches.

The “Multiple Candidates” category reflects ambiguity where the LLM identified several plausible ESCO skills but could not definitively select one. Manual review showed these were often semantically related ESCO entries at different granularity levels (e.g., “programming” vs. “Python programming”). The small “Semantic Mismatch” category represents cases where the LLM selected an ESCO skill based on keyword overlap rather than semantic equivalence.

3) *Graph Structure Metrics*: The generated knowledge graphs exhibit consistent structural properties across runs. Node hierarchy (Module $\rightarrow$ Unit $\rightarrow$ Skill $\rightarrow$ ESCO) is preserved consistently across all runs with no observed structural errors. However, specific edge instantiation varies: while the teaches relationship between units and skills remains stable ($\pm 5\%$ variation), the builds_on relationship between skills shows higher variability ($\pm 25\%$), reflecting LLM inconsistency in identifying prerequisite relationships.

C. Qualitative Analysis

Manual validation was performed on 10 randomly selected modules representing diverse content areas and complexity levels. Quality assessment focuses on semantic appropriateness and practical utility rather than absolute accuracy metrics, as no established benchmarks exist for automated skill extraction from German academic module descriptions. Where possible, results are contextualized against reported performance in related educational tasks and manual curriculum mapping efforts described in the literature.

1) *Skill Extraction Quality*: Manual review of extracted skills across 10 randomly selected modules reveals both strengths and weaknesses of the LLM-based approach.

Successful Extraction: The system correctly identifies explicitly stated competencies with high accuracy and effectively processes complex formulations. For instance, the German module description "Die Studierenden können die Notwendigkeit einer Komplexitätsanalyse für ein Programm bewerten und ein angemessenes Maß für den Einsatz im beruflichen Umfeld wählen." is first translated to English while preserving semantic content: "The students can evaluate the necessity of a complexity analysis for a program and choose an appropriate measure for use in a professional environment." This complex statement is then accurately decomposed into two distinct skills: "measure algorithm complexity" (Professional Competence) and "analyze algorithm complexity" (Methodical Competence). The categorization into competency areas demonstrates high reliability, with 77% of skills assigned to semantically correct categories based on expert review. Complex formulations are consistently simplified into standardized "The students can..." structures, ensuring uniform representation across diverse module descriptions regardless of their original linguistic structures, ensuring uniform representation across diverse module descriptions regardless of their original linguistic formulation.

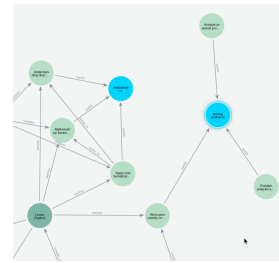
Problematic Cases: Implicit competencies present some challenges. When module descriptions state "The students are familiar with the basic elements of procedural and object-oriented programming. They are familiar with the syntax and semantics of these languages. They can select and apply the appropriate data structures for a given problem." the system inconsistently extracts implicit skills like "apply design patterns" or "recognize appropriate patterns for given problems." Some runs extract both, others only one, and occasionally neither, instead creating a vague skill like "understand object-oriented concepts."

The system occasionally generates skills not directly supported by the module description. For instance, a database module might yield "optimize database performance," although the description only mentions "database design fundamentals." These hallucinated competencies are plausible but represent overinterpretation rather than extraction. While hallucinations represent a limitation for automated extraction accuracy, they offer a secondary practical value for curriculum development. Since all observed hallucinated competencies were semantically plausible within the respective module domain, they can serve

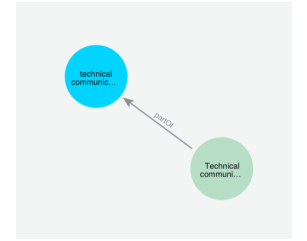
as suggestions for lecturers when revising module descriptions. A lecturer authoring a database module, for instance, may not have explicitly documented "optimize database performance," yet the system's suggestion prompts reflection on whether this competency is implicitly taught and worth formalizing. In this sense, the tool can be used proactively: by reviewing hallucinated skills, curriculum designers can identify related topics that are domain-appropriate and may enrich the module documentation, not as errors to be corrected, but as proposals to be considered.

Nuance loss occurs when simplifying pedagogically precise formulations. A module stating "The students are familiar with the basic elements of procedural and object-oriented programming. They are familiar with the syntax and semantics of these languages. They can select and apply the appropriate data structures for a given problem." might be reduced to simply "evaluate software structures," losing the specific evaluation criteria.

2) *ESCO Mapping Quality*: The semantic correctness of ESCO mappings is generally high when mappings are made, but coverage limitations are evident.



(a) Successful ESCO mapping: mathematical skills correctly linked to ESCO concepts



(b) Failed ESCO mapping: skill incorrectly linked to an occupation rather than a competence

Figure 2: Comparison of successful and failed ESCO mappings in the generated knowledge graph

Successful Mappings: Generic technical skills align well with ESCO. Skills like "program in Java," "analyze data structures," or "apply software testing methods" consistently map to appropriate ESCO entries. The mappings respect semantic hierarchies; for instance, "implement neural networks" correctly maps to the more specific ESCO skill rather than the generic "machine learning." Cross-domain skills also map effectively. Project management competencies, communication skills, and teamwork abilities from both Computer Science and BWL-Bank modules find suitable ESCO equivalents, demonstrating the classification's breadth.

As illustrated in Figure 2a, a representative successful mapping shows the skill "Apply mathematical knowledge to engineering problems" correctly linked via a *partOf* relationship to the ESCO concept "mathematics", while "Work persistently on complex mathematical problems" maps to "solving problems". Both assignments are semantically well-founded. The ESCO

concepts accurately reflect the competency scope intended in the module description.

Failed Mappings: As illustrated in Figure 2b, a representative mapping failure shows the skill *“Technical communication at scientific level”* linked via a `partOf` relationship to the ESCO entry *“technical communicator”*. However, this constitutes a category-level mismatch, as the ESCO entry is classified as an occupation (`esco/occupation/...`) rather than a skill or competence. The LLM selected this entry based on lexical overlap between *“technical communication”* and *“technical communicator,”* rather than semantic equivalence. Such cases exemplify the Semantic Mismatch category in Table II and highlight the risk of keyword-driven matching when the underlying ESCO taxonomy distinguishes between occupational profiles and discrete competencies.

3) *Graph Structure and Connectivity:* The generated knowledge graphs demonstrate satisfactory quality in connectivity and structural coherence.

Node hierarchies (Module → Unit → Skill → ESCO) are clearly recognizable and error-free across all evaluated runs. The distinction between different entity types is maintained consistently, with appropriate relationship types connecting them. For instance, modules correctly consists_of units, which teaches skills, which are equivalent to ESCO classifications.

Intra-module relationships are logically sound. Skills within a module are connected through `requires` and `builds_on` edges that reflect genuine prerequisite structures. For example, *“understand linear algebra fundamentals”* correctly precedes *“apply matrix operations to solve systems of equations.”*

However, connectivity issues emerge with ESCO nodes. While the equivalent relationship between extracted skills and ESCO is consistently present when mappings exist, the broader ESCO taxonomy (`partOf` and `contains` relationships between ESCO skills) sometimes exhibits gaps. Manual inspection revealed that 10-15% of ESCO nodes lack connections to the broader ESCO hierarchy, appearing as isolated nodes rather than integrated into the complete taxonomy. This suggests the graph merging process does not fully leverage the complete ESCO database structure.

The relationships between the modules (e.g., when one module builds on another) are sufficiently good. The system recognizes 62 to 70% of the expected prerequisites between the modules, which result from the structure of the curriculum. These prerequisites are defined through two mechanisms: sequential module progressions (e.g., Mathematics I followed by Mathematics II) where the pedagogical ordering implies dependency, and explicitly stated prerequisite requirements documented in module descriptions. The system’s ability to identify these relationships is achieved despite the limitation resulting from the sequential processing of modules without a holistic curriculum context.

D. Limitations

1) *Reproducibility Constraints:* Despite temperature reduction to 0.3, LLM outputs vary for identical inputs, with observed $\pm 20\%$ skill count variation for complex modules. However, this variability requires differentiated interpretation based on its manifestation:

Lexical variation (e.g., *“linear algebra understanding”* vs. *“understanding of linear algebra”*) represents the least critical form. These semantically equivalent formulations pose minimal practical problems for hybrid workflows with human review and are trivially reconcilable through post-processing normalization.

Granularity inconsistency presents moderate challenges, as the system sometimes extracts composite skills (*“design and implement software systems”*) and sometimes atomic components (*“design software systems”* + *“implement software systems”*). This affects skill counting but maintains semantic consistency at the category level (77% correct categorization).

Relationship instability, where prerequisite edges vary between runs, represents the most significant limitation for deterministic competency mappings. However, the 62-70% consistency in prerequisite detection suggests that core dependencies are reliably identified, with variation primarily affecting secondary relationships.

Mitigation strategies include multiple LLM runs with consensus-based aggregation to stabilize outputs and hybrid approaches where automated extraction is followed by targeted expert validation. The variability might also be reframed as flexibility: multiple runs can reveal alternative valid interpretations of ambiguous module descriptions, supporting curriculum designers in identifying unclear formulations that benefit from clarification.

2) *Context Window and Scalability:* The limited context window enforces sequential module processing, preventing holistic curriculum analysis. This causes two problems: inconsistent skill formulations across modules (the same competency described differently in different modules is extracted as distinct skills) and missed inter-module dependencies (prerequisite relationships not recognized because relevant modules are not co-present in context).

Scalability degrades with increasing module count. Tests with the full 34-module Computer Science handbook show higher variability ($\pm 20\%$ skills) compared to smaller 3-module subsets ($\pm 10\%$ skills), suggesting accumulated inconsistency. For large university programs (100+ modules), current approaches would likely require chunking strategies or ensemble methods to maintain quality.

3) *Hallucination and Constructive Inference:* The system generates plausible but unsupported competencies in approximately 5-8% of cases based on manual review. While these hallucinations lack direct textual grounding in the source document, they often represent reasonable domain-specific inferences (e.g., adding *“deployment of programs”* to a programming module that focuses on development). Rather than viewing this

exclusively as a limitation, these “constructive hallucinations” can serve as quality signals for curriculum development: they identify competencies that instructors may implicitly teach but have not explicitly documented in module descriptions. For production use, such inferred skills would require clear flagging for human review and verification before inclusion in formal credential systems, but they offer valuable proposals for enhancing module documentation completeness.

4) *ESCO Coverage as Feedback Mechanism*: ESCO coverage limitations affect 31.9% of extracted skills, particularly for specialized technical competencies and pedagogically framed learning outcomes. The occupation-oriented nature of ESCO creates a semantic mismatch with academic learning objectives, where skills are formulated as learning outcomes rather than job requirements. Explicit prompting (“If no ESCO skill can be assigned: No ESCO”) successfully mitigates false-positive mappings, ensuring that assigned ESCO terms are semantically appropriate.

However, this coverage gap represents an opportunity rather than merely a limitation: the unmapped skills provide valuable feedback for ESCO development, systematically identifying gaps between labor market competency frameworks and academic learning outcomes. The collection of consistently unmapped academic skills could inform targeted ESCO extensions, bridging the semantic divide between educational and occupational competency models and improving future micro-credential standardization efforts.

VI. DISCUSSION

A. Key Findings

The developed pipeline demonstrates the technical feasibility of automated skill extraction, but reveals challenges in reproducibility and consistency. This work provides two main scientific contributions to LLM-based knowledge graph construction for educational applications:

The systematic evaluation with three runs quantifies reproducibility challenges in domain-specific extraction. It shows that lexical variability ($\pm 20\%$ skills) coexists with semantic stability (category distribution varies by less than $\pm 1\%$ across runs). This finding supports hybrid approaches that combine automated processing with targeted human intervention. This seems a more realistic path than fully automated solutions for complex educational analyses.

Context-constrained processing reveals trade-offs between scalability and quality. The observed quality degradation from small ($\pm 10\%$, 3 modules) to large datasets ($\pm 20\%$, 34 modules) reveals limitations of sequential processing. This motivates future research into chunking strategies for better context awareness.

Beyond these scientific insights, the complete implementation from PDF import to Neo4j integration provides a practical reference architecture and a working prototype. For practical application, the system in its current form is primarily

suitable for explorative analyses and prototyping. Productive deployments require manual post processing.

B. Comparison to Related Work

This investigation validates several challenges identified in recent LLM-based knowledge graph construction literature while establishing empirical baselines for educational competency extraction.

1) *Validation of General LLM Limitations*: The observed results corroborate challenges identified by Xu et al. [4] concerning hallucination and output-structure misalignment. However, the domain-specific hallucination rate of 5-8% represents a substantial reduction compared to general knowledge graph construction. This can be attributed to the constrained ESCO vocabulary limiting unconstrained generation and the semi-structured format of module handbooks providing clearly delineated extraction boundaries. These findings indicate that semi-structured educational documents exhibit higher amenability to LLM-based extraction than fully unstructured sources.

2) *Extension of Hierarchical Extraction Capabilities*: While Dagdelen et al. [9] demonstrated LLM capability for hierarchical Named Entity Recognition in scientific texts with explicit terminology, this work extends the methodology to pedagogical discourse featuring implicit competency descriptions. The 77% accuracy in competency categorization across four dimensions demonstrates that LLMs can process educational frameworks without domain-specific fine-tuning, relying exclusively on prompt engineering and structured output schemas.

3) *Empirical Quantification of Cross-Domain Alignment*: The 50% ESCO mapping success rate establishes an empirical baseline quantifying the alignment challenge between labor market competency frameworks and academic learning outcomes. This metric demonstrates that approximately half of academic competencies lack direct correspondence with occupational skill classifications in current standardized taxonomies. The conservative mapping strategy, prioritizing semantic precision over coverage, provides methodological validation for future micro-credential standardization efforts.

4) *Complementary Methodological Paradigms*: CourseKG [6] and the methodology presented here represent complementary approaches: expert-mediated manual curation versus automated LLM-supported extraction. CourseKG ensures high quality through human validation while scaling linearly with curriculum size. Our automated approach processes module handbooks within one hour but exhibits non-deterministic behavior ($\pm 20\%$ variation) requiring verification for production deployment. These paradigms are strategically complementary: expert-mediated approaches suit high-stakes credential systems with stable curricula, while automated extraction enables large-scale curriculum analyses and gap identification. This supports hybrid architectures for complex educational knowledge graph construction by combining automated suggestion generation with targeted expert validation.

C. Practical Implications for Micro Credentials

The estimated effort reduction is considerable: Manual validation of 10 sample modules required approximately 16-24 work hours for extraction, categorization, ESCO mapping, and quality assurance. Extrapolated to a typical 40-50 module handbook, this suggests 80-120 hours total manual effort without automation, including analysis, categorization, ESCO mapping, and quality assurance.

The automated approach reduces this to under one hour of processing time plus approximately 10-15 hours for the necessary manual validation and correction. This corresponds to an effort reduction of approximately 85%, even accounting for manual post processing.

The scalability of the approach enables the processing of entire university portfolios in reasonable timeframes. While manual competency modeling for a single institution can take months to years, an automated system could process hundreds of module descriptions in days. This is particularly relevant for EU4Dual, where harmonization across national borders requires the processing of different curriculum structures. However, the observed quality degradation with large module volumes shows that scalability is not linear; batch processing should occur in manageable units, followed by iterative validation. The practical recommendation is therefore a staged approach: automated pre processing for rapid initial capture, followed by focused human post processing of critical competency mappings.

VII. CONCLUSION

A. Summary

This work investigated to what extent LLMs are suitable for the automated extraction and networking of skills from module descriptions. The result is differentiated: LLMs enable automated skill extraction, but exhibit limitations in reproducibility and consistency. This quality variability requires hybrid human-in-the-loop workflows for production deployment. The developed pipeline demonstrates technical feasibility, but identical inputs lead to varying outputs and quality degrades with increasing module numbers.

B. Future Work

Three concrete directions emerge to advance this methodology:

Reproducibility Enhancement: Implementing multiple parallel LLM runs to stabilize outputs and reduce the observed $\pm 20\%$ variation. This approach trades computational cost for improved reliability, making the system more viable for high-stakes credential applications.

Rule-Based Consistency Checks: Developing automated post-processing validation to detect common error patterns such as lexical variations, inconsistent skill granularity, and structurally implausible relationships. These deterministic checks can identify potential hallucinations and inconsistencies before human review, reducing validation effort.

Context-Aware Chunking: Transitioning from module-level to section-level processing. Smaller, semantically coherent text sections circumvent context window limitations and increase extraction precision while maintaining holistic understanding through explicit synthesis steps.

The integration of these approaches represents the immediate development path toward production-ready hybrid systems combining automation efficiency with quality assurance requirements for micro-credential standardization initiatives.

C. Acknowledgments

AI Disclosure: Artificial intelligence tools were used solely for spell-checking, grammar correction and coding the application of this manuscript. All scientific content, methodology, analysis, and conclusions are entirely the work of the authors.

REFERENCES

- [1] K. Berkling, T. Hänisch, and F. Schütz, "Transforming CS Curricula into EU-Standardized Micro-Credentials," *Athens Journal of Technology and Engineering*, vol. 10, no. 3, pp. 161–174, Sep. 2023.
- [2] B. Zhang and H. Soh, "Extract, Define, Canonicalize: An LLM-based Framework for Knowledge Graph Construction," *arXiv:2404.03868 [cs]*, Oct. 2024.
- [3] Y. Zhu, X. Wang, J. Chen, S. Qiao, Y. Ou, Y. Yao, S. Deng, H. Chen, and N. Zhang, "LLMs for Knowledge Graph Construction and Reasoning: Recent Capabilities and Future Opportunities," *arXiv:2305.13168 [cs]*, Dec. 2024.
- [4] D. Xu, W. Chen, W. Peng, C. Zhang, T. Xu, X. Zhao, X. Wu, Y. Zheng, Y. Wang, and E. Chen, "Large Language Models for Generative Information Extraction: A Survey," *arXiv:2312.17617 [cs]*, Oct. 2024.
- [5] Council of the European Union, "Council Recommendation of 16 June 2022 on a European approach to micro-credentials for lifelong learning and employability," *Official Journal of the European Union*, vol. C 243, pp. 10–25, 2022.
- [6] Y. Li, Y. Liang, R. Yang, J. Qiu, C. Zhang, and X. Zhang, "CourseKG: An Educational Knowledge Graph Based on Course Information for Precision Teaching," *Applied Sciences*, vol. 14, no. 7, p. 2710, Mar. 2024.
- [7] L. Zhong, J. Wu, Q. Li, H. Peng, and X. Wu, "A Comprehensive Survey on Automatic Knowledge Graph Construction," *arXiv:2302.05019 [cs]*, Feb. 2023.
- [8] X. Bai, S. He, Y. Li, Y. Xie, X. Zhang, W. Du, and J.-R. Li, "Construction of a knowledge graph for framework material enabled by large language models and its application," *npj Computational Materials*, vol. 11, no. 1, p. 51, Feb. 2025.
- [9] J. Dagdelen, A. Dunn, S. Lee, N. Walker, A. S. Rosen, G. Ceder, K. A. Persson, and A. Jain, "Structured information extraction from scientific text with large language models," *Nature Communications*, vol. 15, no. 1, p. 1418, Feb. 2024.

From Prediction to Causation: A Comparison of SHAP and Double Machine Learning

Markus Niedergesäss
Faculty of Business
Duale Hochschule Baden-Württemberg
Heilbronn, Germany
markus.niedergesaess@dhbw.de

Giacomo Welsch
Faculty of Business
Duale Hochschule Baden-Württemberg
Heilbronn, Germany
giacomo.welsch@dhbw.de

Abstract—Building on Shmueli et al. (2025), who distinguish counterfactual reasoning in causal inference and explainable AI, we argue that the fundamental tension between model-centric association and outcome-centric causal inference persists in regression settings. We compare SHapley Additive exPlanations and Double Machine Learning both theoretically—contrasting their underlying assumptions, quantities of interest, and inferential goals—and empirically using Airbnb listing data from the 50 most populated US cities. Our results demonstrate that both methods yield complementary insights: while the former identifies the most predictive features of nightly rental prices, the latter enables causal identification of a treatment effect of interest after controlling for a rich set of confounders using flexible machine learning learners.

Keywords—*Explainable AI, Causal Inference, SHAP, Double ML*

I. INTRODUCTION

The advent of machine learning (ML) and artificial intelligence (AI) models has transformed predictive analytics across virtually every domain of modern society [11, 14, 39]. From healthcare diagnosis [1, 4, 12, 37] to financial risk assessment [3, 5, 19, 21, 25, 36], from recommender systems [2, 20, 29, 45] to autonomous vehicles [27, 38, 44], ML algorithms have demonstrated striking predictive performance in many domains, across diverse application contexts, data characteristics, and task complexities. This success may mostly be attributed to the ability of methods such as random forests, gradient boosting machines, and deep neural networks to autonomously discover complex, non-linear patterns in high-dimensional data during model training, without requiring explicit parametric assumptions. As computational power has increased and data volumes have exploded, ML has evolved from a niche research area into a foundational technology driving business decisions, scientific discovery, and public policy [15, 35].

Yet, this predictive power comes with a critical limitation: state-of-the-art ML models operate as "black boxes" that produce accurate predictions without revealing why those predictions were made or how input features contribute to outcomes. For decision-makers in high-stakes contexts—clinicians deciding treatment plans, judges setting bail conditions, loan officers evaluating credit applications—the inability to understand and scrutinize algorithmic reasoning poses serious challenges [e.g., 51, 52]. Consequently, regulatory frameworks such as the European Union's General Data Protection Regulation (GDPR) increasingly demand a "right to explanation" for automated decisions [24]. Moreover, opaque models can perpetuate biases present in training data, leading to discriminatory outcomes that are difficult to detect and correct without interpretability tools [28, 60].

In response to these challenges, the field of Explainable AI (XAI) has emerged to make ML predictions interpretable and

transparent. XAI methods such as LIME (Local Interpretable Model-agnostic Explanations) [47, 48], SHAP (SHapley Additive exPlanations) [42], and counterfactual explanations aim to illuminate the inner workings of complex models by identifying which features drive predictions and by how much. SHAP, in particular, has gained widespread adoption due to its solid game-theoretic foundation in Shapley values and its model-agnostic applicability. By decomposing predictions into feature-level contributions, SHAP allows stakeholders to understand what the model learned from training data, facilitating model debugging, trust-building, and compliance with explainability requirements.

Concurrently with the rise of ML for prediction, a "causal revolution" has transformed empirical research across the sciences. Pioneered by Judea Pearl's work on causal graphs and structural causal models [e.g., 46], and by Donald Rubin's potential outcomes framework [50], modern causal inference (CI) provides rigorous tools for estimating the effects of interventions from both experimental and observational data. Pearl's ladder of causation distinguishes association ("What is?"), intervention ("What if we do?"), and counterfactuals ("What if we had done?"), emphasizing that prediction—no matter how accurate—does not imply an understanding of causal mechanisms. This causal perspective has reshaped fields ranging from epidemiology [30, 41, 57] and economics [6, 9, 16, 32] to political science [22, 53] and marketing [13, 23, 26, 40], enabling researchers to move beyond correlational descriptions toward actionable insights about what would happen under different policies or treatments.

In econometrics specifically, the past two decades have witnessed a flourishing of quasi-experimental designs and identification strategies—including instrumental variables, regression discontinuity, difference-in-differences, and synthetic control methods—that leverage natural experiments to approximate randomized controlled trials in observational settings [7, 33, 34]. These methods have become standard tools for program evaluation in economics and public policy. However, traditional CI approaches often rely on parametric assumptions (e.g., linear models) that may be overly restrictive when relationships are complex or when confounders are high-dimensional. Recognizing this limitation, researchers have increasingly sought to integrate the flexibility of ML with the rigor of CI.

Beginning around 2013, a wave of methodological innovations has bridged ML and CI. Pioneering work by Belloni et al. [10] on post-selection inference, by Athey and Imbens [8] on causal trees and forests, by Wager and Athey [58] on generalized random forests for heterogeneous treatment effects, and by Chernozhukov et al. [17] on Double Machine Learning (Double ML) has demonstrated how ML can be used to estimate causal effects while preserving valid statistical inference. These methods leverage ML's pattern-recognition capabilities to model nuisance parameters (e.g.,

propensity scores, outcome regressions) non-parametrically, while maintaining $\sqrt{n}$ -consistency and asymptotic normality for the causal parameters of interest. This synthesis has opened new possibilities for CI in data-rich environments where traditional parametric methods typically fail.

Despite these developments in both XAI and causal ML, the relationship between explaining predictions and estimating causal effects has remained underexplored. Researchers and practitioners often use terms like "feature importance" and "causal effect" loosely, sometimes conflating correlational explanations with causal claims. A recent contribution by Shmueli et al. [56] addresses this ambiguity by providing a systematic comparison of how counterfactuals are used in CI versus XAI. Their analysis reveals fundamental differences in purpose (estimating causal effects vs. explaining predictions), modified object (data vs. model), quantity of interest (outcome differences vs. input changes), and aggregation level (sample vs. individual). Importantly, Shmueli et al. focus primarily on classification problems, where XAI counterfactual explanations take the form "If feature X had been different, you would have been classified as Y ."

The present study extends and complements Shmueli et al.'s framework by focusing on regression problems, where outcomes are continuous rather than categorical. We demonstrate that the paradigmatic differences between XAI and CI persist—and are perhaps even more subtle—in the regression context. Specifically, we compare SHAP with Double ML. While both methods operate in data-rich environments and can leverage the same ML algorithms (e.g., random forests, gradient boosting), they address fundamentally different questions: SHAP asks "What did the model learn?" whereas Double ML asks "What is the causal effect?"

To make these differences concrete, we apply both methods to the same empirical setting: predicting AirBnB listing prices and estimating the causal effect of preview image quality on prices. Using data from Welsch and Kowalczyk (2023) on 2,096 AirBnB listings across 50 U.S. cities, we train predictive models (random forest and XGBoost) and apply SHAP to identify which features drive price predictions. SHAP reveals that image quality, measured by Mean Opinion Score (MOS), ranks as the 9th most important feature. We then use Double ML to estimate the causal effect of MOS on prices, controlling for a large set of confounding variables. The Double ML estimates (\$8.99–\$14.54 per MOS point, depending on method) are statistically significant but notably smaller than a naive interpretation of SHAP importance might suggest, highlighting the confounding present in observational data.

Our contribution is threefold. First, theoretically, while Shmueli et al. discuss counterfactuals primarily through classification examples (predicted classes), we demonstrate how their conceptual framework applies equally to regression settings, showing that the paradigmatic tensions between model-centric and outcome-centric reasoning, association and causation, and exploration and confirmation persist regardless of whether outcomes are categorical or continuous. Second, methodologically, we provide a detailed comparison of SHAP and Double ML, clarifying what each method assumes, what questions each can answer, and how they complement rather than substitute for one another. Third, empirically, we demonstrate these differences in a realistic application where

both prediction and CI are relevant, showing that SHAP and Double ML yield qualitatively similar patterns (both indicate image quality matters) but fundamentally different types of knowledge (associative importance vs. causal effect size).

By juxtaposing SHAP and Double ML both theoretically and empirically, we aim to promote methodological clarity in applied research. Too often, researchers treat feature importance from XAI methods as if it were equivalent to causal impact, or conversely, dismiss XAI methods as "merely correlational" without recognizing their value for hypothesis generation and model transparency—each of the authors of this paper has done this themselves. Our analysis shows that both paradigms are essential: SHAP provides the exploratory insights needed to identify candidate causal variables, while Double ML provides the confirmatory estimates needed for intervention planning. Understanding when to use each method—and how to integrate them into a coherent research workflow—is critical as ML increasingly informs decisions in business, medicine, and public policy.

II. TWO PARADIGMS OF EXPLANATION

A. The Fundamental Distinction

ML has achieved remarkable predictive performance across numerous applications, yet the resulting models often operate as "black boxes" whose decision-making processes remain opaque [59]. The demand for interpretability has spawned two fundamentally different approaches to understanding model behavior and relationships in data. These approaches differ not merely in their technical implementation, but in their epistemological foundations, research questions, and practical implications.

Recent work by Shmueli et al. [56] provides a comprehensive comparison of counterfactual reasoning in CI and XAI. Their theoretical framework illuminates key conceptual differences—particularly regarding purpose, causality versus correlation, and aggregation level—with a main focus on classification tasks. Our work seeks to complement their findings by focusing on regression problems. To this end, we provide empirical demonstration of these paradigmatic differences through the comparison of SHAP and Double ML.

Post-hoc Explanation using SHAP: The first paradigm begins with an already-trained prediction model. The workflow follows a data-driven sequence: collect data, train a model to maximize predictive performance, evaluate accuracy, and only then ask what the model has learned. The central question is descriptive: "Which features drive the model's predictions?" This approach is inherently inductive and exploratory. Researchers need not specify hypotheses in advance, instead, they examine the trained model to understand which patterns it has identified. SHAP exemplifies this paradigm by decomposing individual predictions into contributions from each input feature [42].

Importantly, while XAI literature often reserves the term "counterfactual explanation" for classification scenarios (e.g., "If feature X had value x' , you would have been classified differently"), the underlying counterfactual reasoning applies equally to regression: "If feature X had been higher, the predicted outcome Y would have increased." Our analysis focuses on this regression context, where SHAP provides feature attributions that describe associative relationships learned by the model.

A-priori CI using Double ML: The second paradigm starts not with a model, but with a specific research question about the relationship between variables. The workflow is theory-driven: formulate a hypothesis about a treatment effect, specify a causal model including relevant confounders, collect data, and estimate the effect of interest. The central question is inferential: "What is the effect of treatment X on outcome Y ?" This approach is deductive and confirmatory. Researchers must specify what they want to estimate before analyzing data. Double ML represents a modern implementation of this paradigm, combining the flexibility of machine learning for nuisance parameter estimation with the rigor of semiparametric inference for causal effects [17].

Table 1 summarizes the most important differences between SHAP and Double ML in particular, and thus XAI and CI in general. It becomes apparent that the two approaches look at (the same) data from two different directions and with two different goals.

TABLE 1: COMPARISON OF SHAP AND DOUBLE ML PARADIGMS

Dimension	SHAP	Double ML
Starting Point	Trained prediction model	Research question
What is explained?	Predictions of outcome variable	Outcome variable
Timing of question	After model training	Before modeling
Direction of reasoning	Model $\rightarrow$ Explanation	Question $\rightarrow$ Model $\rightarrow$ Estimate
Role of theory	Optional, post-hoc	Essential, a-priori
Scope	All features in model	Specific treatment(s)
Output	Feature attributions	Treatment effect
Type of relationship	Associative (what model learned)	Causal (under assumptions)
Statistical inference	Not available	Confidence intervals, p-values
Primary purpose	Exploration, hypothesis generation	Confirmation, hypothesis testing

In the following, we examine SHAP and Double ML in detail from a theoretical perspective.

B. SHAP: Model-Centric Explanation

Conceptual Foundation: SHAP builds on Shapley values, a concept from cooperative game theory introduced by Shapley [54]. In the game-theoretic setting, players form coalitions and receive payoffs; the Shapley value determines each player's fair share of the total payoff based on their marginal contributions across all possible coalitions. Lundberg and Lee [42] adapted this framework to machine learning: features are "players," the prediction is the "payoff," and the SHAP value of a feature represents its fair attribution of the difference between the prediction and a baseline value.

The key insight is that SHAP provides a unified framework for interpreting any machine learning model by satisfying a set of desirable axiomatic properties [42]. This model-agnostic approach allows researchers to explain predictions from complex models—such as gradient boosting,

random forests, or neural networks—using a consistent mathematical foundation.

Mathematical Formulation: Let $f: \mathbb{R}^p \rightarrow \mathbb{R}$ denote a trained prediction model that maps a feature vector $x = (x_1, \dots, x_p)$ to a predicted outcome $\hat{y} = f(x)$. SHAP aims to explain this prediction by decomposing it into additive contributions from each feature.

Additive Feature Attribution Model: Following Lundberg and Lee [42], SHAP employs an additive explanation model of the form:

$$g(z') = \varphi_0 + \sum_{i=1}^M \varphi_i z'_i$$

where $z' \in \{0,1\}^M$, M is the number of simplified input features, and $\varphi_i \in \mathbb{R}$ represents the attribution of feature i . Here, $z'_i = 1$ indicates that feature i is present (i.e. considered in the model), while $z'_i = 0$ means it is absent. The model g approximates the original black-box model f locally around a prediction of interest.

Shapley Value Definition: For a given instance x , the SHAP value φ_i of feature i is defined as:

$$\varphi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{[|S|! (|F| - |S| - 1)!]}{|F|!} \times [f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_{S(x_S)}]$$

where: $F = \{1, \dots, p\}$ is the set of all features, S is a subset (coalition) of features excluding feature i , $f_{S(x_S)}$ represents the expected prediction when only features in S are observed, and the weights ensure fair attribution across all possible feature orderings. Intuitively, φ_i computes the average marginal contribution of feature i across all possible orderings in which features could be added to form the full feature set.

Efficiency Property: The SHAP explanation satisfies the efficiency axiom:

$$\sum_{i=1}^p \varphi_i = f(x) - \varphi_0$$

where $\varphi_0 = E[f(X)]$ is the baseline prediction (expected value over the training distribution). This ensures that the sum of all feature attributions equals the difference between the prediction and the baseline [42].

Axiomatic Properties: SHAP values are uniquely characterized by four desirable axioms [42]:

- **Efficiency (Local Accuracy):** $\sum_{j=1}^p \varphi_j(x) = f(x) - E[f(X)]$; the sum of attributions equals the difference between prediction and baseline.
- **Symmetry:** If features j and k contribute equally to all coalitions, then $\varphi_j(x) = \varphi_k(x)$.
- **Dummy:** If feature j does not change the prediction regardless of which other features are present, then $\varphi_j(x) = 0$.
- **Additivity:** For two models f and g , $\varphi_j^{(f+g)}(x) = \varphi_j^f(x) + \varphi_j^g(x)$.

These axioms ensure that SHAP provides a theoretically sound and fair attribution method that aligns with human intuition about feature importance [42, 59].

Global Feature Importance: While SHAP values $\varphi_j(x)$ provide local explanations for individual predictions, they can be aggregated to assess global feature importance. A common measure is the mean absolute SHAP value:

$$SHAP_j = (1/n) \sum_{i=1}^n |\varphi_j(x_i)|$$

This metric quantifies the average magnitude of feature j 's contribution across all observations. Features with higher $SHAP_j$ values are considered more important for the model's predictions overall [43]. Dependence plots further enrich interpretation by visualizing how SHAP values vary with feature values, revealing non-linear relationships, threshold effects, and interactions [42].

C. Double ML: Outcome-Centric CI

Conceptual Foundation: CI aims to estimate the effect of a treatment or intervention on an outcome. The fundamental challenge is confounding: in observational data, treatment assignment is typically not random, so simple comparisons between treated and untreated groups conflate the treatment effect with selection bias. Classical approaches address confounding through regression adjustment, matching, or instrumental variables, but these methods often rely on parametric assumptions that may be violated when confounders are high-dimensional or relationships are non-linear [e.g., 34].

Double ML, introduced by Chernozhukov et al. [17], offers a modern solution. The core idea of controlling for confounders through partialling-out is not new. It dates back to Robinson [49] and builds on the classical Frisch-Waugh-Lovell theorem from econometrics [Frisch and Waugh (1933), Lovell (1963)]. What is new in Double ML is the use of flexible machine learning methods to estimate the two nuisance functions for the outcome and the treatment variable in high-dimensional settings, while preserving valid statistical inference for the treatment effect through Neyman-orthogonality and cross-fitting [17].

Unlike SHAP, which begins with a trained model and asks what it has learned, Double ML begins with a causal question: "What is the effect of treatment D on outcome Y , adjusting for confounders X ?"

The Partially Linear Model: Double ML considers the partially linear model

$$Y = \theta D + g(X) + \varepsilon,$$

where $Y \in \mathbb{R}$ is the outcome of interest, $D \in \mathbb{R}$ is the treatment variable, $X \in \mathbb{R}^p$ is a potentially high-dimensional vector of confounders, $\theta \in \mathbb{R}$ is the treatment effect parameter (low-dimensional, target of inference), $g: \mathbb{R}^p \rightarrow \mathbb{R}$ is an unknown nuisance function, and ε is a mean-zero error term [17].

The goal is to estimate θ while allowing $g(\cdot)$ to be arbitrarily complex. Traditional linear regression would impose $g(X) = X'\beta$, which often may be too restrictive. Double ML instead estimates $g(\cdot)$ non-parametrically using machine learning, without sacrificing valid inference about θ .

Estimation via Partialling-Out: The estimation proceeds in three steps, leveraging the classical partialling-out idea [49]:

1. **Estimate the nuisance functions:** Use machine models (e.g., random forests, gradient boosting) to estimate the nuisance functions $\ell(x) = E[Y | X = x]$ and $m(x) = E[D | X = x]$ for the outcome

and treatment variable. Let $\hat{\ell}(\cdot)$ and $\hat{m}(\cdot)$ denote the fitted models.

2. **Compute the residuals:** For each observation i compute the residualized outcome $\tilde{Y}_i = Y_i - \hat{\ell}(X_i)$ and residualized treatment $\tilde{D}_i = D_i - \hat{m}(X_i)$. The residuals $\tilde{Y}$ and $\tilde{D}$ are orthogonal to the confounders X .
3. **Estimate the treatment effect:** Regress the residualized outcome $\tilde{Y}$ on the residualized treatment variable $\tilde{D}$

$$\tilde{Y}_i = \theta \tilde{D}_i + v_i$$

to estimate the treatment effect θ . The Frisch-Waugh-Lovell theorem guarantees this yields the same estimate as including X directly in regression, but with added flexibility of non-parametric estimation for $g(\cdot)$ and $m(\cdot)$ [49].

Why "Double" ML? The term "Double ML" arises because machine learning is used twice: once to estimate $E[Y|X]$ and once to estimate $E[D|X]$. This dual use is essential for achieving Neyman-orthogonality, a property that makes $\hat{\theta}$ robust to small misspecifications in either nuisance function [17]. Specifically, if both nuisance functions estimate their targets at a rate faster than $n^{-1/4}$, then $\hat{\theta}$ achieves $\sqrt{n}$ -consistency and asymptotic normality, even if the nuisance functions are not perfectly specified.

Cross-Fitting: A critical innovation in Double ML is cross-fitting (i.e., sample splitting). When flexible machine learning methods are used to estimate nuisance parameters, regularization bias can arise if the same data are used both for training the models and making predictions, causing those predictions to become overly optimistic and contaminate the final parameter estimate.¹

Cross-fitting solves this by: (1) Randomly partitioning data into K folds (typically $K = 2$ or $K = 5$), (2) For each fold k : train the nuisance functions $\hat{\ell}$ and $\hat{m}$ on all folds except k , predict $\hat{Y} = \hat{\ell}(X_i)$ and $\hat{D} = \hat{m}(X_i)$ for observations in fold k , compute residuals $\tilde{Y}_i$ and $\tilde{D}_i$ for fold k , (3) Pool residuals across all folds and estimate the treatment effect $\hat{\theta}$. This ensures predictions are always out-of-sample, eliminating regularization bias from the nuisance parameter estimates [17].

Statistical Inference: Under regularity conditions (conditional ignorability, sufficient overlap, smoothness of nuisance functions), the Double ML estimator is $\sqrt{n}$ -consistent and asymptotically normal, i.e. $\sqrt{n}(\hat{\theta} - \theta) \rightarrow N(0, \sigma^2)$, where σ^2 can be consistently estimated. This allows construction of confidence intervals and hypothesis testing using classical t-statistics. Importantly, inference remains valid even when the number of confounders p is large relative to n , as long as nuisance functions can be estimated sufficiently well [17].

Causal Interpretation and Assumptions: The parameter θ estimates a conditional average treatment effect (CATE). For causal interpretation, the following assumptions are required, which are adapted to the specific regression context with a numerical outcome and treatment variable:

¹ This applies to flexible Double ML models using methods like XGBoost or Random Forests. Double Lasso, in contrast, addresses bias through its

specific theoretical framework that accounts for the variable selection process rather than through cross-fitting.

- **Conditional Ignorability (Unconfoundedness):** After conditioning on the confounders X , treatment assignment is independent of potential outcomes, formally $Y(d) \perp D \mid X$. This means that conditioning on X removes all confounding—there are no unmeasured confounders affecting both treatment and outcome.
- **Overlap (Common Support):** $0 < P(D = d \mid X = x) < 1$ for all x in the support of X . This ensures that for every observed covariate pattern, there is a positive probability of observing both treated and untreated units, making treatment effect estimation feasible across the entire covariate distribution.
- **SUTVA (Stable Unit Treatment Value Assumption):** (1) No interference between units—one unit's treatment does not affect another unit's outcome, and (2) treatment consistency – there is only one version of each treatment level (no hidden variations in treatment implementation).

Without conditional ignorability, the treatment effect θ does not have a causal interpretation but still represents the "best linear predictor" of Y given D after partialling out X —a useful predictive relationship even if not strictly causal [17].

D. SHAP vs. Double ML: Questions addressed and prerequisites

Having described the technical foundations of SHAP and Double ML, we now synthesize their fundamental differences in terms of the research questions they address and the prerequisites for their application. This synthesis integrates and extends the conceptual framework proposed by Shmueli et al. [56] for the regression context.

Questions Addressed: SHAP operates within the space of a trained predictive model, asking which features contributed to predictions and by how much. It reveals which patterns the model discovered during training, including potential spurious correlations and confounding. SHAP answers questions about the model's internal logic without reference to the real-world data-generating process [56, 59].

Double ML addresses a fundamentally different question that precedes model training: What is the magnitude of the causal effect of treatment D on outcome Y , after accounting for confounders X ? This requires a priori specification of treatment, outcome, and confounders based on domain knowledge. Under appropriate causal assumptions, θ estimates what would happen if we actually changed D in reality [17, 56].

The Direction of Reasoning: SHAP uses an inductive, exploratory process: (1) collect data, (2) train the predictive f model to maximize accuracy, (3) evaluate performance, (4) apply SHAP to understand what f learned. In short: the model comes first, explanation comes second.

Double ML uses a deductive, confirmatory process: (1) formulate a causal question about the outcome Y and the treatment D , (2) identify confounders X , (3) train ML models for the nuisance functions, (4) estimate the causal effect θ . In short: the research question comes first and models are tools in service of answering that question [17, 56].

Prerequisites and Assumptions: The key prerequisites and assumptions for SHAP and Double ML are summarized in Table 2. The critical distinction is that SHAP requires no

causal assumptions and can be applied to any well-performing predictive model, while Double ML requires strong theoretical commitments about the causal structure and relies on untestable assumptions (particularly conditional ignorability). However, only Double ML provides causal estimates suitable for intervention planning.

TABLE 2: PREREQUISITES AND ASSUMPTIONS

Aspect	SHAP	Double ML
Required Input	Trained predictive model $f(x)$	Pre-specified treatment D , outcome Y , confounders X
Key Assumptions	Model well-specified for prediction; features interpretable	Conditional ignorability (unconfoundedness); overlap; SUTVA
Domain Knowledge	Not required (model-driven exploration)	Essential (theory-driven specification of causal structure)
Interpretation	Associative (correlations in model)	Causal (under conditional ignorability)
Testability	Model performance (MSE, R^2) testable	Key assumption (unconfoundedness) untestable; inference via SE, CI

Complementarity: SHAP and Double ML are complementary tools addressing different stages of knowledge generation. SHAP is appropriate when you have a well-performing predictive model and seek to understand its behavior, generate hypotheses about potentially causal variables, or communicate model decisions to stakeholders. It excels in exploratory analysis without requiring prior theoretical commitments.

Double ML is appropriate when you have a specific causal question, can measure relevant confounders (or argue unconfoundedness is plausible), and need actionable estimates for intervention planning with statistical inference (confidence intervals, hypothesis tests). It requires sufficient domain knowledge to specify the causal model a priori.

One way to integrate SHAP and Double ML is to use SHAP for identifying highly predictive features, which can then motivate a targeted Double ML analysis to test for causal relationships. Conversely, Double ML findings help contextualize SHAP results: if Double ML estimates a small causal effect while SHAP attributes large importance to the same feature, this discrepancy reveals potential confounding or spurious correlations that merit further investigation.

Paradigmatic Differences: There also paradigmatic differences between SHAP and Double ML. While SHAP is model-centric, i.e. it explains the black box, taking the model as the object of inquiry, Double ML is outcome-centric explaining the world, using models as tools. This distinction echoes the broader divide between "explaining" and "predicting" in statistical modeling [55]. Both paradigms are essential: SHAP provides transparency and generates insights from complex models; Double ML provides rigor and actionable causal estimates. Together, they form a complete toolkit for extracting knowledge from data.

E. Key Takeaways

SHAP and Double ML embody two distinct paradigms. SHAP operates post-hoc, taking a trained model as given and explaining its predictions through feature attributions. It is model-centric, exploratory, and well-suited for hypothesis generation and model transparency. Double ML operates a-priori, starting with a causal question and constructing an estimation strategy to answer it. It is outcome-centric, confirmatory, and designed for hypothesis testing and intervention planning.

Crucially, SHAP explains the model whereas Double ML explains the world (under causal assumptions). SHAP reveals associations learned by the model, while Double ML estimates the effect of changing a variable in reality. Understanding this distinction is essential for choosing the right tool and interpreting results appropriately. The next section demonstrates these differences empirically using Airbnb listing price data.

III. EMPIRICAL APPLICATION: AIRBNB IMAGE QUALITY AND PRICING

A. Research Context and Data

To demonstrate the paradigmatic differences between SHAP and Double ML, we analyze the effect of preview image quality on Airbnb listing prices. This application is particularly suitable because image quality represents an actionable intervention for hosts—unlike fixed property characteristics such as location or size—making the distinction between associative and causal relationships practically relevant.

Data Collection and Preparation: Following Welsch and Kowalczyk [59], we utilize Airbnb listing data collected via the Airbnb API from the 50 most populated cities in the USA. The initial sample comprised 2,500 listings (50 per city), with price data collected for one year in advance. After removing listings with missing values and duplicates, and excluding 16 outliers with prices exceeding \$500 per night, the final dataset contains 2,096 listings. The data were randomly split into a training ($n=1,886$) and test set ($n=210$).

The analysis examines nightly rental prices (in USD) as the outcome variable. For prediction tasks and SHAP analysis, all 47 features serve as potential predictors, including the Mean Opinion Score (MOS) of the preview image alongside 46 additional characteristics such as location features (distance to city center, population density, city population), property attributes (property type, capacity, bedrooms, bathrooms, amenities), host information (verification status, cancellation policy), and other listing characteristics (number of photos, minimum/maximum nights).

The Mean Opinion Score is a standardized metric for perceived image quality, widely used in image processing and computer vision [31]. The MOS ranges from 1 (lowest quality) to 5 (highest quality) and is computed using a convolutional neural network trained to predict human perceptions of image quality [31]. For each Airbnb listing, we apply the MOS algorithm to the preview image—the first image displayed in search results—which represents a critical determinant of user engagement and booking decisions. Unlike objective image metrics (e.g., resolution, file size), the MOS captures subjective quality dimensions such as lighting, composition, color balance, and visual appeal, making it

particularly relevant for understanding how image presentation affects consumer choices in online marketplaces.

For the Double ML causal analysis, the Mean Opinion Score of the preview image serves as the treatment variable of interest, while the remaining 46 features function as confounders that must be controlled for to identify the causal effect of image quality on rental prices.

B. Prediction Models and Performance

We train multiple machine learning models to predict Airbnb prices, following the approach of Welsch and Kowalczyk [59]. The models include baseline models (Dummy mean/median, Linear Regression, Decision Tree) and advanced ML models with focus on Random Forest (RF) and XGBoost. All models were trained on 1,886 training observations and evaluated on the held-out test set (210 observations). Hyperparameters for RF, XGBoost, and Artificial Neural Network (ANN) were optimized using 5-fold cross-validation combined with grid search.

Model Performance: Table 3 presents the predictive performance of all models, measured by Root Mean Squared Error (RMSE). Lower RMSE values indicate better predictive accuracy.

TABLE 3: MODEL PERFORMANCE (RMSE ON TEST SET)

Model ^a	RMSE ^b
Dummy (mean)	53.47
Dummy (median)	54.38
Linear Regression	44.68
Decision Tree	48.36
Artificial Neural Network	45.66
Random Forest	42.66
XGBoost	41.73

^a Adapted from Welsch and Kowalczyk [59]. ^b RMSE = Root Mean Squared Error

The results reveal that ensemble methods substantially outperform baseline and simple models as well as the ANN. The comparatively weak performance of the ANN, which even falls short of linear regression, may suggest that the chosen network architecture is not well suited to the data, that the sample size is too limited to exploit the ANN's full potential, and that tree-based ensemble methods are better able to capture nonlinear patterns in this tabular setting. XGBoost achieves the lowest RMSE (41.73), followed closely by Random Forest (42.66). These models serve as the basis for our subsequent SHAP and Double ML analyses.

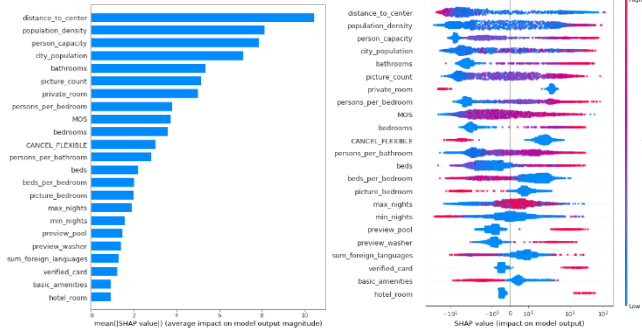
The superior performance of tree ensembles reflects their ability to capture non-linear relationships and complex interactions between features—precisely the characteristics that make them "black boxes" requiring post-hoc explanation methods like SHAP.

C. SHAP Analysis: Post-hoc Explanation

Research Question: Using SHAP, we ask: "Which features are associated with price predictions in the XGBoost model?" This is a model-centric, exploratory question. We do not pre-specify hypotheses; instead, we examine what patterns the trained model has learned.

SHAP Results for MOS: We apply SHAP to the trained XGBoost model to compute feature attributions. Results are exhibited in Figure 1. The SHAP analysis reveals that MOS (preview image quality) ranks 9th in global feature importance among all 47 features. The SHAP dependence plot for MOS shows a positive association: higher MOS values (better image quality) are consistently associated with higher predicted prices. The relationship exhibits non-linearity with diminishing returns—marginal improvements in MOS have larger effects at low-to-medium quality levels than at high quality levels.

Figure 1: Results from SHAP Analysis



Interpretation: The SHAP analysis suggests that image quality is predictively important—the XGBoost model relies on MOS when generating price predictions. Listings with higher-quality preview images tend to have higher predicted prices.

This finding, however, is only associative and not causal. SHAP tells us what the model learned, not what would happen if a host improved their image quality. Several confounding mechanisms could explain this association: hosts with higher-quality images may also have higher-quality properties (confounding via property quality), professional hosts may invest in both better properties and better photography (confounding via host professionalism), or higher-priced listings may attract more professional photography (reverse causality). To answer the causal question—"Does improving image quality cause higher prices?"—we turn to Double ML.

D. Double ML Analysis: A-priori CI

Research Question and Causal Model: Using Double ML, we ask a different question: "What is the causal effect of preview image quality (MOS) on listing prices?" This is an outcome-centric, confirmatory question. We specify the treatment (MOS), outcome (Price), and confounders (all other features) before estimation. The estimated model then is:

$$Price = \theta \cdot MOS + g(X) + \varepsilon,$$

where θ is the average treatment effect (ATE) of MOS on Price, $g(X)$ is a flexible function of the 46 confounders, and we assume conditional ignorability. This assumption is plausible if X includes all variables that jointly determine image quality and price. However, omitted variable (i.e. unmeasured confounding) and selection bias remain a limitation.

Double ML Estimation: We estimate the treatment effect θ using three variants of Double ML: (1) Double Lasso uses Lasso, (2) Double ML (RF) uses Random Forest, and (3) Double ML (XGBoost) uses XGBoost to estimate both nuisance functions. All estimators employ 5-fold cross-fitting to eliminate overfitting bias. Inference is based on heteroskedasticity-robust standard errors.

Results: Results for the treatment effect of MOS on Price using the three Double ML models are displayed in Table 4.

TABLE 4: DOUBLE ML RESULTS—TREATMENT EFFECT OF MOS ON PRICE

Method ^a	Estimate ($\hat{\theta}$) ^b	Standard error (SE) ^c	t-value	p-value ^d
Double Lasso	14.54	4.81	3.02	0.003**
Double ML (RF)	10.23	4.47	2.29	0.022*
Double ML (XGBoost)	8.99	4.36	2.06	0.039*

^a All models use 5-fold cross-fitting. ^b Estimate ($\hat{\theta}$) = average treatment effect of a 1-point increase in MOS on price (USD). ^c SE = standard error. ^d * $p < 0.05$, ** $p < 0.01$.

All three Double ML estimators yield positive and statistically significant effects of image quality on listing prices, controlling for 46 confounding variables:

- **Double Lasso** estimates that a 1-point increase in MOS increases prices by \$14.54. This method uses L1-regularized linear models for nuisance functions, which may be restrictive if relationships are highly non-linear.
- **Double ML (Random Forest)** estimates an effect of \$10.23. Random Forest allows for non-linearities and interactions, yielding a somewhat smaller but still significant estimate.
- **Double ML (XGBoost)** estimates an effect of \$8.99. This is the most conservative estimate, consistent with XGBoost's superior predictive performance potentially capturing confounding more effectively.

Consistency Across Methods: The estimates range from \$8.99 to \$14.54, with overlapping confidence intervals. This consistency suggests robustness—the positive effect of image quality on prices is not an artifact of a particular estimation method. The downward trend from Lasso (14.54) to RF (10.23) to XGBoost (8.99) likely reflects increasingly flexible modeling of confounders, with more sophisticated methods isolating the direct causal effect more precisely by better accounting for non-linear confounding relationships.

Statistical Significance: All three estimates are statistically significant at the 5% level, with Double Lasso significant at the 1% level ($p = 0.003$). We confidently reject the null hypothesis of no effect: preview image quality affects listing prices, even after controlling for observed heterogeneity.

Causal Interpretation of the Treatment Effect: The estimated treatment effect of image quality (MOS) on nightly rental prices may be interpreted as approximately causal under certain conditions. First, we control for a comprehensive set of confounding variables. Second, through Double ML with increasingly flexible machine learning methods (Lasso, Random Forest, XGBoost), we account for complex nonlinear relationships and interactions among confounders, minimizing the risk of functional form misspecification that could bias the treatment effect estimate.

However, even after controlling for all observed features, the possibility of unobserved confounding remains. For instance, hosts may have an unobserved motivation to maximize their rental prices, which could lead them to deliberately invest in higher-quality images. If such unobserved motivations simultaneously affect both image quality and pricing decisions, our estimate would capture not

only the direct causal effect of image quality but also this confounding pathway.

To address such limitations, future research could employ panel data combined with Double ML methods for fixed effects models [18]. These approaches difference out time-invariant unobserved heterogeneity—such as a host's persistent motivation to optimize pricing—thereby enabling more robust causal identification of the treatment effect. Such extensions would represent a promising avenue for strengthening the causal claims in this application domain.

Comparison of SHAP and Double ML analyses:

- **SHAP Analysis (post-hoc, model-centric):** MOS ranked 9th among 47 features in the XGBoost model. The SHAP dependence plot showed a positive association between MOS and predicted prices, with higher-quality images consistently linked to higher predictions. However, SHAP does not quantify a specific marginal effect—it shows relative importance and directional patterns, not effect sizes in interpretable units.
- **Double ML Analysis (a-priori, outcome-centric):** Controlling for all confounders, a 1-point increase in MOS causes prices to increase by \$8.99–\$14.54, depending on the method. The XGBoost-based estimate (\$8.99) is the most conservative and arguably most reliable, given XGBoost's superior accuracy.

The insights show that SHAP and Double ML address different questions: (1) SHAP: "How important is MOS for the model's predictions?" → MOS contributes meaningfully to prediction accuracy, but this includes both causal effects and confounding. (2) Double ML: "What happens if a host improves image quality by 1 MOS point?" → Prices increase by approximately \$9–\$15, holding all else constant.

SHAP identifies MOS as predictively relevant (exploratory finding), but only Double ML quantifies the actionable effect size (confirmatory finding). A host considering hiring a professional photographer can use the Double ML estimate (\$9–\$15 per MOS point) to calculate return on investment—SHAP alone would be insufficient for this decision.

Practical Implication: For intervention planning, use the Double ML estimate (\$8.99–\$14.54 per MOS point). For hypothesis generation or model debugging, use SHAP to identify which features matter for predictions. The two methods are complementary, not substitutes.

IV. DISCUSSION

A. Extending Counterfactual Reasoning to Regression

This paper applies and extends the theoretical framework of Shmueli et al. [56] on counterfactual reasoning in CI and XAI to regression problems. While Shmueli et al.'s framework is general, their XAI counterfactual examples focus on classification tasks—explaining categorical predictions ("If feature X were different, you would have been classified as Y"). Our comparison of SHAP and Double ML in a regression setting reveals that the fundamental tension between model-centric explanation and outcome-centric CI persists regardless of whether the outcome is categorical or continuous.

The AirBnB application illustrates this tension concretely: SHAP identifies image quality (MOS) as the 9th most

important feature for price predictions—a useful exploratory finding that describes what the XGBoost model learned. However, this associative insight cannot answer the intervention question: "By how much would improving image quality increase prices?" Double ML addresses this causal question directly, estimating that a 1-point improvement in MOS increases prices by \$8.99–\$14.54, holding confounders constant. Both methods provide valid but fundamentally different forms of knowledge.

B. Implications for future Research and Practice

We identify three key findings for future research and practice:

Methodological Clarity. Researchers must be explicit about whether they seek to explain model behavior (as done using SHAP) or estimate causal effects (as done using Double ML). Conflating these goals—for instance, interpreting SHAP importance as causal impact—leads to flawed inferences and potentially harmful recommendations. Our results show that SHAP and Double ML can yield qualitatively similar patterns (both indicate MOS matters), but only Double ML provides the quantitative effect size necessary for cost-benefit analysis and intervention planning.

Complementary Workflows. SHAP and Double ML are not competitors but can be complementary stages in a research pipeline. The optimal workflow proceeds from exploration (SHAP identifies candidate variables) to confirmation (Double ML tests causal hypotheses). In our application, SHAP's identification of MOS as predictively important motivated the subsequent Double ML analysis. Without SHAP, we might not have used image quality for causal investigation; without Double ML, we could not quantify the return on investment for professional photography.

Generalizability Beyond the AirBnB Example. While we focus on pricing in online marketplaces, the paradigmatic distinction generalizes to any domain where both prediction and CI are relevant. In healthcare, SHAP might reveal which biomarkers drive a disease risk model's prediction, while Double ML estimates the causal effect of a specific treatment. In marketing, SHAP explains why a customer churn model flags certain users, while Double ML quantifies the impact of retention interventions. The key insight—that models can be predictive without being causal—applies universally.

C. Limitations and Future Directions

Our analysis inherits limitations from both methods. SHAP explanations reflect the training data's correlation structure, including any biases or measurement errors. If hosts who invest in professional photography also differ systematically on unobserved dimensions (e.g., attention to detail in property maintenance), SHAP will attribute importance to MOS even if the causal effect is small. This is not a flaw of SHAP—it accurately describes the model—but a limitation of interpreting associative patterns causally.

Double ML relies on conditional ignorability (unconfoundedness), which requires that all confounders affecting both image quality and prices are measured and included in X. While our 46 control variables cover a broad range of property, location, and host characteristics, unmeasured factors (e.g., host marketing sophistication, property aesthetic consistency beyond the preview image) could bias the estimates. This assumption is untestable, and violations would inflate the estimated effect of MOS.

Sensitivity analyses for unobserved confounding would strengthen future applications.

Future research should explore several directions: (1) Extend the comparison to other XAI methods (LIME, counterfactual explanations for regression) and CI approaches (instrumental variables, regression discontinuity). (2) Investigate settings where SHAP and Double ML diverge more dramatically, which would reveal the extent of confounding. (3) Develop integrated frameworks that combine exploratory and confirmatory phases, potentially using SHAP-based variable selection to inform Double ML specifications. (4) Apply the framework to high-stakes domains (medicine, criminal justice) where distinguishing correlation from causation has ethical implications.

V. CONCLUSION

This paper demonstrates that XAI and CI represent two fundamentally different paradigms for extracting knowledge from data, even when applied to the same problem using the same machine learning algorithms. By comparing SHAP and Double ML in the context of Airbnb pricing, we show that these methods address distinct questions with different assumptions, prerequisites, and interpretations. SHAP operates post-hoc, explaining what a trained predictive model learned from data—including spurious correlations and confounding. Double ML operates a-priori, estimating causal effects by design, requiring theoretical commitments about treatment, outcome, and confounders before seeing the data.

Our contribution applies the theoretical framework of Shmueli et al. [56] to regression problems. While Shmueli et al. illustrate how counterfactuals function differently in CI versus XAI primarily through classification examples, we demonstrate that the same paradigmatic tensions persist—and manifest in nuanced ways—when working with continuous outcomes. In regression settings, XAI counterfactuals (via SHAP) explain model behavior (*"If feature X had been higher, the predicted outcome would increase"*), while CI counterfactuals (via Double ML) estimate causal effects under identifying assumptions (*"By how much would outcome Y change if we intervened to change treatment D?"*). Both involve counterfactual reasoning, yet serve fundamentally different purposes: SHAP explains the model's associations, while Double ML aims to explain real-world causal mechanisms.

The Airbnb application illustrates these differences concretely. SHAP identifies image quality (MOS) as the 9th most important feature among 47 predictors, revealing that the XGBoost model relies heavily on this variable when generating price predictions. This is valuable exploratory insight that could motivate further investigation. However, SHAP cannot answer the intervention question: "Should I hire a professional photographer?" Double ML addresses this directly, estimating that improving image quality by one MOS point increases prices by \$8.99–\$14.54 on average, controlling for property characteristics, location, and host attributes. Both findings are valid within their respective paradigms, but only the causal estimate supports cost-benefit analysis and decision-making.

For researchers and practitioners, our analysis yields three key takeaways. First, methodological clarity is essential: feature importance from SHAP is not equivalent to causal impact, and conflating these concepts leads to flawed inferences and potentially harmful recommendations. Second,

SHAP and Double ML are complementary, not competing: the optimal research workflow proceeds from exploration (SHAP identifies candidates) to confirmation (Double ML tests causal hypotheses). Third, the distinction generalizes beyond Airbnb to any domain where both prediction and CI matter—healthcare, marketing, criminal justice, public policy—making this framework broadly applicable.

ACKNOWLEDGEMENT

Large language models, specifically Claude and ChatGPT, were used solely to support the formulation and linguistic refinement of the manuscript text. The AI systems were not used to generate the study design, research results, analysis, interpretations, figures, or code. All content was reviewed, revised where necessary, and approved by the authors.

References

- [1] S. Abd Elhafeez, C. Torino, G. D'Arrigo, D. Bolignano, F. Provenzano, F. Mattace-Raso, C. Zoccali, and G. Tripepi, "Unveiling the potential of artificial intelligence in revolutionizing disease diagnosis and prediction: a comprehensive review of machine learning and deep learning approaches," *European Journal of Medical Research*, vol. 30, no. 1, p. 59, 2025.
- [2] M. F. Aljunid, A. Manjramkar, and M. Narvekar, "A collaborative filtering recommender systems: survey," *Neurocomputing*, vol. 611, p. 128573, 2024.
- [3] M. F. Aljunid and A. Huchzermeier, "Machine learning powered financial credit scoring: a systematic literature review," *Artificial Intelligence Review*, vol. 58, Art. no. 89, 2025.
- [4] M. Alkhatib, G. Younes, L. Kanaan, and M. O. Diab, "Machine learning in healthcare," *Annual Review of Biomedical Data Science*, vol. 5, pp. 291–313, 2022.
- [5] A. M. Alonso and A. Carballo, "Machine learning for enhanced credit risk assessment: an empirical approach," *Journal of Risk and Financial Management*, vol. 16, no. 12, p. 496, 2023.
- [6] J. D. Angrist and A. B. Krueger, "Does compulsory school attendance affect schooling and earnings?" *The Quarterly Journal of Economics*, vol. 106, no. 4, pp. 979–1014, 1991.
- [7] J. D. Angrist and J.-S. Pischke, "The credibility revolution in empirical economics: how better research design is taking the con out of econometrics," *Journal of Economic Perspectives*, vol. 24, no. 2, pp. 3–30, 2010.
- [8] S. Athey and G. W. Imbens, "Recursive partitioning for heterogeneous causal effects," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 113, no. 27, pp. 7353–7360, 2016.
- [9] S. Athey and G. W. Imbens, "The state of applied econometrics: causality and policy evaluation," *Journal of Economic Perspectives*, vol. 31, no. 2, pp. 3–32, 2017.
- [10] A. Belloni, V. Chernozhukov, and C. Hansen, "Inference on treatment effects after selection among high-dimensional controls," *Review of Economic Studies*, vol. 81, no. 2, pp. 608–650, 2014.
- [11] Y. Bengio, Y. LeCun, and G. Hinton, "Deep learning for AI," *Communications of the ACM*, vol. 64, no. 7, pp. 58–65, 2021.
- [12] M. Bhandari, A. Neupane, and B. Joshi, "Machine-learning-based disease diagnosis: a comprehensive review," *Diagnostics*, vol. 12, no. 7, p. 1697, 2022.
- [13] T. Blake, C. Nosko, and S. Tadelis, "Consumer heterogeneity and paid search effectiveness: a large-scale field experiment," *Econometrica*, vol. 83, no. 1, pp. 155–174, 2015.
- [14] J. Bughin, J. Seong, J. Manyika, M. Chui, and R. Joshi, Notes from the AI frontier: Modeling the impact of AI on the world economy. McKinsey Global Institute, 2018.
- [15] E. Brynjolfsson and A. McAfee, "What's driving the machine learning explosion?" *Harvard Business Review*, Jul. 2017. [Online]. Available: <https://hbr.org/2017/07/whats-driving-the-machine-learning-explosion>
- [16] D. Card and A. B. Krueger, "Minimum wages and employment: a case study of the fast-food industry in New Jersey and Pennsylvania," *American Economic Review*, vol. 84, no. 4, pp. 772–793, 1994.

- [17] V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins, "Double/debiased machine learning for treatment and structural parameters," *The Econometrics Journal*, vol. 21, no. 1, pp. C1–C68, 2018.
- [18] P. S. Clarke and A. PolSELLi, "Double machine learning for static panel models with fixed effects," *The Econometrics Journal*, vol. 29, no. 1, pp. 69–86, 2026.
- [19] X. Dastile, T. Celik, and M. Potsane, "Machine learning-driven credit risk: a systemic review," *Neural Computing and Applications*, vol. 34, no. 17, pp. 14327–14346, 2022.
- [20] A. Da'u and N. Salim, "A collaborative filtering recommendation framework utilizing social networks," *Intelligent Systems with Applications*, vol. 20, p. 200268, 2023.
- [21] E. Dumitrescu, S. Hue, C. Hurlin, and S. Tokpavi, "Machine learning for credit scoring: improving on the age-old credit scoring models," *Journal of Financial Data Science*, vol. 4, no. 3, pp. 72–92, 2022.
- [22] A. C. Eggers, A. Fowler, J. Hainmueller, A. B. Hall, and J. M. Snyder Jr., "On the validity of the regression discontinuity design for estimating electoral effects: new evidence from over 40,000 close races," *American Journal of Political Science*, vol. 59, no. 1, pp. 259–274, 2015.
- [23] P. B. Ellickson, P. L. E. Grieco, and O. Khvastunov, "Measuring competition in spatial retail," *The RAND Journal of Economics*, vol. 53, no. 1, pp. 189–232, 2022.
- [24] European Parliament and Council of the European Union, "Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)," *Official Journal of the European Union*, L 119, pp. 1–88, Apr. 2016.
- [25] G. Fernandes and R. Artes, "Advancing credit rating prediction: the role of machine learning in corporate credit rating assessment," *Risks*, vol. 13, no. 6, p. 116, 2025.
- [26] B. R. Gordon, F. Zettelmeyer, N. Bhargava, and D. Chapsky, "A comparison of approaches to advertising measurement: evidence from big field experiments at Facebook," *Marketing Science*, vol. 38, no. 2, pp. 193–225, 2019.
- [27] S. Grigorescu, B. Trasnea, T. Cocias, and G. Macesanu, "A survey of deep learning techniques for autonomous driving," *Journal of Field Robotics*, vol. 37, no. 3, pp. 362–386, 2020.
- [28] F. Hasanzadeh, C. B. Josephson, G. Waters, et al., "Bias recognition and mitigation strategies in artificial intelligence healthcare applications," *npj Digital Medicine*, vol. 8, Art. no. 154, 2025.
- [29] T. Hassan, A. M. Qamar, and M. A. Dar, "Deep learning-based multi-criteria recommender system for technology-enhanced learning," *Scientific Reports*, vol. 15, p. 6829, 2025.
- [30] M. A. Hernán and J. M. Robins, "Estimating causal effects from epidemiological data," *Journal of Epidemiology & Community Health*, vol. 60, no. 7, pp. 578–586, 2006.
- [31] V. Hosu, H. Lin, T. Sziranyi, and D. Saupe, "KonIQ-10k: an ecologically valid database for deep learning of blind image quality assessment," *IEEE Transactions on Image Processing*, vol. 29, pp. 4041–4056, 2020.
- [32] G. W. Imbens and J. D. Angrist, "Identification and estimation of local average treatment effects," *Econometrica*, vol. 62, no. 2, pp. 467–475, 1994.
- [33] G. W. Imbens, "Nonparametric estimation of average treatment effects under exogeneity: a review," *Review of Economics and Statistics*, vol. 86, no. 1, pp. 4–29, 2004.
- [34] G. W. Imbens and J. M. Wooldridge, "Recent developments in econometrics of program evaluation," *Journal of Economic Literature*, vol. 47, no. 1, pp. 5–86, 2009.
- [35] C. Janiesch, P. Zschech, and K. Heinrich, "Machine learning and deep learning," *Electronic Markets*, vol. 31, pp. 685–695, 2021.
- [36] A. E. Khandani, A. J. Kim, and A. W. Lo, "Consumer credit-risk models via machine-learning algorithms," *Journal of Banking & Finance*, vol. 34, no. 11, pp. 2767–2787, 2010.
- [37] A. Krenzer, M. Banck, K. Makowski, A. Hekalo, D. Fitting, J. Troya, B. Sudarevic, W. G. Zoller, A. Hann, and F. Puppe, "A real-time polyp-detection system with clinical application in colonoscopy using deep convolutional neural networks," *Journal of Imaging*, vol. 9, no. 2, p. 26, 2023.
- [38] S. Kuutti, R. Bowden, Y. Jin, P. Barber, and S. Fallah, *Deep Learning for Autonomous Vehicle Control: Algorithms, State-of-the-Art, and Future Prospects*. Cham, Switzerland: Springer International Publishing, 2021.
- [39] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.
- [40] R. A. Lewis and J. M. Rao, "The unfavorable economics of measuring the returns to advertising," *The Quarterly Journal of Economics*, vol. 130, no. 4, pp. 1941–1973, 2015.
- [41] M. Lipsitch, E. T. Tchetgen Tchetgen, and T. Cohen, "Negative controls: a tool for detecting confounding and bias in observational studies," *Epidemiology*, vol. 21, no. 3, pp. 383–388, 2010.
- [42] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems*, vol. 30, 2017, pp. 4765–4774.
- [43] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee, "From local explanations to global understanding with explainable AI for trees," *Nature Machine Intelligence*, vol. 2, no. 1, pp. 56–67, 2020.
- [44] K. Muhammad, A. Ullah, J. Lloret, J. Del Ser, and V. H. C. de Albuquerque, "Deep learning for object detection and scene perception in self-driving cars: survey, challenges, and open issues," *Array*, vol. 10, Art. no. 100057, 2021.
- [45] M. Nour, Ü. Şahin, and H. Polat, "A fused deep neural network and matrix factorization for multi-criteria collaborative filtering," *Neural Computing and Applications*, 2023.
- [46] J. Pearl, *Causality: Models, Reasoning, and Inference*, 2nd ed. Cambridge, U.K.: Cambridge University Press, 2009.
- [47] M. T. Ribeiro, S. Singh, and C. Guestrin, "Model-agnostic interpretability of machine learning," *arXiv preprint arXiv:1606.05386*, 2016.
- [48] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you?: explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 1135–1144.
- [49] P. M. Robinson, "Root-N-consistent semiparametric regression," *Econometrica*, vol. 56, no. 4, pp. 931–954, 1988.
- [50] D. B. Rubin, "Estimating causal effects of treatments in randomized and nonrandomized studies," *Journal of Educational Psychology*, vol. 66, no. 5, pp. 688–701, 1974.
- [51] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Machine Intelligence*, vol. 1, no. 5, pp. 206–215, 2019.
- [52] C. Rudin and B. Ustun, "Optimized scoring systems: toward trust in machine learning for healthcare and criminal justice," *Interfaces*, vol. 48, no. 5, pp. 449–466, 2018.
- [53] J. S. Sekhon, "Opiates for the matches: matching methods for causal inference," *Annual Review of Political Science*, vol. 12, pp. 487–508, 2009.
- [54] L. S. Shapley, "A value for n-person games," in *Contributions to the Theory of Games, Volume II*, H. W. Kuhn and A. W. Tucker, Eds. Princeton, NJ, USA: Princeton University Press, 1953, pp. 307–317.
- [55] G. Shmueli, "To explain or to predict?" *Statistical Science*, vol. 25, no. 3, pp. 289–310, 2010.
- [56] G. Shmueli, D. Martens, J. Yoo, and T. Greene, "From what ifs to insights: counterfactuals in causal inference vs. explainable AI," *arXiv preprint arXiv:2505.13324*, 2025.
- [57] T. J. VanderWeele and P. Ding, "Sensitivity analysis in observational research: introducing the E-value," *Annals of Internal Medicine*, vol. 167, no. 4, pp. 268–274, 2017.
- [58] S. Wager and S. Athey, "Estimation and inference of heterogeneous treatment effects using random forests," *Journal of the American Statistical Association*, vol. 113, no. 523, pp. 1228–1242, 2018.
- [59] G. Welsch and P. Kowalczyk, "Designing explainable predictive machine learning artifacts: methodology and practical demonstration," *arXiv preprint arXiv:2306.11771*, 2023.
- [60] H. Zhang, A. X. Lu, M. Abdalla, M. McDermott, and M. Ghassemi, "Interpretability and fairness evaluation of deep learning models on MIMIC-IV dataset," *Scientific Reports*, vol. 12, p. 7166, 2022.

The ADM-AI Model: Agile Development Model for AI Competences in European Higher Education Curricula

1st Prof. Dr. Ulf-Daniel Ehlers

School of Economy

Baden-Wuerttemberg Cooperative State University

Karlsruhe, Germany

ulf-daniel.ehlers@dhbw-karlsruhe.de

2nd Julia Marcia Mann

School of Economy

Baden-Wuerttemberg Cooperative State University

Karlsruhe, Germany

juliamarcia.mann@dhbw-karlsruhe.de

Abstract—Artificial intelligence is rapidly transforming disciplinary practices and professional environments, challenging higher education institutions to integrate AI competences into existing study programmes. In European contexts shaped by the Bologna Process, curriculum reform must align with modular structures and accreditation regimes. This paper addresses how AI competences can be embedded into higher education curricula through agile yet governance-compatible development processes. Based on an integrative review of curriculum theory, agile and iterative development theory, educational design research, technology integration frameworks, and competence-based education, we develop a four-phase Agile Curriculum Development Model. The model combines competence clarification, constructive alignment, curriculum mapping, and a time-boxed implementation cycle. It is anchored in the AIComp competence framework, providing a structured orientation for AI-related graduate capabilities. Rather than proposing disruptive reform, the framework enables embedded, iterative innovation within existing programme structures and supports evidence-based institutional learning under regulatory stability. The model specifies phase-specific design artefacts and decision gates that enable “below-threshold” innovation while remaining compatible with accreditation and programme governance.

Index Terms—curriculum development, AI competence model, framework, future skills, agile integration

I. INTRODUCTION

Artificial Intelligence is transforming professional fields at unprecedented speed. Higher education institutions are therefore under increasing pressure to integrate AI competence into existing study programmes. However, European universities operate within the structural logic of the Bologna Process, which emphasises stability, accreditation cycles, and formal programme structures. This creates a structural tension between innovation dynamics and institutional regulation. Most AI-related curriculum initiatives focus on new standalone modules or specialised programmes. What is largely missing is a structured approach for systematically embedding AI competence across existing programmes through iterative, small-scale and participatory development processes below the accreditation threshold.

This paper addresses this gap. It presents and theoretically grounds an agile four-phase model for curriculum development

aimed at integrating AI competence into higher education programmes. The model is based on the AI-Comp competence framework and is anchored in curriculum theory, agile development theory, and European higher education governance structures.

II. METHODOLOGICAL APPROACH

This article is a conceptual paper based on an integrative literature review. Its objective is not empirical validation but theoretical synthesis and model construction. The review aimed to identify process logics, design principles, and governance conditions relevant to embedding AI competences into existing higher education curricula. A structured literature search was conducted across major academic databases (Scopus, Web of Science, ERIC, Google Scholar) focusing on four thematic clusters: (1) Curriculum development in European higher education, (2) Agile and iterative approaches in educational design, (3) Curriculum innovation for emerging technologies and future skills, and (4) AI competence models in higher education. The search yielded 412 records in total; after removing duplicates and clearly irrelevant records, 236 sources were retained for title and abstract screening, of which 74 were assessed in full text and 24 were included in the final synthesis. Sources were selected based on relevance, theoretical contribution, and European applicability. The analysis followed a thematic clustering approach, identifying theoretical foundations that inform the development of the agile curriculum model. In addition to the integrative review, we conducted conceptual model construction. We coded and clustered the literature across the four domains to extract recurring process logics, design principles, and governance constraints, and then synthesised them into a Bologna-compatible four-phase framework. For each phase, we defined a core design artefact (e.g., competence profile, learning-objective matrix, curriculum map, sprint plan) and an explicit rationale linking outputs to subsequent phases. Coding focused in particular on (1) competence-related curricular goals, (2) iterative development mechanisms, (3) forms of alignment between teaching, assessment, and programme structures, and (4) institutional

and governance-related constraints. The search covered Scopus, Web of Science, ERIC, and Google Scholar. Search strings combined curriculum development, programme design, agile/iterative, design-based research, technology integration (TPACK/SAMR), AI competence/AI literacy, future skills, and higher education/Bologna [1]–[10]. Backward and forward citation tracking complemented database results. The search was designed to capture both conceptual and process-oriented contributions relevant to model development. Sources were included if they provided (1) process-oriented contributions relevant to programme-level curriculum development, (2) iterative/agile design or competence-based education perspectives, and (3) applicability to European governance contexts. As a theory-driven conceptual study, this paper does not provide empirical validation; empirical implementation studies are required to examine applicability, scalability, and impact.

III. THEORETICAL FOUNDATIONS

This article addresses the research question: How can higher education curricula be agilely developed to integrate AI competences and enable students' competence development in AI-permeated environments? To answer this question, we synthesise four complementary theoretical domains that—taken alone—remain incomplete, but together offer a coherent foundation for a programme-level, iterative, and accreditation-compatible approach: (1) curriculum theory under Bologna governance, (2) agile and iterative development theory in regulated systems, (3) educational design research and technology integration / instructional design models, and (4) competence-based education and AI competence frameworks. The synthesis below is deliberately written as a theory-to-model bridge: each subsection concludes with explicit implications that inform the conceptual model presented later in Section V.

A. Curriculum theory under Bologna governance: stability, modularity, and constrained innovation

Curriculum development in European higher education is structurally shaped by the Bologna architecture: modular programmes, ECTS-based workload logics, learning-outcome transparency, and quality assurance through cyclical accreditation. These design principles stabilise programme identity, comparability, and quality control—but they also institutionalise comparatively slow structural change, because programme-level redesign is typically “expensive” in governance terms (documentation, committee processes, stakeholder alignment, accreditation implications). Consequently, many innovations—especially in fast-moving domains—must be realised as embedded change within existing module frames, rather than as rapid whole-programme restructuring. In this context, outcome-based education (OBE) and constructive alignment provide a strong theoretical anchor and an established “internal quality logic.” Constructive alignment conceptualises curriculum design as a coherence chain from intended learning outcomes to teaching/learning activities and assessment evidence [11], [12]. Under AI conditions, this coherence function becomes even more critical because “AI

competence” is rarely a single skill; it is typically a compound capability that must be demonstrated through credible evidence in disciplinary contexts (e.g., analysing algorithmic bias in domain data, designing AI-supported workflows, or critically evaluating AI-generated claims) [13], [14]. Put differently: in Bologna-governed systems, the most feasible route to meaningful AI integration is often not the creation of standalone AI modules, but the systematic re-alignment of learning outcomes, learning activities, and assessment within existing modules [15]. European programme development is also frequently described as a staged process (e.g., environmental analysis → goal definition → curriculum design → evaluation and further development). In the German higher education discourse, Seidl and Vonhof [16] characterise programme development as a complex process that includes setting competence and learning goals, choosing teaching/learning forms, and planning evaluation—while explicitly asking how “agile” this process should be under dynamic environmental change. This staged logic matters for our research question because it clarifies where agility can be inserted without breaking governance: not by replacing Bologna logics, but by strengthening the evaluation/feedback function and by enabling earlier, smaller and evidence-driven adjustments within modules.

Implications for our model From Bologna-aligned curriculum theory we derive three requirements for an agile AI integration model:

- Innovation must be feasible below the accreditation threshold, primarily through embedded module-level redesign rather than programme disruption.
- AI competence integration must be designed as programme-level coherence, i.e., longitudinal progression across semesters and modules, not isolated interventions [15].
- Constructive alignment functions as the quality backbone that connects competence selection → learning outcomes → formats → assessment/evidence in a verifiable chain [17].

B. Agile and iterative development theory: governed iteration in regulated systems

Agile theory originates in software engineering [18], but it has become a broader paradigm for organising innovation under uncertainty: iterating in small increments, testing early, collecting stakeholder feedback, and adapting continuously. In public-sector and regulated contexts, contemporary research highlights that agility is possible when reframed as bounded experimentation with clear governance interfaces—rather than permanent disruption [8], [10]. This is particularly relevant for higher education programmes because study programmes can be conceptualised as complex “products” shaped by multiple stakeholder groups (students, teachers, administration, labour market actors, professional bodies, and society). Seidl and Vonhof [16] emphasise that agility should not be reduced to a late-stage method; rather, agile principles can function as drivers and design principles across all phases of programme development. The same text highlights the centrality

of structured feedback: from an agile perspective, placing evaluation only at the end is “too little,” because regular feedback is a prerequisite for quality-oriented development. A practically useful bridge between classic agile ideas and higher-education governance is the adapted set of principles used in the German discourse on agile programme development—drawing on “Forum Agile Verwaltung” and applied to study programme contexts [16]. In the empirical synthesis by Härer [19], agile principles relevant to programme development include: taking a holistic view; building cross-functional teams; experimenting with manageable changes and partial results; involving entitled stakeholders; obtaining regular internal and external feedback; and making the system increasingly appropriate. Härer [19] further emphasise that the “holistic view” in programme development emerges only when structural conditions, actor communication, and student competence development are considered together - a useful theoretical justification for treating curriculum innovation as both a pedagogical and organisational learning process.

Implications for our model: Agile theory, translated into Bologna-regulated higher education, implies that “agile curriculum development” should be conceptualised as a governed iteration system:

- reform proceeds through time-boxed cycles and manageable change units
 - pilots precede scaling, and learning is captured as evidence for next step decisions
 - stakeholder participation is structurally embedded (students/teachers/QA)
 - feedback loops are institutionalised, not optional add-ons
- This directly justifies an explicit implementation cycle (our later Phase 4) as a core design element rather than a “nice-to-have”.

C. Educational design research, rapid prototyping, and sprint logics as educational analogues

Within educational theory, the closest conceptual relatives to agile development are Educational Design Research / Design-Based Research (EDR/DBR) and rapid prototyping approaches. EDR conceptualises innovation as iterative cycles of design → enactment → evaluation → redesign, with the dual aim of improving practice and generating transferable design knowledge [20], [21]. This dual aim is particularly relevant for AI competence integration: the problem is not merely “adding content,” but responding to epistemic shifts in how knowledge is produced, validated, communicated, and ethically governed in disciplines affected by AI [22], [23]. In such contexts, linear planning models often struggle because the target practices (AI tools, organisational norms, assessment legitimacy, ethical requirements) are moving. The DBR/EDR logic also provides a strong theoretical precedent for time-boxed implementation windows combined with structured reflection—an idea mirrored in sprint-based course redesign and intensive iterative redesign models reported in higher-education innovation literature (e.g., sprint-inspired approaches to rapid course (re)design and formative evaluation).

Conceptually, these approaches converge on the claim that speed should not be achieved through simplification alone, but through a disciplined cycle of prototyping, data capture, and reflective revision—compatible with organisational learning theory’s notion of double-loop learning [24].

Implications for our model: EDR/DBR suggests that a programme-level agile model should:

- treat curriculum innovation as iterative design knowledge production, not only as implementation work
- include explicit evidence capture and reflection as design requirements;
- combine “concept formation and alignment” with a bounded enactment window and retrospective learning loop.

This strongly supports a four-phase logic in which Phases 1–3 produce a coherent design architecture, and Phase 4 operates as a time-boxed enactment–evaluation–redesign cycle.

D. Technology integration and instructional design frameworks: mediating content–pedagogy–technology and integration depth

AI competence integration requires theoretical mediators that explain how technology becomes educational practice in disciplinary teaching. Two widely used conceptual lenses are TPACK and SAMR. TPACK conceptualises successful technology integration as the intersection of technological knowledge, pedagogical knowledge, and content (disciplinary) knowledge—a crucial reminder that AI competence is not merely tool literacy but discipline-embedded practice. Recent empirical and review literature continues to use TPACK as a planning and reflective lens for contemporary technology integration and, increasingly, AI-related educational design [2], [25], [26]. For our research question, TPACK strengthens a central claim: AI competence must be distributed and discipline-embedded, not isolated in generic AI skills training. SAMR offers a second lens: it distinguishes between substitution and augmentation (often described as enhancement) and modification and redefinition (often framed as transformation). SAMR is frequently used as a reflective heuristic to discuss the depth of technology integration, but critical scholarship has also cautioned against simplistic “ladder” interpretations and advocated using it as a diagnostic reflection tool rather than a deterministic progress model [5], [6]. For AI competence integration, SAMR is valuable precisely because it forces a key question: does AI merely substitute existing practices (e.g., automated summarisation) or does it reshape epistemic tasks (e.g., critically auditing training data assumptions, evaluating bias and validity, designing AI-supported inquiry workflows)? A complementary operational anchor for rapid iteration is agile instructional design, particularly the Successive Approximation Model (SAM), which emphasises iterative prototyping and stakeholder feedback rather than linear perfection [27]. While SAM originates in e-learning design practice, its logic matches higher-education needs for rapid, testable learning design cycles and supports a structured implementation sprint with prototype artefacts, testing, and revision.

Implications for our model: From technology integration and instructional design frameworks we derive that:

- format selection must be framed as discipline–pedagogy–technology integration, not tool adoption;
- curricula must decide on integration depth (enhancement vs. transformation), ideally linked to evidence expectations and assessment legitimacy;
- sprint-based prototyping is theoretically defensible as an instructional design strategy under uncertainty (supporting Phase 4), while Phases 2–3 can be strengthened by explicitly treating “format choice” and “curriculum mapping” as decisions about integration depth and epistemic task redesign.

E. Competence-based education and AI competence frameworks: normative orientation for agile development

Competence-based education conceptualises competence as a context-dependent disposition for effective action rather than a list of isolated skills [28]. Contemporary competence literature and policy frameworks emphasise competence orientation as a response to complexity and accelerating change [7], [9], [29]–[32]. This matters for agile curriculum development because agility without a normative competence orientation risks becoming process-centric: iterating quickly without clarity about what kind of graduate capability should emerge. AI competence frameworks in higher education have increasingly moved beyond purely technical literacy toward integrative models that include ethical judgement, critical reflection, human–AI collaboration, and societal responsibility [23], [33]. AIComp plays precisely this role: it provides an empirically grounded competence architecture that can be mapped to learning objectives, assessments, and programme structures. The theoretical point is general: a structured competence model is the baseline artefact that enables curriculum gap analysis, programme mapping, and constructive alignment. Without it, “agile integration” lacks direction and evaluation criteria [14], [34]–[36].

Implications for our model: Competence theory contributes three essential functions:

- Normative orientation (what competences matter, and why);
- Mapping capability (gap analysis and distribution across the curriculum);
- Alignment bridging (competences → outcomes → evidence).

Thus, our later model must combine agility (process logic) with competence modelling (substance logic) and constructive alignment (quality logic).

F. Integrative synthesis: towards an agile curriculum architecture for AI competence integration

Across the reviewed domains, no single framework sufficiently answers the research question. Bologna curriculum theory explains the structural constraints and the primacy of alignment; agile theory explains how change can be organised

under uncertainty; EDR/DBR explains why iterative evidence-based redesign is epistemically appropriate in educational innovation; technology integration models explain how AI should become discipline-embedded practice; competence-based education explains what must be developed and why. Taken together, the synthesis yields five meta-principles that inform the conceptual model presented in Section V:

- 1) Embedded integration (no add-on logic). AI competence should be distributed across modules and semesters, integrated into disciplinary learning tasks (TPACK/SAMR), rather than confined to standalone “AI modules.”
- 2) Alignment coherence as quality backbone. Competences must translate into assessable learning outcomes, learning activities, and credible evidence (constructive alignment).
- 3) Bounded experimentation within governance. Agility in Bologna contexts requires time-boxed pilots and manageable change units, not continuous structural disruption.
- 4) Stakeholder participation as a design requirement. Students, teachers, QA and governance actors must be structurally involved; feedback loops are not optional.
- 5) Evidence-based scaling and institutional learning. Iterations must produce evidence that enables decisions (scale / stay / park) and supports institutional learning.

This integrated theoretical foundation positions the four-phase model in Section V as a layered innovation logic: it does not replace Bologna governance; it enables fast learning cycles within slow learning regimes by combining competence orientation, constructive alignment, and governed iteration.

IV. AICOMP FRAMEWORK

The AIComp Framework (AI Competences) is a multidimensional competence model developed by the NextEducation Research Group at DHBW Karlsruhe. It is grounded in a large-scale empirical study with more than 1,600 professionals in Southwest Germany (Baden–Württemberg), conducted within the initiatives AI-Campus and AI-Campus Hub Baden–Württemberg. The study examined the current state of AI-related competences in professional contexts and identified those capabilities that require systematic development in response to accelerating technological transformation [1], [14]. Building on this empirical basis, AIComp formulates twelve “future skills” considered essential for life and work in environments increasingly shaped by artificial intelligence. These competences extend beyond technical literacy and include ethical reflection, critical judgement, creativity, and the ability to actively shape AI-driven change processes. AI competence is therefore understood not merely as tool use, but as responsible, reflective, and innovative engagement with AI systems in professional and societal contexts.

Conceptually, AIComp is embedded in Ehlers’s Future Skills framework, which defines future skills as “competences enabling individuals to act successfully in highly dynamic and complex problem situations” [13]. Accordingly, AIComp represents a specialised subdomain of this broader model, focusing on the conditions of agency in AI-permeated environments. Competence is conceived as the interplay of

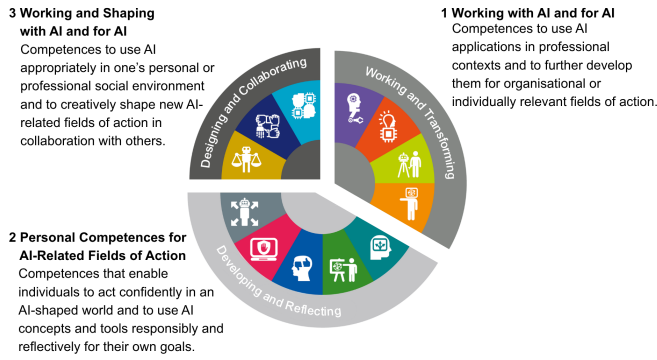


Fig. 1. Competence Structure Model AIComp in three dimensions [37]

knowledge, skills, attitudes, and motivation; it reflects not only what individuals know, but their willingness and ability to act responsibly in technology-mediated contexts.

To operationalize this understanding, AIComp is designed as a coherent competence structure model (fig. 1). At the highest level, the model distinguishes three dimensions of AI-related capability. The first concerns the development of the acting subject and includes reflective AI literacy, ethical sensitivity, and self-regulatory capacities that enable individuals to position themselves responsibly in AI-shaped contexts. The second dimension addresses working and designing with and for AI; it encompasses innovation competence, problem-solving abilities, and the capacity to meaningfully integrate AI into workflows and organisational processes. The third dimension focuses on shaping social and organisational environments with and through AI, including cooperation, communication, and collaborative creation of new AI-supported fields of activity [14], [34], [36], [38], [39].

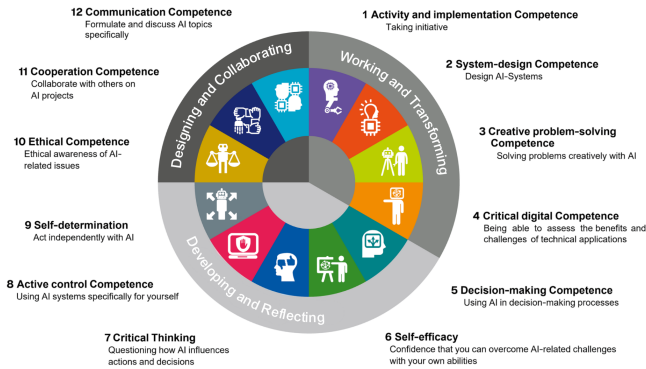


Fig. 2. Competence Structure Model AIComp with 12 future skills [14]

These three dimensions are conceived as interdependent rather than isolated. Effective AI competence emerges when individuals are able to combine self-reflective orientation, task-specific design capabilities, and socially responsible interaction in concrete situations of action. The twelve competence fields (fig. 2) specify these dimensions as thematic dispositions for action, while the thirty-six competence items function as

micro-level descriptors that translate them into teachable and assessable learning outcomes [14]. This layered architecture enables systematic competence mapping, gap analysis, and structured development pathways. The relevance of AIComp for higher education follows directly from this structural clarity. Universities face increasing pressure to integrate AI-related competences in response to technological acceleration, regulatory developments, and labour market transformation. However, agile development approaches primarily address how change processes can be organised—through iteration, feedback, and stakeholder participation—rather than what competences should be developed. A clearly articulated competence model is therefore a necessary precondition for meaningful curriculum reform [34]. AIComp provides a structured and empirically grounded competence architecture that can be mapped onto modules, ECTS allocations, and programme profiles. In addition, each competence field is operationalised through learning objectives differentiated according to Bloom's taxonomy and structured across three progressive levels of cognitive complexity. This three-tiered specification enables the translation of abstract competence descriptions into measurable learning outcomes, ranging from foundational understanding to advanced application and reflective or creative mastery [34]. As a result, the framework not only supports strategic competence mapping but also facilitates constructive alignment between intended learning outcomes, teaching–learning activities, and assessment formats. It thus enables higher education institutions to conduct systematic competence mapping and gap analyses before initiating iterative reform processes. In this sense, AIComp constitutes the analytical baseline for the agile curriculum development model proposed in this paper. It defines the competence landscape and the intended target state of learning that must guide any systematic reform process. The following chapter therefore builds directly on this foundation and presents, in a structured and stepwise manner, the agile programme development model through which AI competences can be integrated into higher education curricula and teaching practice. While AIComp specifies the substantive orientation and competence goals, the subsequent model explains the procedural logic of how these competences can be embedded iteratively and sustainably within study programmes [34].

V. THE AGILE CURRICULUM DEVELOPMENT MODEL (ADM-AI)

The conceptual process model consists of four interconnected phases (fig. 3). Together, they describe how AI-related competences can be systematically integrated into higher education programmes through an iterative yet structurally coherent approach. The model follows a clear logic: it begins with clarifying what competences are relevant for a specific programme, proceeds to defining how these competences should be operationalised pedagogically, translates them into curricular structures, and finally moves into iterative implementation and refinement [34].



Fig. 3. ADM-AI Framework for agile curriculum development in 4 phases [34]

The first Phase focuses on clarifying AI competence needs. It establishes the epistemic and strategic foundation of curriculum development by analysing how artificial intelligence transforms the discipline, identifying relevant competence fields, prioritising them, and translating them into a programme-specific competence profile. Phase 2 translates these prioritised competences into clearly formulated learning objectives and appropriate teaching formats. It defines the intended level of competence development and aligns pedagogical strategies and assessment approaches with the previously defined competence architecture. Phase 3 embeds the defined learning objectives and formats into the existing curriculum. Rather than creating isolated add-on modules, this phase emphasises embedded design: AI competences are distributed across semesters, modules, and assessment formats in a structurally coherent and accreditation-compatible manner. Phase 4 operationalises the reform through a structured 90-day implementation cycle. Selected elements are piloted in concrete teaching settings, observed, reflected upon, and iteratively refined. This phase ensures that curriculum innovation remains adaptive, experience-based, and institutionally anchored [34].

A. Phase 1: Clarifying AI Competence Needs

The first phase of the agile curriculum development model establishes the conceptual and strategic foundation upon which all subsequent reform processes rest. Before learning objectives are defined or curricular structures are redesigned, it is essential to clarify which artificial intelligence (AI)-related competences are relevant to the specific study programme and why they matter. Phase 1 functions as a structured competence clarification process that generates a shared understanding of how AI transforms the discipline, identifies the emerging or modified competences resulting from this transformation, and determines which of these competences merit prioritisation for curricular development. From a curriculum-theoretical perspective, this corresponds to the initial analytical stage of curriculum design in which educational aims and contextual conditions are examined prior to instructional decision-making [40], [41]. In parallel, organisational learning theory emphasises that shared problem framing constitutes a prerequisite

for coordinated collective action [24]. Phase 1 operationalises these insights within a systematic and participatory design framework. AIComp provides the overarching competence architecture for AI-related curriculum development, while Phase 1 of the ADM-AI model identifies and prioritises those competences most relevant for a specific study programme in light of students' current competence profile and anticipated occupational requirements. Although some competence fields, such as critical digital competence, ethical competence, communication competence, and critical thinking, are broadly relevant across disciplines, their weighting and operationalisation vary by disciplinary and professional orientation. For example, technically oriented programmes may place stronger emphasis on the responsible application of AI tools in diagnostics, decision-making, and automation [34].

1) *Step 1: Scoping and Clarifying the Initial Situation:* The process begins with a clarification of the study programme's disciplinary profile and its transformation context. The development team collectively reflects upon three interdependent dimensions: the disciplinary identity of the programme, its societal and professional fields of application, and the manner in which AI is reshaping the discipline's research methodologies and modes of knowledge production. This analytical activity is not limited to description but involves critical reflection on the epistemic evolution within the field. The guiding questions that steer this step include: How does artificial intelligence transform research, methodology, and knowledge production within our discipline, and which scientific competences consequently gain new relevance? And which emerging professional responsibilities and ethical challenges arise, and how should the competence profile adapt in response? These reflective prompts serve as structured scaffolding mechanisms that enable systematic inquiry and guard against superficial reform. The outcome of this step is a documented scoping statement that defines the programme's AI-related development horizon and clarifies structural parameters, such as accreditation requirements, institutional priorities, and resource constraints [34].

2) *Step 2: Collecting and Clustering Relevant Competencies:* Once the disciplinary transformation context has been delineated, relevant AI-related competences are identified and preliminarily conceptualised. These may be drawn from existing competence frameworks, such as AIComp, OECD models, or discipline-specific standards, but they are critically examined for contextual validity rather than adopted wholesale [7], [9], [31]. The clustering of competences serves two core purposes: it mitigates conceptual complexity by grouping related ideas into coherent thematic clusters, and it fosters semantic alignment within the development team. Differing interpretations of competence constructs like "data literacy", "ethical AI judgement" or "human-AI collaboration" are reconciled through collaborative negotiation. This analytical process reflects the principles of competence modelling and collaborative knowledge construction, which view competences as context-dependent dispositions for action rather than isolated skills [28], [42]. Two guiding questions further

refine the analysis: Which AI-related competences visibly transform research, teaching, or professional practice within our field, and which are currently underrepresented in the existing curriculum? Which elements of established competence frameworks are transferable to our context, and where must they be reinterpreted in light of the discipline's internal logic? The outcome of this step is a structured competence landscape that provides a reference framework for subsequent prioritisation [34].

3) *Step 3: Positioning and Prioritising Competences:* Not all competences identified can or should be pursued simultaneously. Strategic prioritisation is therefore essential. Competences are evaluated along two analytical dimensions: their disciplinary relevance and the current competence level of students. This positioning reveals competence areas of highest urgency—those that are both highly relevant and insufficiently developed among learners. The process resembles a targeted gap analysis that combines normative judgement (defining what is most essential to the discipline's future trajectory) with empirical diagnosis (assessing the current competence state of students). The result is a focused subset of key competences—typically three to five—that serve as the focal points for further curricular development. This prioritisation process enhances transparency, coherence, and institutional legitimacy by documenting and justifying decisions within a shared framework of understanding [34].

4) *Step 4: Subject-Specific Contextualisation and Documentation:* In the final step, the prioritised competences are translated into the conceptual and linguistic framework of the specific study programme. Abstract competence formulations are reformulated into discipline-specific outcomes that directly align with the programme's theoretical foundations, methodological approaches, and existing module structure. For instance, a generic competence such as “ethical AI use” might be recontextualised in a psychology curriculum as “critically evaluating algorithmic assessment tools in personnel selection with regard to fairness and bias.” Through this translation, competences attain pedagogical precision and become compatible with accreditation frameworks. This step is grounded in the principle of constructive alignment, according to which learning outcomes should be formulated in ways that are observable, discipline-specific, and assessable [12]. It also reinforces the competence-theoretical insight that competences derive their meaning only when situated within concrete domains of professional and academic action [28]. The outcome of Phase 1 is a clearly articulated, prioritised, and discipline-specific AI competence profile that serves as the analytical foundation for Phase 2, which focuses on the formulation of learning objectives and the design of appropriate instructional formats [34].

B. Phase2: Determining Learning Objectives and Teaching Formats

Building upon the prioritised and discipline-specific AI competence profile established in Phase 1, the second phase transforms these competences into pedagogically operational

forms. Whereas Phase 1 defined what AI-related competences are strategically relevant, Phase 2 specifies how these competences are to be developed within the curriculum. It defines the intended levels of cognitive mastery and determines the instructional arrangements through which students can acquire, apply, and demonstrate these competences in verifiable ways. From a curriculum-theoretical standpoint, this phase corresponds to the articulation of learning outcomes and their alignment with teaching and assessment strategies [11], [12]. It operationalises the principle of constructive alignment by ensuring that competence aims, learning activities, and evaluative procedures remain coherent across all curricular elements. Consequently, the outcome of Phase 2 is not a loose collection of teaching ideas but a systematically derived architecture of learning objectives and corresponding instructional formats forming the binding foundation for curricular embedding in Phase 3 [34].

1) *Step 1: Deriving and Documenting Learning Objectives:* The first step in Phase 2 converts the prioritised AI competences into precise, observable learning objectives. While competences denote enduring dispositions for effective action in complex contexts, learning objectives articulate specific performances that students should be capable of demonstrating at particular stages of the programme. This process requires two complementary forms of refinement: behavioural specification and cognitive differentiation. Learning objectives must state observable learner actions such as analysing algorithmic bias, designing AI-supported interventions, or critically evaluating the epistemic validity of AI-generated results. Such specification aligns with the outcome-based education paradigm, which emphasises clarity, measurability, and transparency across learning design stages. Cognitive levels of performance must likewise be differentiated to capture progression in student understanding. Taxonomies such as the revised Bloom framework and the SOLO taxonomy provide hierarchical structures to describe these progressive stages, ranging from foundational comprehension to complex relational and extended abstract thinking [43], [44]. Within AI competence development, such granularity is particularly salient: students may first explain algorithmic concepts, then apply them within disciplinary methodologies, and ultimately evaluate or design innovative AI-driven solutions [34]. The articulation of learning objectives therefore clarifies three dimensions:

- a) Which specific learning actions students must demonstrate.
- b) At which cognitive level such mastery is expected.
- c) How competence attainment will be evidenced and evaluated.

The outcome of this step is a structured, level-differentiated matrix of learning objectives that functions as an intermediary artefact linking the conceptual competence architecture to practical teaching and assessment design. As an internal design tool, it supports semantic alignment across modules and prevents fragmentation in later curriculum implementation. Guiding questions structure this process of derivation:

How can the selected competences be translated into observable learning actions? At which cognitive level should each competence be achieved within the programme? Which evidence forms best demonstrate attainment? Addressing these questions introduces assessment considerations at an early stage, thereby reinforcing constructive alignment. Anticipated artefacts of learning, like mentioned in the "format palette" [34] may include reflective essays, AI project portfolios, laboratory notebooks, design prototypes, policy briefs, or analytical reports. By engaging assessment design concurrently with learning objective specification, Phase 2 safeguards internal coherence between intended outcomes and evaluation practices [34].

2) *Step 2: Developing and Interlinking Teaching Formats:* Once learning objectives have been articulated, the second step focuses on their pedagogical realisation. The central question shifts from what should be learned to how learning environments can most effectively enable learning progression. The development team systematically identifies and designs teaching formats that render AI competences experientially accessible and pedagogically meaningful. This may involve the adaptation of existing courses, the integration of AI-relevant components into discipline-specific modules, or the creation of novel interdisciplinary and practice-oriented learning arrangements. This format palette provides an orienting framework that illustrates potential configurations without prescribing procedural detail [34]. This diversification of teaching formats reflects constructivist and experiential learning principles, which view learning as an active, context-embedded process of constructing meaning through authentic engagement [45]. Particularly in AI education, knowledge must be dynamically enacted rather than passively received; students advance by experimenting, designing, and critically reflecting within realistic disciplinary scenarios. Simultaneously, innovation must remain compatible with institutional frameworks and regulatory requirements. Accreditation standards, resource conditions, and qualification frameworks need to be integrated into the design process to ensure both compliance and sustainability. Consequently, creative curricular innovation and structural feasibility must advance in tandem. Guiding questions for this step include: How can students meaningfully experience, apply, and demonstrate the targeted competences? Which existing formats can be adapted, and where is genuine innovation necessary? How can quality assurance, transparency, and accreditation compliance be maintained while supporting pedagogical diversity? The result of Step 2 is a coherent format portfolio that is explicitly aligned with the defined learning objectives and designed to enable measurable competence acquisition. This portfolio constitutes the pedagogical blueprint for subsequent curriculum integration in Phase 3 [34].

C. Phase 3: Building the Curriculum

Phase 3 extends the focus from individual modules to the study programme as an integrated whole. Building on the clarified AI competence profile (Phase 1) and the derived learning objectives and teaching formats (Phase 2), this phase

situates these elements within the overarching curriculum architecture of the programme. The central concern shifts from isolated instructional units to the systematic design of a coherent curriculum in which AI competences are structurally and visibly anchored. The aim of this phase is to establish an embedded curriculum in which AI-related competences are progressively developed and reliably assessed, without merely appending stand-alone AI modules or parallel tracks. Rather than constructing separate AI pathways, existing courses are purposefully reconfigured so that AI competences become an organic component of the discipline's epistemic and professional structure. This approach avoids curricular fragmentation and promotes integration at the level of programme identity and graduate profile. Realising this aim requires a comprehensive examination of the entire study programme across semesters, modules, and assessment regimes. The key design question concerns the longitudinal distribution of AI competences: where they are introduced, where they are consolidated and deepened, and where they culminate in advanced application or critical reflection. Such sequencing supports a structured progression of learning and aligns with the idea of programme-level constructive alignment, in which course-level activities and assessments collectively serve the programme's overarching outcomes. A central instrument in Phase 3 is curriculum mapping. The prioritised AI competences are systematically assigned to specific modules, and the teaching formats developed in Phase 2 are embedded accordingly. Simultaneously, appropriate forms of evidence—such as portfolios, project reports, digital badges, or transcript annotations—are defined to render competence development visible and documentable over time. Through this mapping process, programmes obtain a transparent overview of (i) which competences are addressed in which modules, (ii) through which pedagogical formats they are developed, and (iii) by which assessment artefacts achievement is demonstrated. This systemic overview enables the identification of redundancies, gaps, and progression patterns within the curriculum. It clarifies where competences are over-represented, insufficiently addressed, or insufficiently advanced across the study trajectory. At the governance level, curriculum mapping also helps distinguish between adjustments that can be implemented within existing accreditation and regulatory frameworks and those that necessitate more substantial structural modification. In this sense, Phase 3 explicitly mediates between pedagogical innovation and institutional feasibility. Furthermore, integration can extend beyond the boundaries of a single programme. Foundational AI modules, interdisciplinary ethics offerings, or shared competence formats can be designed in polyvalent ways and utilised across multiple study programmes. Such cross-programme structures increase resource efficiency, foster interdisciplinary encounters, and support the development of shared institutional AI literacy. The outcome of Phase 3 is an integrated curriculum blueprint that structurally embeds AI competences within the programme architecture. It ensures coherence between competence goals, learning objectives, pedagogical formats, assessment practices, and regulatory re-

quirements. And Phase 3 marks a transitional stage rather than a final endpoint: structural embedding alone does not guarantee sustainable transformation. Phase 4 therefore focuses on iterative implementation, feedback-driven refinement, and institutional learning processes that stabilise and further develop the newly introduced AI curriculum structures [34], [38].

D. Phase 4: The 90-Day Implementation Cycle

Phase 4 translates the structurally embedded AI competences of Phase 3 into concrete pedagogical practice within a defined temporal framework. Instead of attempting comprehensive reform in a single step, this phase introduces an iterative, time-bound implementation cycle—typically 90 days—that enables controlled experimentation, systematic observation, and evidence-informed refinement. In doing so, it links curriculum design with organisational learning and positions implementation as a deliberate, research-informed process rather than a one-off rollout. At the outset, the development team selects a strategically chosen pilot context in which the newly formulated AI learning objectives and teaching formats can be tested under real conditions. This selection is not incidental: the chosen module or learning setting should combine disciplinary relevance with sufficient flexibility to accommodate innovation, while remaining manageable in scope. By concentrating efforts on a clearly delineated context, the reform process becomes tangible, bounded, and actionable, thereby lowering the threshold for experimentation and fostering ownership among teaching staff. Within this pilot setting, the team specifies which subset of learning objectives, materials, and methods will be implemented during the 90-day period. The focus lies on a coherent, realistic implementation package rather than on exhaustive coverage of all planned innovations. Responsibilities are explicitly allocated: it is defined who prepares content and learning resources, who conducts teaching activities, who documents classroom processes and student artefacts, and who facilitates subsequent reflection. This clarified role distribution enhances transparency, reduces ambiguity during the cycle, and anchors the implementation process in shared, rather than individual, responsibility. Before the cycle begins, practical and organisational prerequisites are systematically addressed. This includes clarifying technical infrastructure and access to AI tools, aligning assessment procedures with the pilot’s learning objectives, informing students about aims and expectations, and, where necessary, coordinating with programme leadership or examination boards. Such anticipatory work reduces the risk of avoidable disruptions and underscores that agile development does not imply improvisation, but rather carefully framed experimentation within clear constraints. The core of the 90-day cycle is the structured collection and interpretation of experience. Teaching is accompanied by targeted observation, student feedback, and systematic documentation of learning artefacts and process notes by teaching staff. Short reflection intervals may be integrated during the cycle, but a central retrospective at its conclusion is essential. In this retrospective, the team examines

not only which elements “worked” or “did not work,” but also why this was the case and which underlying assumptions about students, content, or formats may need revision. This aligns with the notion of double-loop learning, in which both practices and their governing premises are subjected to critical scrutiny [24], [34]. The goal of Phase 4 is thus not the flawless implementation of a final design, but the generation of robust experiential knowledge that feeds back into the curriculum. Successful elements can be stabilised, scaled, or extended to additional modules; less effective approaches can be modified, delimited, or discontinued. Through successive 90-day cycles, the curriculum evolves in an adaptive, evidence-informed manner, and reform becomes a continuous institutional practice rather than a singular project.

VI. THE SIX AGILE PRINCIPLES

Building on the agile and iterative development theory outlined in Section III.B and its adaptation to higher education programme development, the ADM-AI model operationalises agility through six curriculum-specific principles. These principles translate general agile values—iteration, stakeholder orientation, and continuous feedback—into governance-compatible design rules for programme innovation within Bologna structures. They are aligned with the practical translation of agile thinking into study programme development as reflected in the playbook from Ehlers and Mann as well as in related higher-education discourse [34].

- 1) Integrative (Embedded Design): AI and future competences are systematically embedded into existing modules, learning objectives, and assessment formats rather than added as isolated units. This principle reflects the agile commitment to incremental improvement within existing structures and aligns with constructive alignment logic: competence development is integrated into the overall curricular architecture rather than treated as an add-on.
- 2) Participatory (Co-Creation Across Phases): Stakeholders—students, faculty, administrators, and where relevant external partners—are involved from conceptualisation through implementation and evaluation. This principle derives directly from agile theory’s emphasis on cross-functional collaboration and continuous stakeholder feedback [18], [46]. In regulated systems, participation increases both legitimacy and implementation feasibility.
- 3) Iterative (Sprint-Based Development Cycles): Curriculum innovation unfolds in bounded cycles (e.g., 90-day pilots), each ending with structured reflection and decision-making. Iteration operationalises double-loop learning [24], [34] at programme level: not only practices but also assumptions can be reconsidered.
- 4) Experimental (Small, Testable Adjustments): Innovation proceeds through manageable changes within existing module frameworks, enabling experimentation without triggering full re-accreditation. This principle reflects agile governance approaches in regulated environments

[8], [10] and corresponds to “bounded experimentation” rather than disruptive overhaul.

- 5) Visible (Evidence and Competence Transparency): Competence acquisition is made explicit and documentable through artefacts, badges, microcredentials, or transcript entries. Agile processes require observable outcomes; iterations must generate evidence that supports informed decisions (scale / stay / park) and organisational learning.
- 6) Holistic (Systemic Alignment Across Levels): Curriculum development is understood as a systemic process that integrates disciplinary logic, institutional strategy, accreditation constraints, and societal demands. This principle extends agile thinking from project-level optimisation to system-level coherence, ensuring that innovation remains strategically aligned and future-orientated.

VII. DISCUSSION

DISCUSSION The proposed model demonstrates that agility and regulatory stability need not be opposing forces in European higher education. By conceptualising curriculum development as governed iteration rather than structural disruption, the framework reconciles innovation dynamics with Bologna-based accreditation regimes. Instead of positioning AI competence integration as an additional curricular layer, the model reframes it as alignment work: competences are translated into learning outcomes, embedded into modules, and rendered visible through assessable artefacts.

The discussion highlights three central contributions. First, the model shifts attention from isolated AI modules toward programme-level coherence and longitudinal competence progression. Second, it integrates agile development principles into curriculum governance without undermining institutional accountability. Third, it combines competence modelling, constructive alignment, and iterative implementation cycles into a unified architecture.

At the same time, the framework presupposes institutional preconditions. Agile curriculum development requires a culture of experimentation, leadership support, and willingness to generate and interpret evidence from pilot implementations. Without institutional commitment to feedback loops and retrospective learning, the 90-day cycles risk becoming symbolic exercises rather than mechanisms of organisational learning. Moreover, the model’s applicability may vary across disciplines with different epistemic traditions and regulatory strictures.

At this stage, the framework should be understood as a conceptual foundation for subsequent implementation and evaluation rather than as an empirically validated intervention model. A logical next step is its application in real curriculum development processes by study programme leaders, curriculum developers, and institutional actors responsible for programme design. Such implementations could form the basis for user-orientated studies examining how the model is understood, operationalised, and perceived in practice; which elements prove most usable in different disciplinary contexts; and where adaptations are required. At the same time, the

framework builds on prior research on agile study programme development in higher education. Härer’s work shows that agility in study programme development has already been conceptually examined and empirically explored in German higher education contexts, providing an important reference point for further implementation-orientated work. The present framework is also informed by the authors’ own experience with curricular transformation processes, which helped to sharpen its practical orientation and institutional relevance. In this sense, the model is intended as a starting point for iterative institutional testing and refinement [19].

Future research should therefore empirically examine how the model performs across institutional types and disciplinary contexts, how stakeholder participation influences adoption, and which forms of evidence most effectively support scaling decisions. This is particularly important because the model has been derived from integrative synthesis rather than evaluated through multi-site implementation. Furthermore, the feasibility of “below-threshold” change may differ across disciplines and national accreditation practices, requiring contextual tailoring and governance negotiation in real settings.

Future research should therefore empirically examine how the model performs across institutional types and disciplinary contexts, how stakeholder participation influences adoption, and which forms of evidence most effectively support scaling decisions. This is particularly important because the model has been derived from integrative synthesis rather than evaluated through multi-site implementation. Furthermore, the feasibility of “below-threshold” change may differ across disciplines and national accreditation practices, requiring contextual tailoring and governance negotiation in real settings.

VIII. CONCLUSION

This paper has proposed a theoretically grounded and governance-compatible framework for agile curriculum development aimed at integrating AI competences into European higher education programmes. By synthesising curriculum theory, agile governance, educational design research, technology integration models, and competence-based education, the four-phase model offers a structured pathway from competence clarification to iterative implementation. Rather than replacing existing accreditation structures, the framework enables faster learning cycles within slow structural regimes. In doing so, it provides higher education institutions with a practicable strategy for embedding AI competences systematically, coherently, and sustainably in study programmes shaped by regulatory stability and accelerating technological change.

IX. ACKNOWLEDGEMENT

This project was funded by the Ministry of Science, Research and Arts Baden-Württemberg and conducted by the NextEducation Research Group (www.next-education.org) at Baden-Württemberg Cooperative State University (DHBW) Karlsruhe. The artefacts are Open Educational Resources (OER) and can be found under the following websites: ai-comp.org and ai-comp.org/agil.

X. APPENDIX

The final synthesis was based on these sources selected through a structured integrative review. The corpus was defined by conceptual saturation and theoretical relevance: once the four thematic clusters no longer produced substantively new categories for process logics, design principles, alignment structures, and governance conditions, additional screening was discontinued. The resulting set of sources provided sufficient breadth across curriculum theory, iterative/agile development, AI competence frameworks, and European higher education governance to support theory-driven model construction.

[1] K. Beck et al., 'Manifesto for Agile Software Development', *Zwölf Prinzipien Agiler Softwareentwicklung*. Accessed: Mar. 19, 2026. [Online]. Available: <https://agilemanifesto.org/>

[2] J. Biggs, 'Enhancing teaching through constructive alignment', *High Educ*, vol. 32, no. 3, pp. 347–364, Oct. 1996, doi: 10.1007/BF00138871.

[3] J. B. Biggs and C. S. Tang, *Teaching for quality learning at university: what the student does*, 4th edition. in SRHE and Open University Press Imprint. Maidenhead: McGraw-Hill/Society for Research into Higher Education/Open University Press, 2011.

[4] M. Černý, 'AI Literacy in Higher Education: Theory and Design — [springerprofessional.de](https://www.springerprofessional.de)'. Accessed: Mar. 30, 2026. [Online]. Available: <https://link.springer.com/chapter/10.1007/978-3-031-63235-8-24>

[5] J. Delcker, J. Heil, D. Ifenthaler, S. Seufert, and L. Spigri, 'First-year students AI-competence as a predictor for intended and de facto use of AI-tools for supporting learning processes in higher education', *Int J Educ Technol High Educ*, vol. 21, no. 1, p. 18, Mar. 2024, doi: 10.1186/s41239-024-00452-7.

[6] European Association for Quality Assurance in Higher Education (ENQA), European Students' Union (ESU), European University Association (EUA), and European Association of Institutions in Higher Education (EURASHE), 'Standards and Guidelines for Quality Assurance in the European Higher Education Area (ESG)'. Standards and Guidelines for Quality Assurance in the European Higher Education Area, Brussels, Belgium, 2015. doi: 978-9-08-168672-3.

[7] European Education and Culture Executive Agency, Ed., *The European higher education area in 2024: Bologna process implementation report*. Luxembourg: Publications Office, 2024. doi: 10.2797/351309.

[8] European Ministers of Education, 'European Higher Education Area and Bologna Process.', *The Bologna Declaration of 19 June 1999*. Accessed: Mar. 20, 2026. [Online]. Available: <https://ehea.info/page-ministerial-conference-bologna-1999>

[9] Flair Collaboration, 'AI competence frameworks and policies in higher education: Synthesis report', *Vienna University of Economics and Business*, 2025. [Online]. Available: <https://www.wu.ac.at/en/flair/synthesis-report>

[10] E. R. Hamilton, J. M. Rosenberg, and M. Akcaoglu, 'The Substitution Augmentation Modification Redefinition

(SAMR) Model: a Critical Review and Suggestions for its Use', *TechTrends*, vol. 60, no. 5, pp. 433–441, Sep. 2016, doi: 10.1007/s11528-016-0091-y.

[11] F. Härer, G. Herzwurm, and R. Schmidt, 'Agilität in der Studiengangsentwicklung an deutschen Hochschulorganisationen – Eine Bestandsaufnahme', *Gr Interakt Org*, vol. 56, no. 1, pp. 185–195, Mar. 2025, doi: 10.1007/s11612-025-00799-z.

[12] M. Janssen and H. van der Voort, 'Agile and adaptive governance in crisis response: Lessons from the COVID-19 pandemic', *International Journal of Information Management*, vol. 55, p. 102180, Dec. 2020, doi: 10.1016/j.ijinfomgt.2020.102180.

[13] M. Koehler and P. Misha, 'What Happens When Teachers Design Educational Technology? The Development of Technological Pedagogical Content Knowledge', *Journal of Educational Computing Research - J EDUC COMPUT RES*, vol. 32, pp. 131–152, Jul. 2005, doi: 10.2190/0EW7-01WB-BKHL-QDYV.

[14] D. Long and B. Magerko, 'What is AI Literacy? Competencies and Design Considerations', in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, in CHI '20. New York, NY, USA: Association for Computing Machinery, Apr. 2020, pp. 1–16. doi: 10.1145/3313831.3376727.

[15] C. Loughlin, S. Lygo-Baker, and Å. Lindberg-Sand, 'Reclaiming constructive alignment', *European Journal of Higher Education*, vol. 11, no. 2, pp. 119–136, Apr. 2021, doi: 10.1080/21568235.2020.1816197.

[16] S. McKenney and T. Reeves, *Conducting Educational Design Research*, 2nd Edition. 2018.

[17] I. Mergel, S. Ganapati, and A. B. Whitford, 'Agile: A New Way of Governing', *Public Administration Review*, vol. 81, no. 1, pp. 161–165, Jan. 2021, doi: 10.1111/puar.13202.

[18] P. Mishra and M. J. Koehler, 'Technological Pedagogical Content Knowledge: A Framework for Teacher Knowledge', *Teachers College Record*, vol. 108, no. 6, pp. 1017–1054, Jun. 2006, doi: 10.1111/j.1467-9620.2006.00684.x.

[19] OECD, 'The Future of Education and Skills: Education 2030', *OECD Education Policy Perspectives 98*, Jun. 2018. doi: 10.1787/54ac7020-en.

[20] G. Ruge, O. Tokede, and L. Tivendale, 'Implementing constructive alignment in higher education - Cross-institutional perspectives from Australia', *Higher Education Research and Development*, p. 16, Nov. 2018, doi: 10.1080/07294360.2019.1586842.

[21] D. A. Schmidt, E. Baran, A. D. Thompson, P. Mishra, M. J. Koehler, and T. S. Shin, 'Technological Pedagogical Content Knowledge (TPACK): The Development and Validation of an Assessment Instrument for Pre-service Teachers', *Journal of Research on Technology in Education*, vol. 42, no. 2, pp. 123–149, Dec. 2009, doi: 10.1080/15391523.2009.10782544.

[22] P. D. T. Seidl and C. Vohnhof, 'Agile Prinzipien – Was kann die Studiengangsentwicklung davon lernen?', no. 3, *Synergie. Fachmagazin für Digitalisierung und Lehre*,

Jul. 21, 2017. Accessed: Mar. 20, 2026. [Online]. Available: <https://hochschulforumdigitalisierung.de/agile-prinzipien-was-kann-die-studiengangsentwicklung-davon-lernen/>

[23] F. E. Weinert, Concept of competence: a conceptual clarification.. Seattle: Hogrefe & Huber. in D. S. Rychen & L. H. Salganik (eds) Defining and selecting key competencies. Seattle, 2001.

[24] O. Zawacki-Richter, V. I. Marín, M. Bond, and F. Gouverneur, ‘Systematic review of research on artificial intelligence applications in higher education – where are the educators?’, *Int J Educ Technol High Educ*, vol. 16, no. 1, p. 39, Dec. 2019, doi: 10.1186/s41239-019-0171-0.

REFERENCES

- [1] U.-D. Ehlers, M. Lindner, S. Sommer, and E. Rauch, “AICOMP - Future Skills in a World Increasingly Shaped By AI,” *Ubiquity Proceedings*, Oct. 2023. [Online]. Available: <https://account.ubiquityproceedings.com/index.php/up-j-up/article/view/91>
- [2] M. J. Koehler, P. Mishra, and W. Cain, “What is Technological Pedagogical Content Knowledge (TPACK)?” *Journal of Education*, vol. 193, no. 3, pp. 13–19, Oct. 2013. [Online]. Available: <https://doi.org/10.1177/002205741319300303>
- [3] M. A. Thohir, J. Jumadi, and W. Warsono, “Technological pedagogical content knowledge (TPACK) of pre-service science teachers: A Delphi study,” *Journal of Research on Technology in Education*, vol. 54, no. 1, pp. 127–142, Jan. 2022, _eprint: <https://doi.org/10.1080/15391523.2020.1814908>. [Online]. Available: <https://doi.org/10.1080/15391523.2020.1814908>
- [4] D. A. Schmidt, E. Baran, A. D. Thompson, P. Mishra, M. J. Koehler, and T. S. Shin, “Technological Pedagogical Content Knowledge (TPACK): The Development and Validation of an Assessment Instrument for Preservice Teachers,” *Journal of Research on Technology in Education*, vol. 42, no. 2, pp. 123–149, Dec. 2009, _eprint: <https://doi.org/10.1080/15391523.2009.10782544>. [Online]. Available: <https://doi.org/10.1080/15391523.2009.10782544>
- [5] R. R. Puentedura, *SAMR: A Model for Educational Technology Integration*. Harvard University Press, 2014.
- [6] E. R. Hamilton, J. M. Rosenberg, and M. Akcaoglu, “The Substitution Augmentation Modification Redefinition (SAMR) Model: a Critical Review and Suggestions for its Use,” *TechTrends*, vol. 60, no. 5, pp. 433–441, Sep. 2016. [Online]. Available: <https://doi.org/10.1007/s11528-016-0091-y>
- [7] OECD, “The Future of Education and Skills: Education 2030,” OECD Education Policy Perspectives 98, Jun. 2018, series: OECD Education Policy Perspectives Volume: 98. [Online]. Available: <https://www.oecd.org/en/publications/the-future-of-education-and-skills-54ac7020-en.html>
- [8] M. Janssen and H. van der Voort, “Agile and adaptive governance in crisis response: Lessons from the COVID-19 pandemic,” *International Journal of Information Management*, vol. 55, p. 102180, Dec. 2020. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0268401220309944>
- [9] European Ministers of Education, “European Higher Education Area and Bologna Process.” Jun. 1999. [Online]. Available: <https://eha.info/page-ministerial-conference-bologna-1999>
- [10] I. Mergel, S. Ganapati, and A. B. Whitford, “Agile: A New Way of Governing,” *Public Administration Review*, vol. 81, no. 1, pp. 161–165, Jan. 2021. [Online]. Available: <https://onlinelibrary.wiley.com/doi/10.1111/puar.13202>
- [11] J. B. Biggs and C. S.-k. Tang, *Teaching for quality learning at university: what the student does*, 4th ed., ser. SRHE and Open University Press Imprint. Maidenhead: McGraw-Hill/Society for Research into Higher Education/Open University Press, 2011.
- [12] J. Biggs, “Enhancing teaching through constructive alignment,” *Higher Education*, vol. 32, no. 3, pp. 347–364, Oct. 1996. [Online]. Available: <https://doi.org/10.1007/BF00138871>
- [13] U.-D. Ehlers, *Future Skills: Lernen der Zukunft - Hochschule der Zukunft*, ser. Zukunft der Hochschulbildung - Future Higher Education. Wiesbaden: Springer Fachmedien Wiesbaden, 2020. [Online]. Available: <http://link.springer.com/10.1007/978-3-658-29297-3>
- [14] U.-D. Ehlers and E. Rauch, “AICOMP - An Investigation into Future Skills for a World Shaped Through AI,” in *DHBW AI Transfer Congress 2024 - Proceedings*. Stuttgart: Duale Hochschule Baden-Württemberg Baden-Wuerttemberg Cooperative State University, Jul. 2024, pp. 63–71.
- [15] European Education and Culture Executive Agency, Ed., *The European higher education area in 2024: Bologna process implementation report*. Luxembourg: Publications Office, 2024.
- [16] P. D. T. Seidl and C. Vohnhof, “Agile Prinzipien – Was kann die Studiengangsentwicklung davon lernen?” no. 3, Jul. 2017. [Online]. Available: <https://hochschulforumdigitalisierung.de/agile-prinzipien-was-kann-die-studiengangsentwicklung-davon-lernen/>
- [17] C. Loughlin, S. Lygo-Baker, and Lindberg-Sand, “Reclaiming constructive alignment,” *European Journal of Higher Education*, vol. 11, no. 2, pp. 119–136, Apr. 2021, _eprint: <https://doi.org/10.1080/21568235.2020.1816197>. [Online]. Available: <https://doi.org/10.1080/21568235.2020.1816197>
- [18] K. Beck, M. Beedle, A. van Bennekum, A. Cockburn, W. Cunningham, M. Fowler, and J. Grenning, “Manifesto for Agile Software Development,” 2001. [Online]. Available: <https://agilemanifesto.org/>
- [19] F. Härer, G. Herzwurm, and R. Schmidt, “Agilität in der Studiengangsentwicklung an deutschen Hochschulorganisationen – Eine Bestandsaufnahme,” *Gruppe. Interaktion. Organisation. Zeitschrift für Angewandte Organisationspsychologie (GIO)*, vol. 56, no. 1, pp. 185–195, Mar. 2025. [Online]. Available: <https://doi.org/10.1007/s11612-025-00799-z>
- [20] S. McKenney and T. Reeves, *Conducting Educational Design Research*, 2nd ed., 2018.
- [21] G. Ruge, O. Tokede, and L. Tivendale, “Implementing constructive alignment in higher education - Cross-institutional perspectives from Australia,” *Higher Education Research and Development*, p. 16, Nov. 2018.
- [22] N. Selwyn, *Should robots replace teachers?: AI and the Future of Education*. Polity Press, Nov. 2019. [Online]. Available: <https://research.monash.edu/en/publications/should-robots-replace-teachers-ai-and-the-future-of-education/>
- [23] O. Zawacki-Richter, V. I. Marín, M. Bond, and F. Gouverneur, “Systematic review of research on artificial intelligence applications in higher education – where are the educators?” *International Journal of Educational Technology in Higher Education*, vol. 16, no. 1, p. 39, Dec. 2019. [Online]. Available: <https://educationtechnologyjournal.springeropen.com/articles/10.1186/s41239-019-0171-0>
- [24] C. Argyris and D. A. Schön, “Organizational Learning: A Theory of Action Perspective,” *Reis*, no. 77/78, pp. 345–348, 1997. [Online]. Available: <https://www.jstor.org/stable/40183951>
- [25] D. A. Schmidt, B. Alboloushi, A. Thomas, and R. Magalhaes, “Integrating artificial intelligence in higher education: perceptions, challenges, and strategies for academic innovation,” *Computers and Education Open*, vol. 9, p. 100274, Dec. 2025. [Online]. Available: <https://linkinghub.elsevier.com/retrieve/pii/S2666557325000333>
- [26] P. Mishra and M. J. Koehler, “Technological Pedagogical Content Knowledge: A Framework for Teacher Knowledge,” *Teachers College Record*, vol. 108, no. 6, pp. 1017–1054, Jun. 2006. [Online]. Available: <https://doi.org/10.1111/j.1467-9620.2006.00684.x>
- [27] M. W. Allen and R. Sites, *Leaving ADDIE for SAM: An Agile Model for Developing the Best Learning Experiences*. American Society for Training & Development, Sep. 2012, google-Books-ID: DSQBJOt2_0cC.
- [28] F. E. Weinert, *Concept of competence: a conceptual clarification.. Seattle: Hogrefe & Huber.*, ser. D. S. Rychen & L. H. Salganik (eds) Defining and selecting key competencies, Seattle, 2001.
- [29] M. Mulder, “Competence-based Education and Training,” *The Journal of Agricultural Education and Extension*, vol. 18, no. 3, pp. 305–314, Jun. 2012, _eprint: <https://doi.org/10.1080/1389224X.2012.670048>. [Online]. Available: <https://doi.org/10.1080/1389224X.2012.670048>
- [30] R. Wesselink, H. Biemans, J. Gulikers, and M. Mulder, “Models and Principles for Designing Competence-based Curricula, Teaching, Learning and Assessment,” in *Competence-based Vocational and Professional Education: Bridging the Worlds of Work and Education*, M. Mulder, Ed. Cham: Springer International Publishing, 2017, pp.

533–553. [Online]. Available: https://doi.org/10.1007/978-3-319-41713-4_25

- [31] European Association for Quality Assurance in Higher Education (ENQA), European Students' Union (ESU), European University Association (EUA), and European Association of Institutions in Higher Education (EURASHE), "Standards and Guidelines for Quality Assurance in the European Higher Education Area (ESG)," Brussels, Belgium, 2015. [Online]. Available: <https://ehea.info/media.ehea.info/file/ESG/00/2/ESG2015616002.pdf>
- [32] D. He, "The Impact of Generative Artificial Intelligence on Higher Education," in *Proceedings of the 2nd International Conference on Intelligent Education and Computer Technology*. New York, NY, USA: Association for Computing Machinery, Nov. 2025, pp. 324–329. [Online]. Available: <https://dl.acm.org/doi/10.1145/3764206.3764256>
- [33] D. Long and B. Magerko, "What is AI Literacy? Competencies and Design Considerations," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, ser. CHI '20. New York, NY, USA: Association for Computing Machinery, Apr. 2020, pp. 1–16. [Online]. Available: <https://dl.acm.org/doi/10.1145/3313831.3376727>
- [34] U.-D. Ehlers and J. M. Mann, "Playbook Agile Studiengangentwicklung," Karlsruhe, Jan. 2026. [Online]. Available: <https://ai-comp.org/downloads/PlaybookAgileStudiengangentwicklung1.1.pdf>
- [35] J. Delcker, J. Heil, D. Ifenthaler, S. Seufert, and L. Spirgi, "First-year students AI-competence as a predictor for intended and de facto use of AI-tools for supporting learning processes in higher education," *International Journal of Educational Technology in Higher Education*, vol. 21, no. 1, p. 18, Mar. 2024. [Online]. Available: <https://doi.org/10.1186/s41239-024-00452-7>
- [36] Flair Collaboration, "AI competence frameworks and policies in higher education: Synthesis report," Vienna University of Economics and Business, Tech. Rep., 2025. [Online]. Available: <https://www.wu.ac.at/en/flair/synthesis-report>
- [37] U.-D. Ehlers and M. Lindner, "DHBW AI Transfer Congress 2024 Proceedings," no. DHBW AI Transfer Congress 2024 Proceedings, pp. 63–71, 2024.
- [38] M. Černý, "AI Literacy in Higher Education: Theory and Design | springerprofessional.de," 2024. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-031-63235-8_4
- [39] U.-D. Ehlers, "NextBlog - Unveiling the Truth of AI and Education," Nov. 2026. [Online]. Available: <https://next-education.org/de/team-blog/ne-teamblog/unveiling-the-truth-of-ai-and-education-from-red-pill-to-red-will/2026-03-11/41/>
- [40] R. M. Harden, "Approaches to curriculum planning," *Medical Education*, vol. 20, no. 5, pp. 458–466, Sep. 1986.
- [41] R. M. Harden, S. Sowden, and W. R. Dunn, "Educational strategies in curriculum development: the SPICES model," *Medical Education*, vol. 18, no. 4, pp. 284–297, Jul. 1984.
- [42] C. Bereiter, *Education and mind in the knowledge age*. Mahwah, N.J. : L. Erlbaum Associates, 2002. [Online]. Available: <http://archive.org/details/educationmindink0000bere>
- [43] A. LW, K. DR, A. PW, C. KA, R. Mayer, P. PR, J. Rath, and W. MC, *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives*, Jan. 2001.
- [44] J. B. Biggs and K. F. Collis, *Evaluating the Quality of Learning: The SOLO Taxonomy (Structure of the Observed Learning Outcome)*. Academic Press, May 2014, google-Books-ID: xUO0BQAAQBAJ.
- [45] D. A. Kolb and M. S. Plovnick, *The experiential learning theory of career development*. Cambridge, M.I.T., 1974. [Online]. Available: <http://archive.org/details/experientiallear00kolb>
- [46] J. Schmitt and S. T. Simon, "Participatory, Agile, Co-creative: Identifying Topics for a Future-Oriented, Innovative Research on Digital Transformation," *Weizenbaum Journal of the Digital Society*, vol. 3, no. 1, Dec. 2023. [Online]. Available: <https://ojs.weizenbaum-institut.de/index.php/wjds/article/view/316>

From Lead to Loyalty: A Systematic Review of AI Use Cases for SMEs along the Customer Journey

1st Tessa Steinigke

Information Systems, Data Science and AI
DHBW Heilbronn
Heilbronn, Germany
tessa.steinigke@dhbw.de

2nd Patrick Föll

Information Systems, Data Science and AI
DHBW Heilbronn
Heilbronn, Germany
patrick.foell@dhbw.de

Abstract—Artificial intelligence (AI) offers potential for small and medium-sized enterprises (SMEs), yet adoption rates remain low compared to the documented benefits. This systematic literature review synthesises 51 publications to map AI use cases in SME sales and marketing along the three stages of the customer journey: prepurchase, purchase, and postpurchase. The review identifies and clusters AI applications across these stages, with cross stage tools, particularly AI-powered customer relationship management (CRM) systems and chatbots, emerging as the most analysed category. The prepurchase stage attracts the greatest research attention, centred on content creation, search engine optimization (SEO), personalisation and demand forecasting. The purchase stage covers AI-driven product recommendations, dynamic pricing and UX innovations in e-commerce. The post-purchase stage provides insights into customer sentiment and engagement, while also enabling the prediction of churn and retention, along with automatic countermeasures. A recurring finding is the persistent gap between AI awareness and actual adoption, driven primarily by financial constraints, lack of technical expertise, and organisational inertia. Based on the synthesised evidence, this paper provides a decision-oriented entry map of AI use cases and derives actionable recommendations to guide SME practitioners and policymakers in overcoming adoption barriers.

Index Terms—AI, SME, Sales, Marketing, Customer Journey, CRM, Chatbots, Adoption

I. INTRODUCTION

Artificial intelligence (AI) is among the most rapidly expanding fields in technology today [17]. The solution landscape has become increasingly fragmented, with a rapidly growing number of AI-enabled tools and services. This abundance, however, rarely translates into actionable orientation for decision makers. The choice is overwhelming for managers, with many reporting that they do not adopt AI precisely because they lack structured guidance [11]. A significant share even confuse basic automation with AI or experiment with tools like ChatGPT without a clear strategic direction [19].

While large corporations increasingly integrate AI into their operations and gain benefits, small and medium-sized enterprises (SMEs) continue to lag behind [19]. According to figures from the European Commission, 99.8% of all enterprises in the EU are SMEs, contributing 53.6% of total added real value [12]. Despite this pivotal role, SMEs are constrained by scarce resources and face considerable challenges in adopting AI technologies [23, 24]. Research from

emerging markets confirms a notable gap between awareness and actual usage: although awareness of AI tools among SMEs is moderate, practical application remains significantly low [10]. The underlying barriers are persistent — limited financial and human resources, a lack of technical expertise, and insufficient digital infrastructure [35, 27, 10].

Customer-facing functions such as marketing and sales are frequently discussed as pragmatic entry points for SMEs, as they involve many standardised tasks and provide direct links to customer value creation [27, 25]. In this domain, AI has demonstrated the ability to generate revenue and enhance the performance of business organisations [2].

The existing body of literature addresses individual AI applications in isolation, yet there is a lack of synthesised overviews that map these findings into a coherent map of AI use cases in marketing and sales from an SME perspective. The objective is to develop a decision-oriented entry map through a structured literature review that synthesizes AI application areas along the marketing and sales process for SMEs, identifies barriers and enablers, and provides actionable orientation for practitioners. This study pursues two interconnected goals:

- (i) To synthesise the current state of research on AI in SME marketing and sales by identifying and clustering use cases along the customer journey (prepurchase, purchase, and postpurchase), while highlighting cross stage applications as a distinct category.
- (ii) To identify the main barriers to adoption and synthesise recommendations reported in the literature.

The customer journey framework proposed by [20] serves as the organising structure for goal (i). Unlike frameworks such as TAM, UTAUT and the TOE, which focus on why organisations adopt technology, the customer journey offers a process-oriented, customer-centric approach that maps AI applications to the concrete stages at which SMEs deploy them. This perspective directly supports the decision-oriented intent of this review.

The remainder of this paper is structured as follows: The methodology employed is outlined, followed by a detailed report on the execution of the research in the field of AI in SMEs. The results are then interpreted and discussed in order

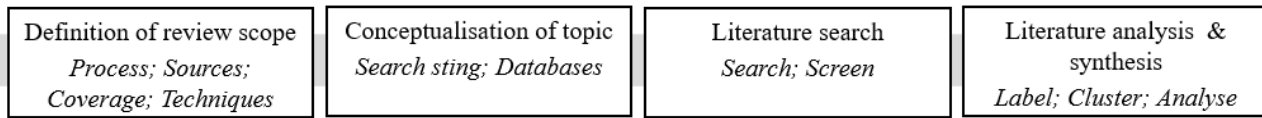


Fig. 1. Methodology overview.

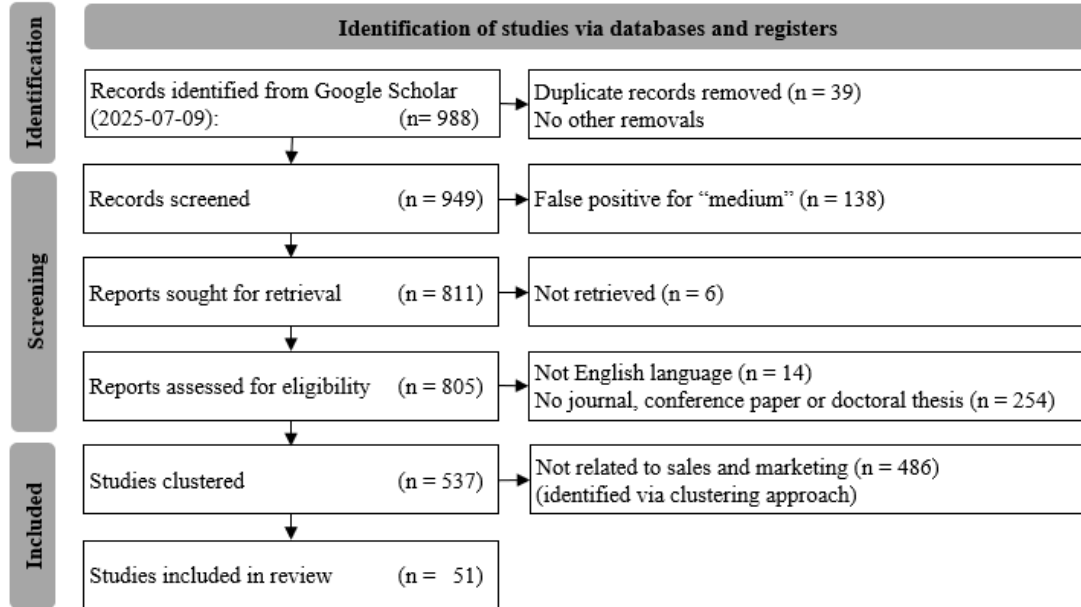


Fig. 2. Literature search process.

to extract actionable insights for researchers and practitioners. The paper concludes with a synthesis that addresses the stated goals and provides strategic recommendations for advancing AI integration in SME marketing and sales.

II. RESEARCH APPROACH AND IMPLEMENTATION

The methodological approach for this literature analysis follows the process proposed by vom Brocke [36]. An overview is given in figure 1.

(i) The scope of the review is defined according to establishing clear boundaries and objectives for the research. Adopting an outcome-oriented perspective, this study produces a stand-alone review that classifies and categorizes existing research on AI in SMEs within marketing and sales. The process begins with a sequential systematic literature search, followed by an iterative definition of analysis dimensions. As the primary source, a citation indexing service serves and a comprehensive coverage approach is pursued to collect as many relevant publications as possible rather than focusing exclusively on seminal works. The search techniques employ keyword searches complemented by reference-based backward searches.

(ii) The topic is conceptualised by identifying key terms and relevant databases by reading sources that are likely to provide topic summaries and overviews of the key issues. As it is the goal to provide a broad overview of the available literature on AI in SMEs which can later be analysed for the

context of marketing and sales, it was opted for the following more generic research string: *allintitle: (AI OR "artificial intelligence" OR "a.i.") (SME OR "s.m.e." OR SMEs OR medium OR "medium-sized")*. The search is restricted to titles only, representing a deliberate methodological decision that prioritises precision over recall. While this reduces the risk of including studies where AI and SMEs are only peripheral aspects, it may systematically exclude relevant work in which these topics appear prominently in abstracts or keywords but not in titles. To partially compensate for this limitation, backward and forward citation searches were conducted for all publications in the final sales and marketing cluster.

(iii) The literature search is conducted within the broader context of AI in SMEs and employs search strategies by a database search using a predefined keyword string. The search process was conducted via Google Scholar on July 9, 2025, without applying restrictions regarding publication year or other parameters. The defined search string initially yielded 949 results after removing duplicates. Additional academic databases were piloted, but they yielded only marginal incremental results for the deliberately broad SME/AI scope. Therefore it is proceeded with Google Scholar as the primary indexing source, complemented by citation chasing. During the screening phase, only minimal filtering criteria were applied in order to capture the widest possible scope of relevant research. The inclusion criteria were limited to peer-reviewed journal articles, conference proceedings, and doctoral theses.

No restrictions were imposed regarding journal rankings or conference quality criteria. Publications were excluded if they were not written in English, were inaccessible, or represented false positives. The latter occurred primarily due to the ambiguous use of the term “medium” and were subsequently removed from the dataset. No additional data sources were incorporated, except for selected methodology and framework papers. After the initial screening, 537 publications were retained for the first stage of analysis. Publications are labelled with characteristics in multiple dimensions defined through a strong iterative process over multiple iterations. A data-driven clustering based on assigned labels enabled the identification of a distinct cluster focusing on sales and marketing. A subsequent backward and forward citation search was conducted for the publications within this cluster. All sources explicitly identifying their subjects as SMEs were included, irrespective of the underlying size thresholds, sectoral context, business model, or geographic region applied by the respective authors. The resulting final corpus comprises 51 publications. The complete process is illustrated in Figure 2.

(iv) The identified literature is subsequently analysed by systematically structuring topic-related concepts into distinct units of analysis. In this process, publications concerning sales and marketing are grouped into a dedicated cluster and examined collectively. Within the cluster backward and forward citation searches are carried out. This approach ensures that not only explicitly relevant studies, but also those with an implicit connection to the topic, are considered. The publications are analysed to identify the current areas of focus and use cases in AI research for SMEs in sales and marketing. To achieve this, the research areas in this field are categorised according to the three stages of the customer journey: prepurchase, purchase and postpurchase – following the framework proposed by Lemon and Verhoef [20] – with additional consideration given to cross stage applications that span multiple stages. All four stages are shown in figure 3.

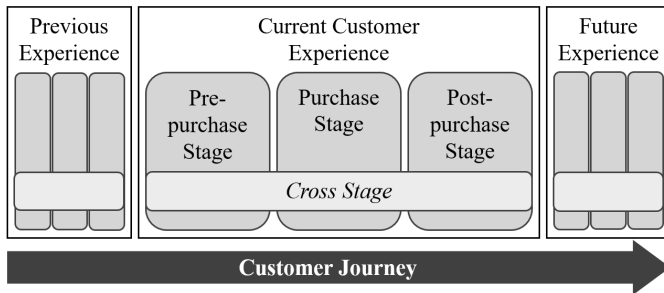


Fig. 3. Customer Journey following Lemon and Verhoef (2016) [20] with additional cross stage.

The prepurchase stage encompasses all interactions before the transaction, including need recognition, information search, and consideration of alternatives. The purchase stage covers the actual purchasing event-choice, ordering, and payment. Finally, the postpurchase stage includes consumption, service encounters, and behaviours such as word-of-mouth

and repurchase decisions. An application was classified as cross stage when the reviewed literature documented its use across at least two distinct stages of the customer journey. The assignment of individual publications to the respective stages was carried out by both authors independently, based on the substantive content of each paper. Disagreements were resolved through discussion and consensus.

(v) According to [36] the last step would be the development of a research agenda; since developing a comprehensive research agenda is beyond the scope of this article, this step has been excluded.

III. RESULTS

The following subsections present and discuss the findings of the publications, structured according to the customer journey. The results of the study are concluded with a discourse on adoption barriers, readiness, and opportunities.

A. Use cases

First, the cross stage use cases are shown, followed by an analysis of the prepurchase, purchase, and postpurchase stages.

1) Cross Stage

Several AI applications extend beyond individual stages of the customer journey, exerting influence across the entire purchase process. Instead of examining their impact at each stage separately, an overview is provided here and illustrated in Table I.

TABLE I
AI USE CASES – CROSS STAGE

Use Case	Tools / Technologies	Refs
Chatbots & Virtual Assistants for 24/7 Customer Support & Q&A	NLP chatbots, rule-based and ML-based chatbots; Rufus (Amazon)	[33], [38], [16], [40]
AI-driven CRM for Automating the Customer Journey & Restructuring Sales Funnels	HubSpot CRM, Zoho CRM (with Zia), Salesforce Einstein; OKKI xiaoman technology	[3], [7], [40], [14]
AI-assisted Social Listening for Synthesising Multi-Channel Engagement Data	aiSnoop (AI-powered social listening tool for SMEs), NLP, sentiment analysis	[13]

Note: Tools/ technologies are examples mentioned in the reviewed studies.

AI-Powered CRM Systems. AI-powered customer relationship management (CRM) systems have been shown to streamline the entire customer journey. Studies found that platforms such as HubSpot, Zoho, Salesforce Einstein and Workflows (like Zapier) help automate customer interactions, optimise marketing strategies and personalise customer experiences across sectors [3, 14]. A case study of African SMEs specifically highlighted how these systems streamline lead nurturing, customer support and personalised marketing simultaneously [3]. Survey evidence confirms that AI-driven

CRM practices — encompassing customer data management, interaction automation, segmentation, predictive analytics and sales optimisation — significantly enhance sustainable growth [7]. An important gap in existing tooling is addressed by the *aiSnoop* social listening tool, designed specifically for SME managers to synthesise multi-channel engagement data into actionable insights, addressing the common issue that widely available tools are too costly and complex for SMEs [13].

Chatbots and Virtual Assistants. Chatbots constitute the most empirically examined AI application in the cross stage category. A review established that AI chatbots, enabled by advances in natural language processing (NLP) and machine learning, have transformed customer service through 24/7 availability, routine task automation and user-data-driven personalisation [38]. Across the empirical literature, perceived ease of use and perceived usefulness consistently emerge as significant drivers of user satisfaction and adoption [15, 16]. Resolution rate was found to have the strongest effect on customer retention among all chatbot efficiency metrics, followed by response time and user engagement, suggesting that SMEs should prioritise issue resolution over mere responsiveness [33]. Chatbot characteristics such as personalisation, accuracy and credibility were found to significantly influence brand perception, with customer attitude and satisfaction serving as key mediators [37]. In a Thai organic farming cooperative context, an AI chatbot was integrated into a dedicated e-commerce platform to facilitate communication between producers and buyers, receiving high satisfaction scores from both experts and end users [31]. In human-AI customer support communities, domain experts were found to prefer teaching chatbots through demonstration over sorting or labelling paradigms, a finding with practical implications for SMEs seeking to train and supervise AI-assisted support tools [6].

AI-Driven Personalisation. A comprehensive framework outlined how machine learning, NLP and predictive analytics enable dynamic customer segmentation and personalised recommendations [25]. Building on dynamic capabilities theory, a theoretical framework linking generative AI to customer personalisation development concluded that Generative AI enables context-oriented personalisation with downstream effects on interactive marketing, value co-creation and consumer loyalty [1].

2) Prepurchase Stage

The prepurchase stage is arguably where AI has attracted the most research attention. A study of generative AI and ChatGPT usage in an SME found that the marketing department experienced the greatest productivity gains [39]. This underscores that the prepurchase domain is where AI's value proposition is most immediately tangible for SMEs. Table II summarizes the use cases for this stage.

Content Creation and SEO. An analysis of German SME marketing applications identified AI tools as being most valued for content creation, search engine optimization (SEO), social media and blog marketing, with users particularly appreciating

TABLE II
AI USE CASES – PREPURCHASE STAGE

Use Case	Tools / Technologies	Refs
Dynamic & Personalised Content Creation	ChatGPT; Facebook, Instagram, and WhatsApp platform AI; Canva; Jasper; Adobe Firefly; Microsoft Copilot; Midjourney	[41], [19], [39], [40], [4]
Text Production	ChatGPT; DeepL; Neuroflash; Jasper AI; Microsoft Copilot	[32], [39], [27], [19]
AI Live-Streaming as a Sales Channel	Taobao Live, TikTok/Douyin, AI avatar streaming platforms	[28]
AI-powered SEO & Search Engine Marketing	ChatGPT, DeepL, Neuroflash; Google Ads/Console, SEMRush; ChatGPT/Bard/Perplexity for keyword research	[27], [14], [17]
Advertising Campaign Optimisation & Audience Targeting & Strategy Development	Google Ads, LinkedIn Ads, Meta/Facebook Ads, SEMRush; generic AI advertising automation, Big data AI analytics	[14], [4], [44]
Predict Market Development (Demand & Trends)	AI monitoring system based on historical sales data / ML models; generic forecasting models / big data analytics	[9], [5], [4]

Note: Tools/ technologies are examples mentioned in the reviewed studies.

time savings and reduced dependence on external service providers [27]. AI-assisted text production tools such as ChatGPT, DeepL and Neuroflash have become central to this workflow [32, 27, 39]. AI-driven content diversity was shown to increase average time spent on websites, with personalisation significantly enhancing lead-to-sales conversion rates [41]. These findings collectively suggest that dynamic, personalised content creation represents one of the most mature and widely adopted AI use cases for SMEs in the prepurchase phase.

Advertising and Targeting. AI-based tools such as Google Ads, Meta Ads and SEMRush support advertising campaign optimisation and audience targeting [14, 4]. One case study from an American medium-sized business illustrated how AI was used to systematically restructure a marketing strategy and build a more effective sales funnel [14]. Big-data AI analytics have further been shown to enhance digital marketing capabilities and support international market strategy, particularly for SMEs in cultural and creative industries [44].

Market Prediction. AI also enables SMEs to predict market developments. An AI-enabled centralised monitoring system was proposed to help SMEs observe and forecast upcoming demand based on historical data and continuous machine learning [9]. Generic forecasting models and big data analytics serve a similar function in several other studies [5, 4].

AI Live-Streaming as a Sales Channel. An emerging phenomenon in the prepurchase phase is AI live-streaming, particularly in the Chinese cross-border e-commerce context. A dual-analysis study of 260 Chinese e-commerce SMEs found that perceived ease of use, perceived usefulness, management support, employee competency and market trends significantly

influence the adoption of AI live-streaming on platforms such as Taobao Live and TikTok/Douyin [28]. Competitive pressure did not show a significant effect [28].

3) Purchase Stage

Research on AI at the actual point of purchase is comparatively thinner, with most contributions focusing on predictive models and UX innovations in e-commerce. Table III summarizes the use cases for this stage.

TABLE III
AI USE CASES – PURCHASE STAGE

Use Case	Tools / Technologies	Refs
Personalized Product and Service Suggestions / Predictions	DCNN-SPM (Deep Convolutional Neural Network – Sales Prediction Model), ML recommendation algorithms, generative AI	[43], [25], [1]
AI-Driven Cross-Selling and Upselling	ML recommendation engines, AI-powered CRM (HubSpot, Zoho, Salesforce Einstein), predictive lead scoring	[25], [7], [35], [42]
AI-based UX Solutions for E-Commerce Innovations	AI clustering algorithms (Skin-Tone Detection), deep learning models	[30]
Dynamic Price Optimization	Generic price optimization algorithms	[4]

Note: Tools/ technologies are examples mentioned in the reviewed studies.

Personalised Product Recommendations and Sales Prediction. A Deep Convolutional Neural Network-based Sales Prediction Model was proposed for analysing SME e-commerce utilisation, achieving a customer satisfaction ratio of 98.7% and a customer buying behaviour analysis accuracy of 97.6% [43]. The model extracts relevant features from raw log data — including browsing history, user reviews, and purchase patterns — to forecast product sales in real time and deliver personalised product suggestions at the point of sale [43]. Complementing this, ML recommendation algorithms and generative AI further enable context-sensitive product suggestions that adapt dynamically to individual customer profiles during the purchasing process [25, 1]. The ability of conversational AI agents to support purchase decisions was empirically confirmed among SME customers: perceived ease of use and perceived usefulness were found to shape customer attitudes, which in turn positively predicted the actual use of conversational agents on SME e-commerce platforms [16].

AI-Driven Cross-Selling and Upselling. Beyond static product suggestions, AI-based recommendation engines have emerged as effective tools for identifying and acting on cross-selling and upselling opportunities at the moment of purchase [42]. By analysing purchase histories and real-time browsing behaviour, these systems surface complementary products or premium alternatives that align with individual customer preferences, thereby increasing average transaction value [25, 42]. AI-powered CRM systems further support this process by enabling sales teams to prioritise high-potential up-sell

opportunities through predictive lead scoring and automated customer segmentation [7]. SME managers explicitly identified AI-based recommendation engines as a mechanism for boosting cross-selling, noting that enhanced personalisation not only improves conversion rates but also strengthens brand loyalty [35].

AI-Based UX Innovations. A user experience analysis of an AI-based skin tone clustering system for clothing SMEs demonstrated that AI can address highly specific functional UX challenges in e-commerce, achieving a user acceptance rate of 90.67% [30].

Dynamic Pricing. Dynamic price optimisation algorithms represent a further purchase stage use case, though this remains under-explored empirically in the field of SMEs [4].

4) Postpurchase Stage

The post-purchase stage, which encompasses customer satisfaction tracking, retention, and churn prevention, is the least extensively discussed area in the current literature, often being embedded in broader studies. Table IV summarizes the use cases for this stage.

TABLE IV
AI USE CASES – POSTPURCHASE STAGE

Use Case	Tools / Technologies	Refs
AI-Driven Sentiment Analysis	NLP sentiment analysis engines, social media monitoring tools	[18]
Customer Engagement & Satisfaction Tracking	Generic marketing AI; NLP, ML, data analysis modules	[21]
Predictive Analytics for Customer Behaviour Forecasting	ML models (churn prediction), predictive analytics modules in CRM systems (generic)	[25], [7]
Personalised Retention Offers for Churn Prevention	AI-based recommendation engines for automated retention offers	[25]

Note: Tools/ technologies are examples mentioned in the reviewed studies.

AI-Driven Sentiment Analysis. NLP-based sentiment analysis enables SMEs to monitor customer feedback, reviews, and social media interactions in real time after a purchase, providing actionable signals about satisfaction levels and brand perception [18]. Negative sentiment can trigger targeted service recovery before dissatisfaction escalates, while positive sentiment helps identify loyal segments and product strengths to inform marketing and product decisions [18].

Customer Engagement Tracking. The effect of AI on customer engagement remains empirically contested. While marketing AI has been shown to enhance sustainable financial performance indirectly through customer engagement [21] and generative AI was found to positively influence customer engagement [10], other studies report weaker effects: AI-driven engagement strategies showed only a limited correlation with sustainable growth [7], and AI did not moderate the relationship between customer experience and business performance [34]. This suggests that AI alone is not a sufficient condition

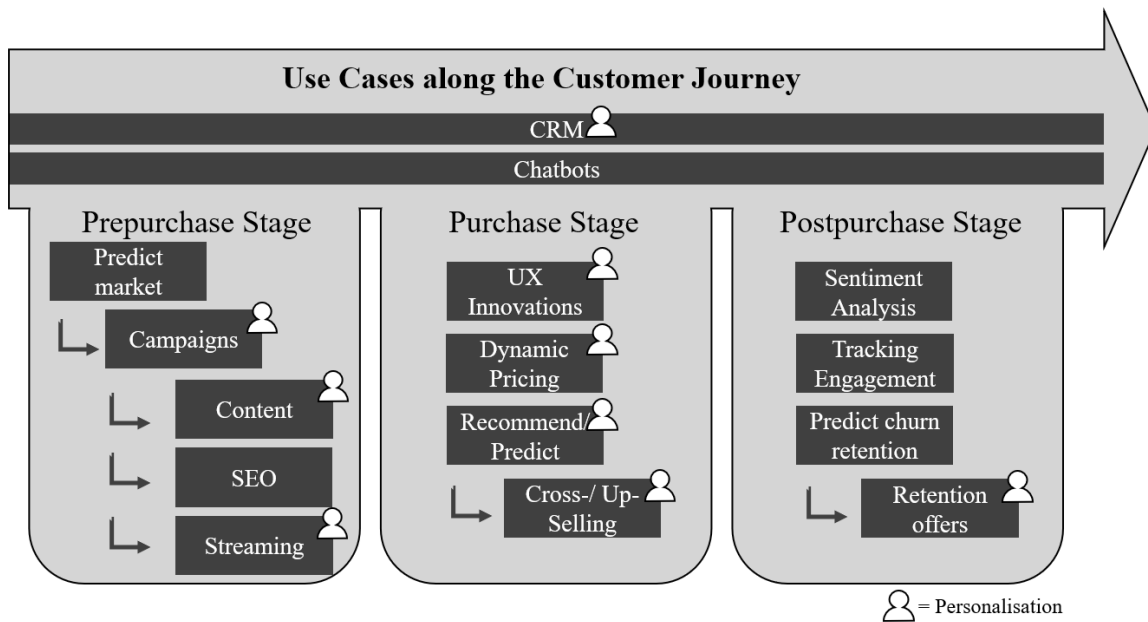


Fig. 4. Use cases in sales and marketing for SMEs along the customer journey.

for engagement improvements; contextual and organisational factors matter significantly.

Predictive Analytics and Personalised Retention Offers for Churn Prevention. ML-based churn prediction models and predictive analytics modules integrated into CRM systems are identified as relevant tools for behavioural forecasting in the postpurchase phase [25, 7]. AI recommendation engines use these predictions to automatically trigger personalised retention interventions, such as individual discount codes, loyalty rewards, or alternative product suggestions, when disengagement signals are detected [25]. A survey of 450 SMEs confirmed that predictive analytics as part of AI-driven CRM practices significantly enhances sustainable growth, though engagement strategies showed a comparatively weaker effect [7]. This suggests that while AI can improve retention through prediction, its impact is most pronounced when embedded in broader CRM architectures rather than deployed in isolation.

B. Adoption Barriers, Opportunities and Readiness

Beyond the customer-journey-specific use cases, the literature highlights cross-cutting factors related to AI adoption barriers and enablers in SMEs, organisational readiness, and business performance effects. Since these dimensions apply across all application areas and align with the second objective of this review, they are analysed separately in this section.

Awareness-Adoption Gap. Although SMEs increasingly utilise generative AI tools to overcome challenges such as resource constraints and high marketing costs, they continue to lag behind larger firms in adoption [19]. The gap between awareness and actual AI adoption is a persistent finding. In southeastern Nigeria, while SMEs demonstrated moderate awareness of generative AI and chatbots, practical application remained significantly low [10]. In Bangkok, 65% of SMEs

were aware of AI technologies, but only 25% had started incorporating them into marketing plans, with the majority remaining in early research and development phases [35]. In Nigeria's retail sector, 70% of SMEs had adopted AI to some extent, but only 20% had fully integrated it into their operations, with customer behaviour analysis (60%) and inventory management (50%) as the most common applications [26].

Barriers. The key barriers to adoption (Figure 5) are consistently identified across geographies. Financial constraints emerge as the most prominent obstacle, with over 70% of SMEs citing limited financial capacity as a barrier [35, 26]. Lack of technical expertise is the second most frequently cited challenge, as many SMEs lack dedicated IT departments or access to qualified AI professionals [35, 23]. Cultural resistance and organisational inertia represent a third barrier category, with employees and managers often reluctant to deviate from established practices [35]. Additional barriers include insufficient government support, data privacy concerns, lack of guiding standards and dependence on outdated data [23, 22]. In the German context, criticism of AI tools was relatively minimal, primarily related to possible errors and content inconsistencies, suggesting that practical usability barriers may be less severe than perceived adoption barriers [27]. The observed adoption barriers align with the Technology-Organisation-Environment (TOE) framework. [29] extend the classical TOE model by introducing AI integration, innovation culture, and customer tech-savviness as additional adoption determinants, finding that all three factors significantly influence e-commerce adoption and, in turn, marketing performance.

Opportunities. Despite these challenges, the evidence for positive performance effects (Figure 5) is strong. AI adoption has been linked to improved decision-making (65% of respon-

dents) and enhanced operational efficiency (60%) in Nigerian retail SMEs [26]. A study in Saudi Arabia confirmed that AI-assisted social media marketing increases customer bases and profitability [8]. In Ghana, AI Marketing was shown to have a significant impact on multiple performance dimensions [2]. Generative AI and chatbot usage jointly accounted for 98.1% of CRM capability variance [10]. The Jordanian study found a 42.5% enhancement in sustainable financial performance driven by marketing AI adoption [21].

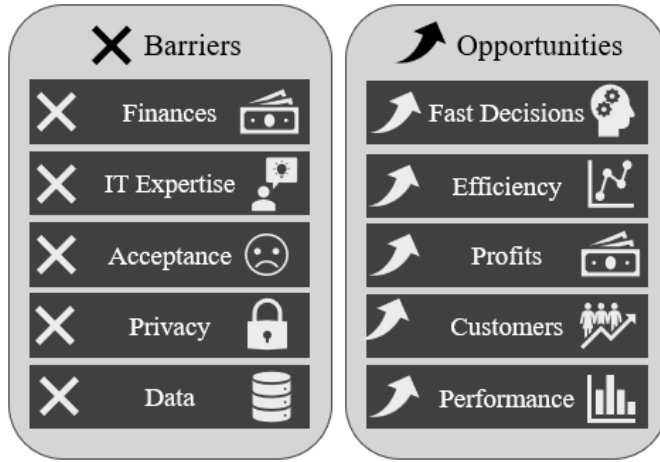


Fig. 5. Barriers and Opportunities of AI in sales and marketing for SMEs.

Recommendations. [17] proposes a four-step framework for marketing strategy innovation covering environment analysis, channel selection, strategy alignment, and effectiveness evaluation. With more focus on AI use case creation [14] structures AI implementation into two concrete phases: first, a Customer Segment Assessment, in which existing marketing channels and CRM data are analysed to identify potential areas of application; second, an AI Use Case Identification, in which tasks that are repetitive or consume a disproportionate share of working time are systematically evaluated for their ROI potential and prioritised accordingly.

For SMEs seeking to overcome adoption barriers, the literature converges on several recommendations also shown in figure 6.

Phase I – Customer Segment Assessment:

- Analyse the environment and data of existing marketing channels, the CRM, technical capabilities, operational capacity and the customers [14, 17]
- Integrate AI overall into SME development policies [10]
- Check for subsidised access and offered government incentives [26, 35]
- Block time for optimisation meetings, evaluation of AI and decision making [14]

Phase II – Use Case Identification & Implementation:

- First build on existing software subscriptions [14]
- Build affordable SaaS and mobile-first solutions [3]
- Partner with technology providers [35]
- Execute targeted training programmes [35, 23], e.g. generative AI for text creation [32]

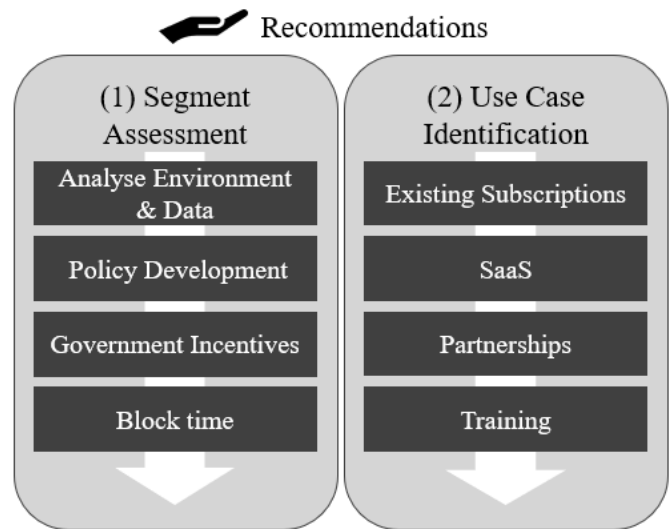


Fig. 6. Recommendations for AI Adoption in SMEs.

IV. DISCUSSION AND IMPLICATIONS

The reviewed literature confirms that AI applications in SME sales and marketing are most mature in the prepurchase stage, where content creation, SEO and CRM automation have accumulated the most consistent empirical support. Cross stage tools – above all AI-powered CRM systems and chatbots – emerge as the practically most impactful category, with personalisation as a unifying mechanism that cuts across all three stages of the customer journey. Purchase and postpurchase stage applications, by contrast, remain comparatively under-explored, and dynamic pricing or churn prediction have received limited empirical validation in SME-specific settings.

The evidence converges on a staged approach to AI integration. A strategy framework suggested in the literature proposes beginning with low-risk, task-specific tools before progressing to more complex generative applications [10], a logic that maps naturally onto the customer journey structure identified in this review. In concrete terms, this means starting with accessible SaaS solutions for content creation and CRM automation in the prepurchase stage, where productivity gains are most immediately tangible and switching costs remain low. As internal competency grows through targeted training, SMEs can extend AI adoption into the purchase stage – through personalised product recommendations and dynamic pricing – and eventually into postpurchase applications such as sentiment analysis and churn prediction, which demand more mature data infrastructure and analytical capability. It should be noted that the literature does not converge on a single entry point: while some studies emphasise content creation and generative AI as the most accessible starting tools [27, 39], CRM-oriented research prioritises chatbot and customer data integration as the higher-impact first step [18, 3]. A third strand argues that any tool-level recommendation is premature without prior investment in workforce skills and organisational culture [35, 29]. This suggests that the optimal entry point is context-dependent and cannot be prescribed universally.

Several findings nonetheless warrant critical scrutiny. The positive performance effects reported across the literature must be interpreted with caution, since the precise impact of AI adoption is notoriously difficult to isolate and quantify. Performance is an inherently multidimensional construct, and the application of traditional financial metrics alone has been widely criticised for its limitations in capturing the full range of AI-driven value creation [2]. Whether AI drives performance improvement or whether better-resourced, higher-performing SMEs are simply more likely to adopt AI in the first place remains an open empirical question. The geographic concentration of the evidence base in Nigeria and South-east Asia compounds these concerns, as findings from these contexts may not transfer readily to European SMEs, where regulatory environments such as GDPR, market maturity, and resource profiles differ substantially. Longitudinal and quasi-experimental designs, objective performance measurement, experiences out of real implementations and greater attention to developed-market SME contexts represent the most pressing directions for future work.

This review is itself subject to limitations that should temper the confidence with which its conclusions are drawn. The exclusive use of Google Scholar and the restriction of searches to title fields increase thematic precision but reduce recall, and no formal quality assessment was applied to the included studies, meaning methodologically weaker contributions enter the synthesis on equal footing with more rigorous ones.

V. CONCLUSION

This review set out to achieve two interconnected objectives: to provide a comprehensive, structured overview of AI use cases in SME sales and marketing, and to identify the key barriers to adoption alongside evidence-based recommendations for practitioners.

Regarding the first objective, the analysis demonstrates that AI applications are distributed unevenly across the customer journey. The prepurchase stage is the most research-rich domain, with content creation, SEO optimisation, personalised advertising and market forecasting representing well-documented, practically accessible entry points for SMEs. Cross stage applications — particularly AI-powered CRM systems and chatbots — constitute the empirically strongest category, with consistent evidence of positive effects on customer retention, satisfaction and operational efficiency. The purchase stage is characterised by very powerful AI-driven product recommendations, dynamic pricing and UX innovations in e-commerce. The postpurchase stage provides insights into customer sentiment and engagement, while also enabling the prediction of churn and retention, along with the automatic implementation of countermeasures. Personalisation emerges as a unifying principle that cuts across all three stages and represents a key mechanism through which AI creates measurable value.

Regarding the second objective, financial constraints, limited technical expertise and cultural resistance are consistently identified as the most significant barriers across di-

verse geographic and sectoral contexts. The gap between AI awareness and practical adoption is a persistent and cross-cultural finding. The evidence shows that when adoption does occur, it has an impact on performance — improving decision-making, enhancing customer engagement, increasing revenues and strengthening operational efficiency.

For SME practitioners, the findings suggest a staged approach to AI integration: beginning with accessible, low-cost tools for content creation and CRM automation, building internal competency through targeted training, and progressively extending AI capabilities along the customer journey as organisational readiness increases. Affordable SaaS solutions, mobile-first strategies and partnerships with technology providers offer pathways to overcoming resource constraints.

For researchers, this review highlights several areas requiring further investigation. Longitudinal and quasi-experimental study designs are needed to establish causal relationships between AI adoption and business performance. European and other developed-market SMEs remain under-represented in the evidence base. The marketing literature of SMEs lacks advanced AI technologies, including computer vision, voice interfaces and reinforcement learning. Detailed, replicable implementation frameworks with embedded cost-benefit analyses represent a critical gap that future research should address.

The question for most SMEs is no longer whether AI delivers value, but how and where to begin. Even modest first steps today build the organisational readiness required for deeper integration tomorrow. In this sense, AI adoption is less a technological leap than a strategic journey — one that SMEs, despite their resource constraints, are increasingly well-equipped to undertake.

ACKNOWLEDGMENT

AI tools were used in this work to support the labeling of papers following the initial literature search. All classifications were subsequently reviewed and verified manually to ensure accuracy and relevance. AI was used to assist in drafting parts of the text and improving the linguistic quality of the manuscript. All AI-generated content was carefully checked, revised, and validated by the authors.

REFERENCES

- [1] Kwabena Abrokwa-Larbi. “The Role of Generative Artificial Intelligence (GAI) in Customer Personalisation (CP) Development in SMEs: A Theoretical Framework and Research Propositions”. In: *Industrial Artificial Intelligence* 1.1 (2023), pp. 1–11. DOI: 10.1007/s44244-023-00012-4.
- [2] Kwabena Abrokwa-Larbi and Yaw Awuku-Larbi. “The Impact of Artificial Intelligence in Marketing on the Performance of Business Organizations: Evidence from SMEs in an Emerging Economy”. In: *Journal of Entrepreneurship in Emerging Economies* 16.4 (2024), pp. 1090–1117. DOI: 10.1108/JEEE-07-2022-0207.

- [3] Justina Eweala Abubakar, Samuel Omokhafa Yusuf, and Prosper Onagie Yusuf. "From Branding to Sales Conversion: How AI-Powered CRM Systems Shape the Customer Journey for African SMEs". In: *International Journal of Science and Research Archive* 14.2 (2025), pp. 494–506. DOI: 10.30574/ijrsra.2025.14.2.0369.
- [4] Carla Guillermina Mendoza Arce et al. "Optimizing Business Performance: Marketing Strategies for Small and Medium Businesses Using Artificial Intelligence Tools". In: *Migration Letters* 21.S1 (2023), pp. 193–201. DOI: 10.59670/ml.v21iS1.6008.
- [5] Monika Arora. "Adoption of Ai and E-Commerce Improving Marketing Performance of SMEs". In: *Academy of Marketing Studies Journal* (2023).
- [6] Zahra Ashktorab et al. "SME-in-the-loop: Interaction Preferences When Supervising Bots in Human-AI Communities". In: *Proceedings of the 2023 ACM Designing Interactive Systems Conference*. Pittsburgh PA USA: ACM, 2023, pp. 2281–2303. ISBN: 978-1-4503-9893-0. DOI: 10.1145/3563657.3596100.
- [7] Joseph Azage and Peter Chiazor Ikepeazu. "Ai-Driven Customer Relationship Management Practices and Sustainable Growth of Nigerian Smes". In: *International Journal of Business and Management Research* 5.3 (2025).
- [8] Wael Basri. "Examining the Impact of Artificial Intelligence (AI)-Assisted Social Media Marketing on the Performance of Small and Medium Enterprises: Toward Effective Business Management in the Saudi Arabian Context". In: *International Journal of Computational Intelligence Systems* 13.1 (2020), pp. 142–152. DOI: 10.2991/ijcis.d.200127.002.
- [9] M. Kazem Chamran, Nariman Fouladinezhad, and Feras Fathi Albdiwy. "An AI-Enabled Centralized Monitoring System to Predict SME Inventory Level". In: *2025 17th International Conference on Computer and Automation Engineering (ICCAE)*. 2025, pp. 400–404. DOI: 10.1109/ICCAE64891.2025.10980557.
- [10] Arachie Augustine Ebuka et al. "Artificial Intelligence Adoption and Business Performance: Evidence from Small and Medium Enterprises in Emerging Markets". In: *International Journal of Public Administration and Management Research* 11.2 (2025), pp. 69–86. ISSN: 2350-2231.
- [11] Rachid Ejjami. "Launching AI Marketing Solutions for Small and Medium Enterprises in Morocco: A Teaching Case Study". In: *International Journal of Teaching and Case Studies* 14.3 (2024), pp. 276–292. DOI: 10.1504/IJTCS.2024.137512.
- [12] European Commission. Joint Research Centre. and European Commission. Directorate General for Internal Market, Industry, Entrepreneurship and SMEs. *Annual Report on European SMEs 2024/2025: SME Performance Review 2024/2025*. LU: Publications Office, 2025. DOI: 10.2760/7714438.
- [13] Lukas R. G. Fitz et al. "aiSnoop: An AI-assisted Social Listening Tool for SME Managers". In: *AMCIS 2025 Proceedings*. Montreal, Canada, 2025.
- [14] Green Urbaczewski, Angela and Urbaczewski, Andrew. "The Use of Artificial Intelligence to Accelerate Small and Medium Sized Business Customer Journey Development". In: *Drake Management Review* (2024).
- [15] Erwin Halim et al. "Optimizing Artificial Intelligence (AI) Chatbot Customer Service in Small and Medium Enterprises (SMEs) in E-Marketplace". In: *2025 19th International Conference on Ubiquitous Information Management and Communication (IMCOM)*. 2025, pp. 1–6. DOI: 10.1109/IMCOM64595.2025.10857522.
- [16] Alexander A. Hernandez et al. "Predicting the Use Behavior on Conversational Artificial Intelligence Enabled Online Shopping in Small and Medium Enterprises: A Structural Equation Modeling Approach". In: *2023 IEEE 11th Conference on Systems, Process & Control (ICSPC)*. 2023, pp. 270–275. DOI: 10.1109/ICSPC59664.2023.10420022.
- [17] Chunli Huang. "Research on Marketing Strategy Innovation of Xi'an Small and Medium-sized Private Enterprises in the Era of New Media and Artificial Intelligence". In: *Proceedings of the 2019 3rd International Conference on Education, Economics and Management Research*. Singapore: Atlantis Press, 2019. ISBN: 978-94-6252-871-0. DOI: 10.2991/assehr.k.191221.113.
- [18] Toluwalase Vanessa Iyelolu et al. "Improving Customer Engagement and CRM for SMEs with AI-Driven Solutions and Future Enhancements". In: *International Journal Of Engineering Research And Development* 20.8 (2024).
- [19] Nir Kshetri et al. "Harnessing Generative Artificial Intelligence: A Game-Changer for Small and Medium Enterprises". In: *IT Professional* 26.6 (2024), pp. 84–89. DOI: 10.1109/MITP.2024.3501552.
- [20] Katherine N. Lemon and Peter C. Verhoef. "Understanding Customer Experience Throughout the Customer Journey". In: *Journal of Marketing* 80.6 (2016), pp. 69–96. DOI: 10.1509/jm.15.0420.
- [21] Ihab K. Magableh et al. "The Role of Marketing Artificial Intelligence in Enhancing Sustainable Financial Performance of Medium-Sized Enterprises Through Customer Engagement and Data-Driven Decision-Making". In: *Sustainability* 16.24 (2024), p. 11279. DOI: 10.3390/su162411279.
- [22] Maisha Morshed, Samiha Mahbub, and Jarin Akber. "Application of Artificial Intelligence (AI) in Small and Medium Enterprise (SME) and Its Future". In: *Global Scientific Journals* 12.11 (2024). ISSN: 2320-9186.
- [23] A. Malgeldiyev and R. Bibassarova. "Improving the Competitiveness of SMEs in the Republic of Kazakhstan with the Use of Artificial Intelligence Technologies". In: *Progress in Science* 5 (2024).
- [24] Michela Matarazzo and Alfredo Celentano. "The Role of Artificial Intelligence in SMEs: Insights from a Sys-

- tematic Literature Review”. In: EURAM Conference. Florenz, Italy, 2025. ISBN: 978-2-9602195-7-9.
- [25] Nnenna Ijeoma Okeke et al. “AI-driven Personalization Framework for SMES: Revolutionizing Customer Engagement and Retention”. In: *World Journal of Advanced Research and Reviews* 24.1 (2024), pp. 2019–2035. DOI: 10.30574/wjarr.2024.24.1.3208.
- [26] Olatunbosun Onaopemipo Olalekan. “The Role of AI in Optimizing Decision-Making for Small and Medium Enterprises (SMEs) in Nigeria’s Retail Industry”. In: *International Journal of Management & Entrepreneurship Research* 7.4 (2025), pp. 328–340. DOI: 10.51594/ijmer.v7i4.1883.
- [27] Jonas Polfuß. “Navigating the AI Hype: Perceptions and Opportunities in German SME Marketing Applications”. In: *Digital Management and Artificial Intelligence*. International Scientific-Practical Conference. Springer Proceedings in Business and Economics (SPBE). 2025. DOI: 10.1007/978-3-031-88052-0_32.
- [28] Luoxi Pu et al. “Investigating How Chinese E-Commerce SMEs Would Adopt AI Live-Streaming via a Dual Analysis: Trendy and Transformative”. In: *Journal of Organizational and End User Computing (JOEUC)* 36.1 (2024), p. 22. DOI: 10.4018/JOEUC.349934.
- [29] Omar Hasan Salah and Mohannad Moufeed Ayyash. “E-Commerce Adoption by SMEs and Its Effect on Marketing Performance: An Extended of TOE Framework with Ai Integration, Innovation Culture, and Customer Tech-Savviness”. In: *Journal of Open Innovation: Technology, Market, and Complexity* 10.1 (2024), p. 100183. DOI: 10.1016/j.joitmc.2023.100183.
- [30] Wina Permana Sari et al. “User Experience (UX) Analysis of Skin Tone Clustering with AI for Clothing Small and Medium Enterprise (SME) System”. In: *2025 3rd International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT)*. 2025, pp. 230–236. DOI: 10.1109/IDCIOT64235.2025.10914855.
- [31] Jakkrit Suttanurak et al. “Development of Artificial Intelligence-Based Application Promoting SMCE and SME for Thailand Organic Valley”. In: *Suan Sunandha Science and Technology Journal (SSSTJ)* 9.1 (2022), pp. 17–24. ISSN: 2539-5742.
- [32] Akhmad Syaifuddin and Luthfia Qurrota A’yun. “Empowering SMEs: Implementing Generative AI for Enhanced Copywriting in Bantul’s Kusuka Ubiku”. In: *Society : Jurnal Pengabdian Masyarakat* 4.1 (2025), pp. 84–90. DOI: 10.55824/jpm.v4i1.518.
- [33] Nana Usman and Ibrahim Yusuf. “Artificial Intelligence and Enterprise Performance: Investigating the Effect of Chatbots and Virtual Assistants on Smes’ Customer Retention in Fct Abuja”. In: *Abuja Journal of Business and Management* 3.2 (2025). DOI: 10.7118/80swcm16.
- [34] Andyan Utama et al. “Customer Citizenship Behavior on Business Performance: Artificial Intelligence as Moderation in Smes”. In: *Jurnal Riset Bisnis dan Manajemen* 18.1 (2025), pp. 53–69. DOI: 10.23969/jrbm.v18i1.17383.
- [35] Uday Shankar Verma, Sudawan Somjai, and Ananta Rusmee. “Readiness of Small and Medium-Sized Enterprises in the Bangkok Metropolitan Area to Integrate AI into Their Marketing Strategies”. In: *International Journal of Development Administration Research* 7.2 (2024), pp. 168–182. ISSN: 2730-3063.
- [36] Jan vom Brocke et al. “Reconstructing the Giant: On the Importance of Rigour in Documenting the Literature Search Process”. In: *Proceedings of the European Conference on Information Systems*. Verona, Italy, 2009.
- [37] Vuong Gia Ky Nam et al. “The Impact of AI Chatbot and Brand Personality on Gen Z Consumer Behavior in Vietnamese SMEs”. In: *Journal of Development and Integration* (2025). ISSN: 2815-6234.
- [38] Wagobera Edgar Kedi et al. “AI Chatbot Integration in SME Marketing Platforms: Improving Customer Interaction and Service Efficiency”. In: *International Journal of Management & Entrepreneurship Research* 6.7 (2024), pp. 2332–2341. DOI: 10.51594/ijmer.v6i7.1327.
- [39] Fabian Walke, Lee Klopfers, and Till J. Winkler. “Leveraging Generative AI and ChatGPT in SMEs: A Grounded Model”. In: *Proceedings of the 58th Hawaii International Conference on System Sciences (HICSS)*. 2025. DOI: 10.24251/HICSS.2025.658. HDL: 10125/109506.
- [40] Li Xiyang and T. Ramayah. “Exploring Cross-Border E-Commerce Development Strategies for Chinese SMES in the Age of Generative AI”. In: *International Journal of Academic Research in Business and Social Sciences* 14.11 (2024), Pages 2629–2642. DOI: 10.6007/IJARBS/v14-i11/23435.
- [41] Ibrahim Yusuf et al. “Ai-Driven Dynamic Content Creation and Micro-Small and Medium Enterprises Market Performance in the Federal Capital Territory, Abuja”. In: *Abuja Journal of Business and Management* 3.2 (2025). DOI: 10.7118/ads2sb69.
- [42] Xiaohong Zhang. “Brand Product Recognition Model of Small and Medium-Sized Enterprises Based on AI Algorithm”. In: *2023 IEEE 6th International Conference on Knowledge Innovation and Invention (ICKII)*. 2023, pp. 273–276. DOI: 10.1109/ICKII58656.2023.10332678.
- [43] Yijie Zhong. “E-Commerce Utilization Analysis and Growth Strategy for Smes Using an Artificial Intelligence”. In: *Journal of Intelligent & Fuzzy Systems* 45.5 (2023), pp. 7619–7629. DOI: 10.3233/JIFS-232406.
- [44] Zupan Zong et al. “Big-Data AI Analytics in Value-Chain Innovation and International Marketing Strategy: Insights from SMEs in Cultural and Creative Industries”. In: *International Marketing Review* (2025). DOI: 10.1108/IMR-02-2024-0049.

Discovering skill requirements using AI: DataGEMS, an EU Horizon Project

Shobhit Agarwal
Wirtschaftsinformatik

Baden-Wuerttemberg Cooperative State University Stuttgart
Stuttgart, Germany
shobhit.agarwal@dhbw-stuttgart.de

Friedemann Schwenkreis
Wirtschaftsinformatik

Baden-Wuerttemberg Cooperative State University Stuttgart
Stuttgart, Germany
Friedemann.Schwenkreis@dhbw-stuttgart.de

Abstract—This paper introduces DataGEMS, an EU Horizon-funded research project focused on data exploration and data discovery. The platform introduces various AI-driven approaches to bridge the gap between data and data-based question-answering systems. DataGEMS is an ongoing project and currently supports four use cases. In this paper, we detail the lifelong learning use case, which compares individuals’ and organizational skill models with reference skill models to identify skill gaps and recommend targeted educational offerings to close them. The use case utilizes natural language questions and a structured set of comparisons to calculate proficiency deltas and match the missing competencies with appropriate educational programs. Additionally, data from diverse sources, including skill taxonomies (ESCO, O*NET), organizational skill models, and educational offerings, support DataGEMS’s data-driven nature. The work presented in this paper highlights skill modeling research and provides a practical AI-driven framework for HR professionals and individuals to navigate continuous up/re skilling requirements driven by automation, AI adoption, and demographic workforce changes.

Index Terms—Lifelong learning, AI, Skill management

I. INTRODUCTION

A. Background and Motivation

Rapid technological advancement, demographic change, and economic restructuring are fundamentally transforming labor markets and skill requirements worldwide. A significant body of research shows that digital innovation and automation have influenced and altered job task compositions by reducing the demand for routine work in certain occupations, while increasing the demand for analytical problem-solving and interpersonal skills, [1], [2], and [3].

Furthermore, the introduction of artificial intelligence (AI) in recent times has further intensified this transformation of skill requirements. Empirical studies demonstrate that AI complements human labor in some tasks while substituting for it in others, leading to shifting skill requirements rather than uniform job destruction [4]. In particular, AI has created a demand for skills that enable effective human-machine interaction, such as digital literacy, adaptability, and collaborative problem-solving [5]. These trends place sustained pressure on individuals and organizations to continuously acquire and update skills. Societal and demographic factors compound these technological pressures. Workforce aging, increased labor market mobility, and changing career trajectories challenge

traditional education models that focus on learning in early life stages [6].

Research in human resource development emphasizes that lifelong learning, defined as continuous, self-directed, and context-sensitive learning across the lifespan, is critical for maintaining employability and organizational resilience [7]. However, participation in lifelong learning remains uneven, often constrained by limited access to relevant learning opportunities and insufficient alignment between training provision and evolving skill demands [8]. From an organizational perspective, workforce up-skilling is further complicated by the lack of integrated, data-driven approaches for understanding skill evolution. As a result, organizations struggle to translate external technological and economic signals into actionable insights for workforce development. The researcher in [9] proposes a transdisciplinary approach to managing transitions in skill requirements, offering practical strategies for organizations to foster lifelong learning while ensuring fairness and inclusion across diverse demographics.

Together, these societal, technological, and economic dynamics underscore the need for data-centric, AI-supported lifelong learning systems. Such systems must enable continuous skill assessment, anticipate emerging skill needs, and support evidence-based up-skilling strategies at both individual and organizational levels.

B. The DataGEMS Platform

DataGEMS^{1 2} is a research and innovation action funded by the European Union under the Horizon Europe Research and Innovation Programme. DataGEMS provides tools and mechanisms, powered by AI and data-centric approaches, to address the challenges outlined in the previous section.

DataGEMS is a data discovery platform with Generalized Exploratory, Management, and Search capabilities. It is built on the principles of data FAIRness, openness, and re-use. It aims to seamlessly integrate data sharing, discovery, and analysis into a system that addresses the whole data lifecycle, i.e., sharing, storing, managing, discovering, analyzing, and reusing (data and/or metadata), bridging the gap between the

¹DataGEMS project has received funding from the European Union’s Horizon Europe framework programme under grant agreement No. 10118841

²<https://datagems.eu/>

data provider and the data consumer. DataGEMS is a next-generation data discovery and management ecosystem that engulfs different types of data (structured, unstructured, real-time and historical) and enables users to (a) enrich data through powerful data profiling mechanisms (b) seamlessly discover and analyze data across and within datasets using user-intuitive discovery and analysis mechanisms, such as using natural language and patterns, and (c) effectively explore and combine data with the help of stepwise guidance mechanisms during dataset discovery and analysis. The effective and efficient functioning of these mechanisms will be enabled by a data and model management layer that decouples low-level data management from higher-level data analytics.

C. Lifelong Learning as a Use Case

Individuals and organizations are constantly challenged to adapt continuously to remain effective and competitive. Given this backdrop, lifelong learning is no longer an individual-centric concept but an organizational requirement, supported by structured models, reliable data, and analytical mechanisms.

DataGEMS will be initially tested and deployed to benefit diverse user communities and user types across core domains: education, meteorology, and language data infrastructures, where the lifelong learning use case is part of the education domain. As one of the use cases for the DataGEMS platform, lifelong learning addresses the aforementioned challenges by analyzing individual and organizational skill models against future skill demands and thereby suggesting appropriate educational offerings to close the skill gap. The underlying idea is to treat skills not as static but as dynamic entities that can be influenced by external events such as technological innovation, legal changes, and shifts in public perception. The overall vision is to enable organizations to better anticipate their skill needs, identify emerging skill gaps, and provide educational solutions to meet these requirements.

The use case builds on the assumption that explicit skill models created using skill management tools represent the skills an individual possesses (given a specific occupation or context). A collection of individual skill models can represent a team or, at a higher level, any organizational unit. Furthermore, these skill models can be reasoned with, i.e., we can perform calculations to get a skill delta, etc. These skill models serve as the foundation for comparison across time, organizational units, and reference frameworks. Some of the key requirements for the comparison include the ability to (i) assess external changes that may affect skill relevance, (ii) assess the impact of such changes on existing skill models, (iii) identify mismatches between current and target skill sets, and (iv) relate identified skill gaps to suitable educational offerings. To achieve the required goals, the use case requires interoperable datasets, such as skill models, skill forecasts, and metadata describing learning opportunities, as well as mechanisms for comparing these models over time. Another key requirement related to skill analysis is the explainability. A stakeholder may not be interested in binary answers to important questions, such as whether certain skills are sufficient or obsolete, but

will also be interested in the rationale behind it, which events, trends, or reference models lead to a particular assessment.

D. Problem Statement and Contributions

The expected outcome of the project and the use case is a comprehensive analytical framework to support human resource planning and workforce development. We envision linking skill model evolution, gap analysis, skill model algebra, and educational recommendations. The use case aims to provide actionable insights for a diverse set of stakeholders, including human resource executives, skill development experts, and individuals. In the process, we contribute to ongoing research on skill modeling, skill gap detection, and AI approaches to lifelong learning while implementing a tool to support continuous up/re-skilling.

II. RELATED WORK

The following section briefly summarizes related work on developing a data-driven learning and knowledge exploration system. Such a system would be at the intersection of multiple domains, including data discovery, question-answering systems, competency-based learning, and adaptive educational technologies. This section reviews the foundational work in these areas.

A. Data Discovery and Knowledge Exploration

The work on data discovery and exploratory data analysis has been central to database research and human-computer interaction; some of the earliest work includes [10], which introduced dynamic queries and visual exploration tools that enabled users to explore large datasets interactively. Furthermore, research has examined reducing the cognitive load of data exploration through intelligent recommendation and automation systems. The authors in [11] developed a mixed-initiative system that combines manual specification with automated chart recommendations for exploratory visual analysis.

Similarly, the AIDE framework proposed by [12], an active-learning-based approach, enables users to explore data interactively and to automatically generate and refine query recommendations. This allows users to iteratively narrow down on their analytical request without requiring an explicit query formulation. The authors in [13] demonstrated that natural language conversational interaction can lower barriers to data analysis for non-technical users. This would enable non-technical users through incremental dialog-driven query refinement. Furthermore, systems like Eviza, presented by the authors [14] at ACM CHI, leverage large language models (LLMs) to enable exploratory question-answering over tabular data. This approach would bridge the gap between structured databases and users asking unstructured questions. They also highlighted the possible use of a conversational paradigm to ease data access. The authors also assume that users possess sufficient domain knowledge to formulate meaningful queries.

B. AI-Based and Data-Rooted Questioning Systems

Question-answering (QA) systems have evolved significantly, from their early template-based approaches to current LLMs, as they have moved from structured and semi-structured data. The authors in [15] developed semantic parsing techniques that map natural-language questions to executable logical forms over knowledge bases, enabling the system to answer complex queries over structured data sources such as Freebase and DBpedia.

The authors in [16] continued the work on semantic parsing for semi-structured tables by introducing the WikiTableQuestions dataset and a logical-form parsing algorithm that can answer complex questions over Wikipedia tables, addressing the dual challenges of broader domain coverage and deep logical computation. More recently, hybrid approaches have combined neural generation with non-parametric retrieval. [17] Introduced Retrieval-Augmented Generation (RAG) at NeurIPS, a framework that uses pre-trained parametric memory with a dense non-parametric retrieval index; this process significantly improves the performance on information-intensive question answering tasks. The interface for QA systems introduces additional complexity by requiring context tracking, referencing resolution, and a way to clarify the query. [18] developed and proposed a conversational QA system that maintains dialogue memory across turns, which enables users to ask follow-up questions without rephrasing the context.

Similarly, there have been hybrid approaches that combine retrieval from unstructured text with querying structured knowledge bases. [19], who proposed GRAFT-Net, a graph neural network model that fuses knowledge base entities and text passages into a unified question-specific subgraph to answer open-domain questions. Although there have been advancements in existing systems, QA systems face various challenges. Particularly in reasoning and representation of structured domain knowledge, as highlighted by [20] in their exhaustive review of knowledge graphs (KDs), their history, and their role in organizing world knowledge for AI systems

C. Skill-Based Learning and Competency-Driven Platforms

The landscape of educational technology has witnessed a monumental shift from content-centric to skill-centric learning models. The authors in [21] laid the foundational work for evidence-centered design for learning, proposing a framework that links observable student behaviors to underlying competencies through probabilistic reasoning. Further building on this, competency-based education (CBE) models prioritize the mastery of specific skills, enabling personalized learning pathways, as surveyed by [22] in the Journal of Competency-Based Education. Additionally, [23] proposed an ontology-based competency framework in the Journal of Interactive Learning Research that formalized the relationships among tasks, skills, and learning resources, facilitating advanced content recommendation and gap analysis. There has been significant research on competency frameworks and skill taxonomies across domains. The European e-Competence Framework (e-CF), developed by the CEN Workshop on ICT Skills [24],

and O*NET [25] in the United States, provides standardized taxonomies of professional skills.

Jovanovic et al. [26] used learning analytics to model a learner's competencies dynamically. The system used trace data from the interactions with the digital platform to infer skill levels and predict performance. Furthermore, Khosravi et al. [27] published research on an adaptive learning system with intelligent feedback mechanisms that enable real-time personalization based on learners' engagement patterns and performance.

Positioning relative to prior work. The skill gap algebra introduced here builds on earlier studies but remains distinct from prior competency-related models. For example, Paquette's ontology-based framework [23] standardises professional and educational competencies digitally. In this framework, skills are modeled as ontological entities, supporting reasoning over them. This ontology-driven architecture enables personalised lifelong learning. Our approach differs by introducing gap operators as a first step towards an algebra defined for skill models. Our main contributions: (i) taxonomy-agnostic projection operators for extracting and normalising skills, (ii) proficiency-aware skill gap operators ($\sqcap$ and $\sqcap^-$) that distinguish missing from under-proficient skills, and (iii) combining identified gaps with educational offerings via a natural-language querying pipeline. DataGEMS makes gap calculation algebraically explicit and auditable, thereby supporting explainability.

D. Data Sources for Lifelong Learning Systems

Lifelong learning systems require diverse data sources to model learners' competencies (skill levels), recommend learning pathways, and align educational offerings with labor market demands. Skill models and competency frameworks, such as those derived from ESCO (European Skills, Competences, Qualifications and Occupations) [28] and [29], analyzed skill taxonomies from job boards such as LinkedIn and Glassdoor to provide structured representations of skill requirements. These frameworks, among others, help design and enable systems to map a learner's profile to the requirements for a job and identify gaps. Janssen et al. [30] explored the use of organizational learning data, including performance reviews, project assignments, and peer feedback, to identify skills and competencies that a formal assessment may not capture.

The data generated and captured through learners' interactions with learning management systems, MOOCs, and workplace platforms offer a great opportunity for personalized recommendations. Pardos and Heffernan [31] demonstrated that fine-grained clickstream data can be used to reveal learning patterns through User Modeling and User-Adapted Interaction. Frey and Osborne [32] published their seminal work on technological forecasting and social change regarding the future of employment, highlighting the skills at risk of automation. Further, studies by Deming [33] in the Quarterly Journal of Economics and the World Economic Forum's Future of Jobs Report [34] track emerging competencies in areas such as AI literacy, data science, and human-centered design.

Additionally, integrating such skill forecast data into learning systems will enable proactive skill development, as de Fareri et al. [35] demonstrated in their study on AI-driven career guidance. The work we carry out as part of DataGEMS unifies these conversational frameworks by leveraging skill models. Working with these diverse research streams and data sources, DataGEMS aims to address critical gaps at the intersection, enable data exploration, and build a QA system bounded by competencies to users’ learning requirements.

III. METHODOLOGY

A. Design Principles and Assumptions

The DataGEMS platform comprises multiple components. For each component, a boundary has been defined with clear inputs and outputs. Given that the platform is distributed, spanning multiple contributors and partners, the platform components are designed for loose coupling. The platform is designed using a microservice architecture paradigm.

The Model Context Protocol (MCP) [36] is a universal protocol that standardizes the definition, discovery, and structured interaction of AI models, particularly LLMs, with external tools and execution environments. In the proposed architecture for DataGEMS, the MCP servers provide a registry of tools with their standard description and specify each tool’s purpose, expected inputs, and outputs. On the client side, the MCP is responsible for correctly reading the registry and invoking the appropriate tools based on the task at hand. Within the architecture of the DataGEMS platform, MCP will be used to expose a set of tools that provide access and data preprocessing capabilities. These tools can be invoked by the components based on user queries. Another core idea of the architecture is the MoMA (Model of Metadata and data) logical model. It is a graph-based data model that represents datasets, metadata, machine-learned models, and all the operations that can be performed on the data. Moma enables the platform to handle heterogeneous datasets (e.g., text, images, sensor data, and structured SQL, NoSQL, and graph data).

B. Conceptual Architecture of the DataGEMS Platform

The primary software deliverable of the DataGEMS project is the DataGEMS platform, which will also be part of the EOSC³ infrastructure. DataGEMS will support users in publishing, discovering, and using datasets across diverse domains. As mentioned earlier, the platform has a microservices architecture, where all components expose REST API endpoints to ensure scalability, modularity, and ease of integration.

As illustrated in the logical architecture overview in Figure 1, the platform’s coarse-grained structure reflects its core responsibilities and interaction model. There are two primary user roles for the platform: The consumer, who searches for datasets to answer dataset exploration and discovery questions, or to support data-driven tasks. The second group of users is the providers, who contribute datasets to the platform. Within the DataGEMS project, the key providers also originate from

the use cases. The platform currently supports the following use cases: Mathematics, Weather, Language, and Lifelong Learning (DHBW). Furthermore, there are two main entry points for the DataGEMS platform: the DataGEMS UI and the DataGEMS API. The DataGEMS UI provides a single-page application that offers a cohesive, interactive experience. This interface enables key platform functionalities, including dataset ingestion, search, exploration, and metadata enrichment. All UI interactions are mediated by the DataGEMS Gateway, which serves as the platform’s central point of interaction.

In addition, the DataGEMS platform offers developers and third-party integrators programmatic access via its DataGEMS API, which the DataGEMS Gateway exposes. The DataGEMS Gateway authenticates users via the DataGEMS AAI, enforces access control policies, and orchestrates the user’s journey through the platform’s services. Behind the gateway, the platform is composed of a set of semantically grouped microservices, organized into four main categories:

- Dataset Profiling: analyzing and extracting metadata from contributed datasets and linking them with each other.
- Dataset Discovery: discover datasets, ask questions on data, and produce datasets.
- User Assistance: assist the user via recommendations of datasets and queries, provenance of data discovery actions.
- Data Storing & Management: handling storage, metadata persistence, and data governance.

Finally, we have the data set layer, which stores all data and metadata. In Figure 1, all microservices can access the datasets directly if required, as per their requirements.

C. Separation of data ingestion, processing, and User interface

At the core of the DataGEMS platform is the data and management layer. This layer is fundamental to realizing the platform’s vision of seamless data discovery and analysis. It achieves this by decoupling low-level data management, responsible for storing and brokering data and machine learning models, from higher-level analytics and discovery functionalities. By decoupling the LLM’s prompts and logic from the tool’s implementation details, MCP simplifies integration and reduces the need for custom, tightly coupled solutions. Through this architectural separation, the platform can interconnect the components that are responsible for:

- Enriching dataset metadata through data profiling and quality monitoring.
- Cross-dataset and in-dataset discovery operators powered by ad-hoc indexes as well as stored machine learning models.
- Personalized support of data exploration workflows via query and dataset recommender systems.
- Polyglot querying capabilities that unify access to heterogeneous data stored in files, databases, and real-time streams.

³<https://eosc.eu/>

- Storing, cataloging, and serving machine learning models and their connections to datasets.

The Data Storage, Management, and Access components form the technical backbone on which various components are implemented to provide these functionalities.

The DataGEMS architecture is designed to be general in the functionality it offers, and the components provide use-case-agnostic functionality. The combination of domain-specific datasets, any required pre-processing steps, and use-case-specific interfacing components enables the general DataGEMS platform to deliver tailored functionality that addresses use-case-specific requirements. Outside of these domains, the DataGEMS operators are designed to be extensible and composable, enabling the onboarding, discovery, and exploration of data from different use cases.

The User interface component is still under review; however, the current version is illustrated in Figure 2 where the user can explore the stored datasets and ask questions directly to the DataGEMS platform.

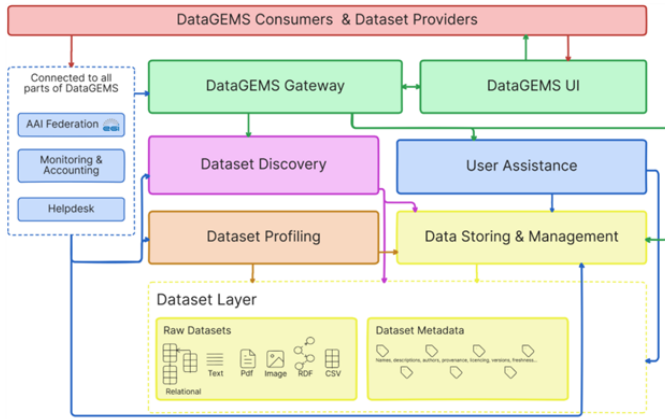


Fig. 1. DataGEMS architecture overview (image retrieved from internal DataGEMS project documentation).

IV. CASE STUDY (LIFELONG LEARNING)

A. Workforce Up-skilling Challenges and Skill Models

The unprecedented shift in technology, particularly the adoption of automation and AI, is continuously reshaping workforce requirements, meaning that a significant portion of the current workforce needs re-skilling to remain employable [37]. Furthermore, there are organizational hurdles, such as inadequate funding and resistance to change, which often undermine upskilling and reskilling initiatives [38]. Studies also show that current training programs may lack relevance to current labor market requirements, which, in turn, causes a persistent skills shortage [39]. Furthermore, misalignment between educational curricula and industry demands continues to challenge comprehensive up/re-skilling strategies [40]. To address these issues, the lifelong learning use case as part of the DataGEMS project, has devised the following strategy, where an individual can submit their skill model and get an updated list of education offerings that can fulfill the missing

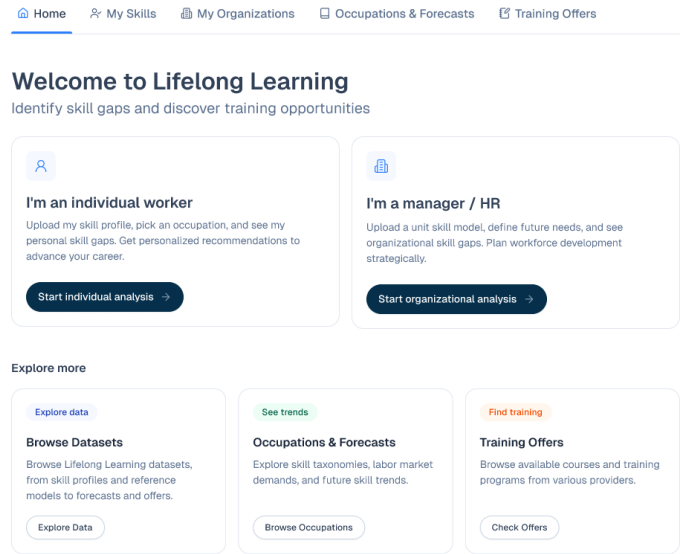


Fig. 2. Lifelong learning User interface (image retrieved from internal DataGEMS project documentation).

skill gaps or organizational head trying to up/re-skill their work force could upload an organizational skill model and receive a detailed report to address the organizational skill gap. The following sections go into detail about the individual scenario as an example.

The European Skills, Competences, Qualifications and Occupations organization (ESCO) [41] defines:

- Skill: Is the ability to apply knowledge and use know-how to complete tasks and solve problems.
- Competence: Is the proven ability to use knowledge, skills, and personal, social, and/or methodological abilities, in work or study situations and in professional and personal development.
- Occupation: is a grouping of jobs involving similar tasks and which require a similar skills set.

Furthermore, a skill is usually not a single entity or ability. It is a multifaceted construct with several components. The authors in [42] take a very comprehensive view of the definition of a skill. It has a proficiency level, which is achieved through practice and repetition. A skill is not only a technical ability; it also involves cognitive and behavioral components. This comprehensive definition motivates us to explore further and understand skill gaps and how to mitigate them. Further definitions relevant for the case study:

- Skills taxonomy: is a structured list of skills.
- Skill model/profile: is a subset of skills (from a skill taxonomy) that an individual possesses, defined in a structured format.
- Reference skill model: is an up-to-date list of skills for any given job, derived from a specific taxonomy,

The Figure 3 depicts a custom skill model prepared using the 'myskillsplus' tool [43]. and Figure 4 refers to the skill requirements (reference model) for a data scientist according

to the O*NET taxonomy [25]. Both images contain only a part of the entire skill model.

Tools, Technologies, Applications

Edge AI and On-Device AI SDKs	Basic/Learning	
Computer Vision Tools	Expert/In-depth	
Computer Vision and Image Processing Libraries	Expert/In-depth	
Time-Series Analysis and Forecasting Tools	Proficient/High Level Knowledge	
Quantum AI and Quantum Machine Learning Tools	Basic/Learning	
Graph Neural Networks (GNN) Frameworks	Familiar/Low Level Knowledge	
Reinforcement Learning Tools	Familiar/Low Level Knowledge	
AI/ML Frameworks and Libraries	Proficient/High Level Knowledge	
Prompt Engineering Tools and IDEs	Familiar/Low Level Knowledge	

Fig. 3. Snapshot of a skill model for a data scientist [43]. On the left-hand side are the skills, and a proficiency level is attached to each.

Technology Skills

15-2051.00 - Data Scientists

Section	Category	Example
Technology	Analytical or scientific software	IBM SPSS Statistics
Technology	Analytical or scientific software	SAS
Technology	Analytical or scientific software	TensorFlow
Technology	Analytical or scientific software	The MathWorks MATLAB
Technology	Application server software	Docker
Technology	Application server software	GitHub
Technology	Application server software	Kubernetes
Technology	Business intelligence and data analysis software	Alteryx software
Technology	Business intelligence and data analysis software	Apache Spark
Technology	Business intelligence and data analysis software	Microsoft Power BI
Technology	Business intelligence and data analysis software	Tableau
Technology	Cloud-based management software	Amazon Web Services AWS SageMaker
Technology	Cloud-based management software	Google Cloud software
Technology	Content workflow software	Atlassian JIRA

Fig. 4. Snapshot of a reference skill model for a data scientist according to O*NET [25].

B. Use Case Setup and Data

The use case is intended primarily for hiring managers and human resource professionals. Hence, the underlying assumption is that platform users will have access to skill management tools (e.g., Muschskills⁴) for creating skill models. The Figure 5 depicts the process of creating a skill model for an individual. Furthermore, a collection of such individual skill models (for each employee) would create an organizational skill model.

Another important pillar for the use case is educational offerings, which help fill skill gaps. Apart from skill models (individual, organizational, and reference), educational offerings are the other essential data of the use case. An educational

⁴<https://www.muchskills.com/>

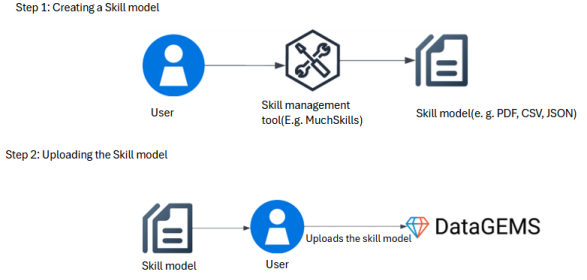


Fig. 5. Skill model creation process.

offering may consist of information about an institution's courses.

C. User Interaction and Natural Language-Based Querying

This section details the user's interaction with the DataGEMS platform. As mentioned previously, the assumption is that the user possesses their skill model. So, the required input data are the skill model, Target Occupation (the occupations the user wants to fulfill skills for), and a set of reference models. The user then asks a question in natural language; Figure 6 explains the process. Since natural language questions can be ambiguous, a disambiguation component is required to identify the user's intent. DataGEMS provides a query disambiguation mechanism that is triggered when the question is identified as ambiguous or too challenging to process as-is. There are other requirements, such as support for questioning in multiple languages, understanding the context of the questions, and understanding slang, etc. The query disambiguation also handles these. Once the question has been disambiguated and processed internally. The response is generated, containing the skill gaps for each available reference model. Furthermore, the user can continue to ask questions about ways to address the missing skills, and the expected output is a list of educational offerings that can be filtered, as shown in Figure 7. Lastly, the user can select the most convenient educational offering and, based on that selection, ask further questions about which skills each offering covers, as shown in Figure 8.

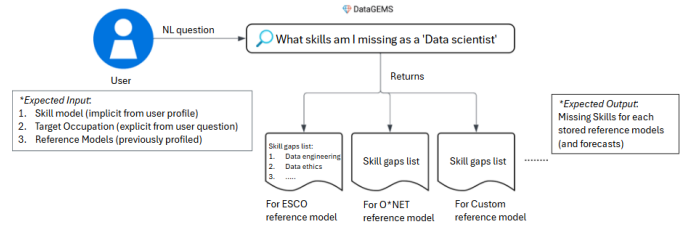


Fig. 6. The user provides their skill model, target occupation(s), select reference model and provide a natural language question

D. Analytical Logic

The core analytical mechanism behind the use case is a structured comparison between a reference skill model and

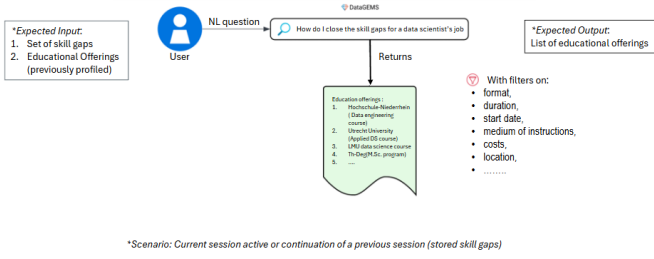


Fig. 7. The user can continue and ask how to fill the missing gaps.

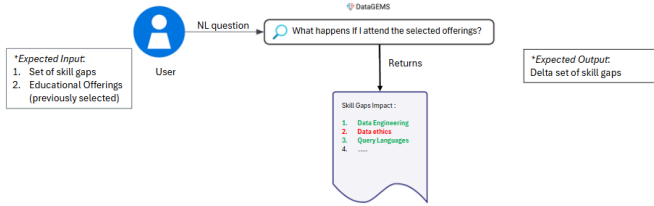


Fig. 8. The user can still refine the request by asking about each educational offering and the skills provided by it.

an individual skill profile. Since the user-provided model may originate from heterogeneous taxonomies, skill identifiers are first aligned to a common reference framework, and the further process is defined in the following sections.

E. Skill Set Definitions

Given a skill identifier 's' and a skill proficiency level ' $l(s)$ '. We define *skill proficiency sets* (skill models/profiles) as P as sets of pairs $(s, l(s))$:

$$P = \{(s_i, \ell(s_i))\}$$

In the following we will differentiate skill proficiency sets of individuals I and reference skill proficiency sets R . We assume that a skill identifier s is only part of a single pair $(s, l(s))$ in a skill proficiency set. The proficiency level associated with a skill identifier s in a skill proficiency set P is denoted by $\ell_p(s)$.

$$R = \{(s_i, \ell_R(s_i)) \mid s_i \text{ required for a target role}\}$$

$$I = \{(s_i, \ell_I(s_i)) \mid s_i \text{ possessed by an individual}\}$$

Furthermore, we define $\mathcal{S}$ as a set of skill identifiers and a skill projection operator π , which extracts the set of skill identifiers from a skill proficiency set

$$\pi(P) = \{\forall s_i \mid \exists (s_i, \ell(s_i)) \in P\}$$

Applying π to R and I yields the skill identifier sets $\mathcal{S}_R = \pi(R)$, $\mathcal{S}_I = \pi(I)$

F. Skill gap calculation

We define *basic skill gap intersection* operator $\sqcap$ that compares skill sets at the identifier level:

$$A \sqcap B = \{(s_i, \ell(s_i)) \mid s_i \in B \wedge s_i \notin A\}$$

We further define the *proficiency gap intersection* intersection that compares skill sets at both the identifier and proficiency level:

$$A \sqcap^- B = \{(s_i, \ell_B(s_i)) \mid s_i \in (S_A \cap S_B) \wedge \ell_A(s_i) < \ell_B(s_i)\}$$

which captures skills present in both sets, but where proficiency is below the required level.

1. Set-Level Skill Gap (Missing Skills)

Skills that are fully absent from the individual profile — neither the skill nor any proficiency level is present:

$$G_{\text{set}}(I, R) = I \sqcap R$$

That is, G_{set} contains all required skill-proficiency pairs $(s_i, \ell_R(s_i))$ for which no entry exists in I whatsoever.

2. Level-Based Skill Gap (Proficiency Below Requirement)

Skills present in the individual profile but whose proficiency does not meet the requirement:

$$G_{\text{level}}(I, R) = I \sqcap^- R$$

The skill exists in the individual's profile, but $\ell_R(s_i) > \ell_I(s_i)$ indicates the proficiency is below what the role demands.

3. Outdated Skills

Skills present in the individual's profile but not required by the target role neither the skill nor its associated proficiency level appears in R :

$$G_{\text{out}}(I, R) = R \sqcap I$$

A skill is considered outdated if at least one of the following conditions holds:

- 1) The skill is not required in the current reference model due to technological innovation and legal changes.
- 2) The skill has low or zero demand frequency across the organization's reference models.
- 3) The skill is marked as deprecated in the reference taxonomy (e.g., replaced by a successor skill).

Calculate the skill gap: Once skills and proficiency levels are aligned, the system computes the skill gap as the union of G_{set} and G_{level} . Furthermore, these calculations are limited to individuals and can be generalized to organizations.

G. Worked Example

To illustrate the formal definitions, consider a simplified scenario with three skills s_1 (Python p), s_2 (Machine Learning), and s_3 (Communication), where proficiency levels range from 1 (beginner) to 5 (expert). Let the reference skill set for a data scientist role R and an individual's profile I be:

$$R = \{(s_1, 4), (s_2, 3), (s_3, 2)\}$$

$$I = \{(s_1, 2), (s_3, 2)\}$$

Applying π yields $\mathcal{S}_R = \{s_1, s_2, s_3\}$ and $\mathcal{S}_I = \{s_1, s_3\}$.

Set-level gap (missing skills): $s_2 \notin \mathcal{S}_I$, so $G_{\text{set}}(I, R) = \{(s_2, 3)\}$.

Level-based gap (insufficient proficiency): $s_1 \in \mathcal{S}_R \cap \mathcal{S}_I$ and $\ell_R(s_1) = 4 > \ell_I(s_1) = 2$, so $G_{\text{level}}(I, R) = \{(s_1, 4)\}$.

Outdated skills: all skills in $\mathcal{S}_I$ are present in $\mathcal{S}_R$, so $G_{\text{out}}(I, R) = \emptyset$.

The overall skill gap is $G_{\text{set}} \cup G_{\text{level}} = \{(s_2, 3), (s_1, 4)\}$, which the system then uses to retrieve matching educational offerings.

H. Workflow Overview

The workflow begins by retrieving two skill models. The stored reference skill model, which is associated with an occupation or a role, as well as the user's individual skill model. And secondly, the educational offerings, further process is as follows:

- *Extraction of skill and proficiency levels*: from both the reference and individual skill models, the system extracts a list of skills and their associated proficiency levels.
- *Calculate the skill gap*: Once skills and proficiency levels are aligned, the system calculates the skill gap by comparing required levels against current levels for each skill
- *Match missing skills with training offerings*: For each identified gap, the system searches educational repositories and metadata sources to identify training offerings that address the missing or insufficient skills.
- *Generate a suitable list of education offerings*: Finally, the system compiles an appropriate list of educational offerings that can fill the gap. The list will support filters such as formats, duration, language, and cost.

V. LIMITATIONS AND FUTURE WORK

The DataGEMS platform and the lifelong learning use case are currently under development. Hence, certain limitations should be noted. The proposed framework and architecture will be tested in real-world scenarios. As a result, assertions around usability and effectiveness require further validation and empirical evaluation.

Additionally, the skill gap operators assume skill identifiers come from heterogeneous taxonomies such as ESCO, O*NET, and custom organizational models. These identifiers can be aligned into a common reference space. In practice, the alignment process can introduce errors and affect the accuracy of gap calculations.

Currently, proficiency levels are directly comparable as integers across taxonomies. However, proficiency levels could be defined differently. In that case, a rigorous cross-taxonomy normalization or mapping is necessary. Using alternative definitions may reduce the accuracy and reliability of the gap calculation.

Planned evaluation steps include deploying the platform within the EOSC infrastructure. Subsequently, the platform will be evaluated with DHBW's partner organizations. These activities will include multiple pilot studies. Specifically, key themes are: (i) a user study with HR professionals and individuals to assess usability for individual and organizational skill profiles; (ii) evaluation of the accuracy of retrieved information in simple and complex scenarios; and (iii) end-to-end assessment of recommended educational offerings. Together, these evaluations will validate the proposed approaches and provide empirical support for further research and platform development.

VI. CONCLUSION

There has been an accelerated pace of adoption of digital technologies such as AI and automation. This high-paced adoption requires both organizations and individuals to continuously reassess their skill portfolios. To address these challenges, the paper introduces the DataGEMS platform and the Lifelong learning use case. The DataGEMS platform's data-centric approach and AI-assisted nature are used to detect dynamic skill gaps and provide targeted educational recommendations. The proposed approach explores combining structured skill models, taxonomies, and educational offerings with unstructured natural-language questions for data-driven workforce planning. The structured set comparison and the calculation of proficiency delta/skill gap allow individuals and organizations to identify partial and full skill gaps. The skill gap calculations help prioritize human resource development needs and align with upskilling programs.

As the platform is powered by AI-supported natural language querying capabilities, the barrier to interacting with the system is lower, enabling intuitive interaction and the ability to ask complex questions. From an organizational perspective, the platform supports proactive workforce planning, strategic re-skilling and up-skilling, and improved allocation of training resources. For educational institutions and training providers, it enables listing educational offerings and promoting skill-aligned course descriptions, helping better align courses with labor market demands. Formally, the integration of data discovery, skill modeling, and AI-enabled search advances research at the intersection of data management, question-answering systems, and lifelong learning technologies.

ACKNOWLEDGMENT

DataGEMS is an ongoing research project under development. Any information presented in this paper should be treated as intermediate information and not the final report. The DataGEMS project has received funding from the European Union's Horizon Europe framework

AI-GENERATED CONTENT ACKNOWLEDGEMENT

We acknowledge the use of Grammarly (<https://www.grammarly.com/>) to correct grammatical errors and the academic tone of the work.

REFERENCES

- [1] C. Park, J. Kim. "Education, Skill Training, and Lifelong Learning in the Era of Technological Revolution," in SSRN Electronic Journal, 2020.
- [2] S. Zekan, P. Pietrzak, P. Palada. "The impact of digitalization on future leaders skills," in *Zbornik radova Veleučilišta u Šibeniku*, vol. 17, no. 3-4, pp. 25–35, 2023.
- [3] L. Haefner, R. Sternberg. "Spatial implications of digitization: State of the field and research agenda," in *Geography Compass*, vol. 14, no. 12, pp. e12544, 2020.
- [4] D. Acemoglu, P. Restrepo. "The Wrong Kind of AI? Artificial Intelligence and the Future of Labor Demand," National Bureau of Economic Research, Tech. Rep., 2019.
- [5] Elina Makela & Fabian Stephany, 2024. "Complement or substitute? How AI increases the demand for human skills," Papers 2412.19754, arXiv.org, revised Feb 2025.
- [6] Marois, G., et al. "Projecting Labour Market Imbalances and Skill Mismatch Under Demographic Change in the EU," in *European Journal of Population*, vol. 42, no. 1, pp. 4, 2026.
- [7] T. Schultheiss, U. Backes-Gellner. "Different degrees of skill obsolescence across hard and soft skills and the role of lifelong learning for labor market outcomes," in *Industrial Relations: A Journal of Economy and Society*, vol. 62, no. 3, pp. 257–287, 2023.
- [8] M. Günen, Ö. Vural. "LIFELONG LEARNING BARRIERS: A THEORETICAL RESEARCH," in *Journal of Multidisciplinary Studies in Education*, vol. 7, no. 4, pp. 164–180, 2023.
- [9] Morandini, S., et al. "The Impact of Artificial Intelligence on Workers' Skills: Upskilling and Reskilling in Organisations," in *Informing Science*, vol. 26, pp. 39-68, 2023.
- [10] C. Ahlberg and B. Shneiderman, "Visual information seeking: Tight coupling of dynamic query filters with starfield displays," in *Proc. ACM SIGCHI Conf. Human Factors in Computing Systems (CHI)*, Boston, MA, USA, 1994, pp. 313–317.
- [11] K. Wongsuphasawat, D. Moritz, A. Anand, J. Mackinlay, B. Howe, and J. Heer, "Voyager: Exploratory analysis via faceted browsing of visualization recommendations," *IEEE Trans. Vis. Comput. Graph.*, vol. 22, no. 1, pp. 649–658, Jan. 2016.
- [12] K. Dimitriadou, O. Papaemmanouil, and Y. Diao, "AIDE: An active learning-based approach for interactive data exploration," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 11, pp. 2842–2856, Nov. 2016.
- [13] R. Mitra, A. Narechania, A. Endert, and J. Stasko, "Facilitating conversational interaction in natural language interfaces for visualization," in *Proc. IEEE Visualization and Visual Analytics (VIS)*, Oklahoma City, OK, USA, 2022, pp. 6–10.
- [14] V. Setlur, S. E. Battersby, M. Tory, R. Gossweiler, and A. K. Chang, "Eviza: A natural language interface for visual analysis," in *Proc. ACM Symp. User Interface Software and Technology (UIST)*, Tokyo, Japan, 2016, pp. 365–377.
- [15] J. Berant, A. Chou, R. Frostig, and P. Liang, "Semantic parsing on Freebase from question-answer pairs," in *Proc. Conf. Empirical Methods in Natural Language Processing (EMNLP)*, Seattle, WA, USA, 2013, pp. 1533–1544.
- [16] P. Pasupat and P. Liang, "Compositional semantic parsing on semi-structured tables," in *Proc. 53rd Annu. Meeting Assoc. Computational Linguistics (ACL)*, Beijing, China, 2015, pp. 1470–1480.
- [17] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-augmented generation for knowledge-intensive NLP tasks," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, 2020, pp. 9459–9474.
- [18] P. Christmann, R. S. Roy, A. Abujabal, J. Singh, and G. Weikum, "Look before you hop: Conversational question answering over knowledge graphs using judicious context expansion," in *Proc. ACM Int. Conf. Information and Knowledge Management (CIKM)*, Beijing, China, 2019, pp. 729–738.
- [19] H. Sun, B. Dhingra, M. Zaheer, K. Mazaitis, R. Salakhutdinov, and W. W. Cohen, "Open domain question answering using early fusion of knowledge bases and text," in *Proc. Conf. Empirical Methods in Natural Language Processing (EMNLP)*, Brussels, Belgium, 2018, pp. 4231–4242.
- [20] V. K. Chaudhri, C. Baru, N. Chittar, X. L. Dong, M. Genesereth, J. Hendler, A. Kalyanpur, D. B. Lenat, J. Sequeda, D. Vrandečić, and K. Wang, "Knowledge graphs: Introduction, history and perspectives," *AI Mag.*, vol. 43, no. 1, pp. 17–29, Mar. 2022, doi: 10.1002/aaai.12033.
- [21] R. J. Mislevy, L. S. Steinberg, and R. G. Almond, "On the structure of educational assessments," *Meas.: Interdiscip. Res. Perspect.*, vol. 1, no. 1, pp. 3–62, 2003.
- [22] J. Gervais, "The operational definition of competency-based education," *J. Competency-Based Educ.*, vol. 1, no. 2, pp. 98–106, Jun. 2016.
- [23] G. Paquette, "An ontology and a software framework for competency modeling and management," *J. Educ. Technol. Soc.*, vol. 10, no. 3, pp. 1–21, 2007.
- [24] CEN Workshop on ICT Skills, "European e-Competence Framework 3.0: A Common European Framework for ICT Professionals in all Industry Sectors," European Committee for Standardization (CEN), Brussels, Belgium, Tech. Rep. CWA 16234-1, 2016.
- [25] "O*NET OnLine Help: The Database." O*NET OnLine, National Center for O*NET Development, www.onetonline.org/help/onet/database. Accessed 09 February 2026.
- [26] Jovanović, J., Gašević, D., Brooks, C., Devedžić, V., Hatala, M. (2007). LOCO-Analyst: A Tool for Raising Teachers' Awareness in Online Learning Environments. In: Duval, E., Klamma, R., Wolpers, M. (eds) Creating New Learning Experiences on a Global Scale. EC-TEL 2007.
- [27] H. Khosravi, S. Sadiq, and D. Gašević, "Development and adoption of an adaptive learning system: Reflections and lessons learned," in *Proc. 51st ACM Technical Symp. Computer Science Education (SIGCSE)*, Portland, OR, USA, 2020, pp. 58–64.
- [28] Mike Zhang, Kristian Nørgaard Jensen, Barbara Plank (2024). Dataset: ESCO. <https://doi.org/10.57702/n2bcncsv>
- [29] Bhola, A., Halder, K., Prasad, A., & Kan, M.Y. (2020). Retrieving Skills from Job Descriptions: A Language Model Based Extreme Multi-label Classification Framework. In Proceedings of the 28th International Conference on Computational Linguistics (pp. 5832–5842). International Committee on Computational Linguistics.
- [30] J. Janssen, F. Berlanga, H. Vogten, and R. Koper, "Towards a learning path specification," *Int. J. Continuing Eng. Educ. Life-Long Learn.*, vol. 18, no. 1, pp. 77–97, 2008.
- [31] Z. A. Pardos and N. T. Heffernan, "Modeling individualization in a Bayesian networks implementation of knowledge tracing," in *Proc. Int. Conf. User Modeling, Adaptation, and Personalization (UMAP)*, Big Island, HI, USA, 2010, pp. 255–266.
- [32] C. B. Frey and M. A. Osborne, "The future of employment: How susceptible are jobs to computerisation?" *Technol. Forecast. Soc. Change*, vol. 114, pp. 254–280, Jan. 2017.
- [33] D. J. Deming, "The growing importance of social skills in the labor market," *Q. J. Econ.*, vol. 132, no. 4, pp. 1593–1640, Nov. 2017.
- [34] World Economic Forum, "The Future of Jobs Report 2025," World Economic Forum, Geneva, Switzerland, Rep., 2025.
- [35] S. Fareri, N. Fantoni, F. Chiarello, E. Coli, and A. Barandoni, "Estimating industry 4.0 impact on job profiles and skills using text mining," *Comput. Hum. Behav.*, vol. 107, pp. 1–11, Jun. 2020.
- [36] Anthropic. 2024. Introducing the Model Context Protocol. <https://www.anthropic.com/news/model-context-protocol>.
- [37] Li, L. Reskilling and Upskilling the Future-ready Workforce for Industry 4.0 and Beyond. *Inf Syst Front* 26, 1697–1712 (2024). <https://doi.org/10.1007/s10796-022-10308-y>
- [38] Hasan, M., Haque, M. A., Nishat, S. S., & Hossain, M. M. (2024). Upskilling and reskilling in a rapidly changing job market: Strategies for organizations to maintain workforce agility and adaptability. *European Journal of Business and Management Research*, 9(6), Article 2502.
- [39] Prakash, A., & Devi, S. (2024). Upskilling and reskilling: Effective strategies to bridge skill gaps and tackle workforce shortages in Kerala's evolving digital landscape. *African Journal of Biomedical Research*, 27(4S), 7461–7468. <https://doi.org/10.53555/AJBR.v27i4S.5016>
- [40] Khan, A., Sethi, S., Mustafa, S. N., & Bibi, M. (2025). Bridging the skills in the age of AI gap: Strategies for upskilling and reskilling in higher education. *Dialogue Social Science Review*, 3(2), 894–909.
- [41] ESCOmedia, <https://esco.ec.europa.eu/en/about-esco/escopedia/escopedia> (accessed Feb. 10, 2026).

- [42] Lamri, J., & Lubart, T. (2023). Reconciling Hard Skills and Soft Skills in a Common Framework: The Generic Skills Component Approach. *Journal of Intelligence*, 11(6).
- [43] My Skills Plus, <https://apps.myskillsplus.com/>, (accessed Feb. 10, 2026).

Author Index

Agarwal, Shobhit	116
Antal, Margit	37
Bonnici West, Lorna	21
Buza, Krisztian	37
Debono, David	21
Deguarra, David	21
Delicata, Francis	21
Dix, Marcel	47
Ehlers, Ulf	93
Fang, Ziming	47
Föll, Patrick	106
Galea Vella, Marisa	21
Gatt, Alan	21
Gauci, Shirley Ann	21
Jerkovic, Ema	30
Kuhn, Marc	5
Ma, Yuliang	47
Mann, Julia Marcia	93
Montalto, Marco	21
Morozov, Andrey	47
Nanopoulos, Alexandros	37
Neumann, Madeleine	5
Niedergesäss, Markus	83
Pallmann, Florian	73
Pazer, Maximilian	10, 63
Rauen, Daniela	55
Reit, Vanessa	5
Scerri, Daren	21
Schwenkreis, Friedemann	116
Schütz, Franziska	73
Selinka, Sarah	5

Soni, Hardik	63
Steinigke, Tessa	106
Theuma, Lisa	21
Thirukketheeswaran, Sinu	5
Welsch, Giacomo	83
Zeitler, Max	10, 63

Keyword Index

AI	106, 116
AI Ethics	55
AI Judges	10
AI competence model	93
AI literacy	30
AI-driven Mobility Pricing	5
Artificial Intelligence Tools and Applications	63
Assistive Technologies	55
Autism Spectrum	55
Autonomous Systems	10
agile integration	93
artificial intelligence	21
 Benchmarks	 10
 CRM	 106
Causal Inference	83
Chatbots	106
Collective Communication	63
Customer Journey	106
competency extraction	73
curriculum analysis	73
curriculum development	93
 Data Engines	 10
Distributed Training	63
Document Ranking	37
Double ML	83
disciplinary enculturation	30
 ESCO classification	 73
Embedding-Based Retrieval	37
Evaluation Framework	10
Explainable AI	83
education technology	21
epistemic agency	30
 FVD	 10
framework	93
future skills	93
 GPU Synchronization	 63
Generative AI	10

Hubness	37
industrial AI	47
information extraction	73
knowledge graphs	73
Lifelong learning	116
label noise	47
large language models	73
large language models (LLMs)	47
ML model robustness	47
Marketing	106
micro-credentials	73
NCCL	63
Nearest-Neighbor Search	37
Participatory Design	55
Physical AI	10
professional identity	30
Retrieval-Augmented Generation	37
research ethics	21
robustness testing	47
SHAP	83
SME	106
Sales	106
Shared Autonomous Vehicles	5
Skill management	116
Social Equity	5
Systematic Literature Review	10
supervision	21
Trustworthy AI	55
test automation	47
Video World Models	10
virtual reality	21
Workplace Inclusion	55
writing pedagogy	30

Program Committee

Slim Abdennadher	German University in Cairo
Serviola Beqiraj	AI with Impact
Alexander Dück	DHBW Mannheim
Patrick Föll	DHBW Heilbronn
Eva Gengler	Friedrich-Alexander-Universität Erlangen-Nürnberg, FemAI - Center for Feminist Artificial Intelligence
Sabine Gillner	Axon Ivy
Beate Grawemeyer	London Knowledge Lab, Birkbeck College, University of Lon- don
Anke Hutzschenreuter	DHBW Heidenheim
Thomas Kessel	DHBW Stuttgart
Monika Kochanowski	DHBW Stuttgart
Falko Kötter	DHBW Stuttgart
Bozena Lamek-Creutz	DHBW Mannheim
Stojan Maleschlijski	SAP
Prof. Dr. Carsten Mueller	DHBW Mosbach
Aikaterini Nakou	DHBW Lörrach
Dennis Pfisterer	DHBW Mannheim
Christiane Plociennik	DFKI
Dirk Reichardt	DHBW Stuttgart
Jennifer Schoch	DHBW Karlsruhe
Stephan Schulz	DHBW Stuttgart
Carmen Somm	Labhard Medien
Marcus Strand	DHBW Karlsruhe
Philippe Thomas	Institut für Informatik, Humboldt-Universität zu Berlin
Aliona von der Trenck	DHBW Heilbronn
Andreas Weber	DHBW Karlsruhe
Katja Wengler	DHBW Karlsruhe
Philipp Wicke	LMU, Ludwig Maximillians Universität Munich

Book of Abstracts

A part of the conference was a poster session.
The book of abstracts gives an overview on the discussed topics.

Decoupled Generative Fault Injection for Autonomous Robots via LLM-LDM

1st Mehdi Azarafza

Department of Computer Science
Hamm-Lippstadt University. of Applied
Sciences
Lippstadt, Germany
Mehdi.azarafza@hshl.de

2nd Faezeh Pasandideh

Department of Computer Science
Hamm-Lippstadt University. of
Applied Sciences
Lippstadt, Germany.
faezeh.pasandideh@hshl.de

3rd Ali Ehteshami Bejnordi

Department of Computer Science
Hamm-Lippstadt University. of
Applied Sciences
Lippstadt, Germany.
ali.ehteshami-bejnordi@hshl.de

4th Stefan Henkler

Department of Computer Science
Hamm-Lippstadt University. of
Applied Sciences
Lippstadt, Germany
stefan.henkler@hshl.de

5th Achim Rettberg

Department of Computer Science
Hamm-Lippstadt University. of
Applied Sciences
Lippstadt, Germany
achim.rettberg@hshl.de

EXTENDED ABSTRACT

Advancing Responsible AI within Smart Systems Development requires a shift from ideal-condition testing to proactive hazard anticipation at the edge. This research addresses the safety-compute bottleneck by introducing a decoupled generative framework that offloads intensive fault synthesis to an offline cloud phase. A real-world robotic vision dataset consisting of 796 RGB images was first collected. Separately, in an offline setting, GPT-OSS[1] deployed via Ollama [2] was used to generate fault scenarios in textual form. A subset of these scenarios was then provided as prompts to Stable Diffusion 2.1 [3] in an image-to-image configuration, enabling fault injection into the collected images. Finally, CLIP[4] (ViT-L/14) was employed as a semantic filtering tool to validate the alignment between the generated faulted images and their corresponding scenario descriptions, thereby reducing the impact of random artifacts or noise on the evaluation pipeline.

Figure 1 illustrates an overview of the proposed framework: a two-phase fault injection system where the Offline Phase has a Control Unit orchestrating the Fault Generator Unit (LLM generates scenarios, LDM creates faulty datasets, and VLM (CLIP) performs semantic comparison) to build an expected output table with validation, and the Online Phase deploys this table to the Target Unit on edge robots, which captures camera images, runs an ML model for lane following, compares against expected outputs for proactive fault detection and warnings, and then executes navigation actions.

Empirical evaluation of the ResNet18 waypoint regressor reveals limited robustness. Although a strong baseline performance is achieved on clean data ($R^2=0.940$), testing across 17 fault scenarios shows major degradation, with fog and occlusion causing an approximate 13% drop in R^2 . These results demonstrate that the proposed decoupled generative approach is a deployable necessity for reliable testing and operation of autonomous agents in Smart City environments.

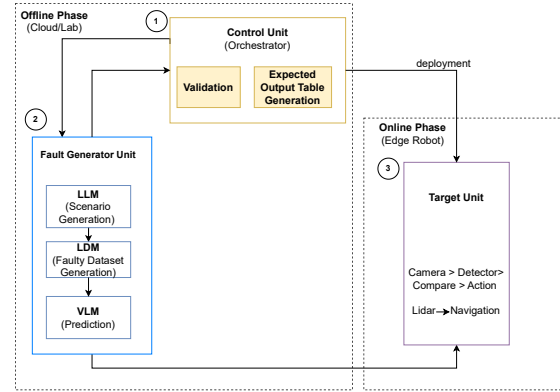


Fig1. Proposed decoupled framework architecture

ACKNOWLEDGMENT

This paper is the result of preliminary work by Hamm-Lippstadt University of Applied Sciences, Germany on the project. This work was supported by the EdgeAI-Trust project "Decentralized Edge Intelligence: Advancing Trust, Safety, and Sustainability in Europe" and project 'AKI4KMU - Automatisiertes KI-Framework für KMU' supported by the Ministry of Economic Affairs, Industry, Climate Action and Energy of the State of North Rhine-Westphalia, Co-funded by the European Union, Grant number EFRE-20800498.

REFERENCES

- [1] OpenAI, "gpt-oss-120b & gpt-oss-20b Model Card," arXiv preprint arXiv:2508.10925, Aug. 2025. [Online]. Available: <https://arxiv.org/abs/2508.10925>
- [2] Ollama Contributors, "Ollama," GitHub repository, 2026. [Online]. Available: <https://github.com/ollama/ollama>.
- [3] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), New Orleans, LA, USA, Jun. 2022, pp. 10684–10695.
- [4] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, and I. Sutskever, "Learning transferable visual models from natural language supervision," in Proc. Int. Conf. Mach. Learn. (ICML), Jul. 2021, pp. 8748–8763.

Birdstrike: Risk Assessment with Deep Learning

1st Alexander F. Auch
DHBW Mosbach
alexander.auch@
mosbach.dhbw.de

2nd Herbert Neuendorf
DHBW Mosbach
herbert.neuendorf@
mosbach.dhbw.de

3rd Tom Morschheuser
DHBW Mosbach

4rd Tom Götzing
DHBW Mosbach

I. PROJECT DESCRIPTION

Bird strike, collisions between birds and aircraft, poses a significant risk to both civil and military aviation. In collaboration with the Center for Geoinformation of the German Armed Forces (Euskirchen), Department of Applied Geosciences / Division of Biology and Ecology, we are investigating to what extent it would be possible to replace an existing manual radar-based bird strike risk assessment system [1] with an AI-driven approach.

II. PREVIOUS WORK

Based on student projects from the Business Informatics [2] and Applied Informatics programs, a system for visualizing and automatically classifying bird movements in individual grid cells (GeoRef) in German military airspace using radar data has been developed and implemented as a prototype. The AI components are based on a master's thesis [3] by a staff member at the Center for Geoinformation of the Bundeswehr (ZGeoBW), that evaluated various neural network architectures, including 3D Convolutional Neural Networks, with hyperparameter optimization for supervised learning using historical radar data [4]. These approaches enable even weakly structured radar patterns to be used for automatic bird strike risk assessment. The training is performed on the hardware of the High Performance Computing Center of the DHBW Mosbach [5].

III. IMPLEMENTATION

A. Current state of the User Interface

The current UI includes various controls to improve the user experience: compass, zoom, national borders overlay, and collapsible elements. The implemented time-lapse feature processes 12 radar frames per hour (5-minute intervals).

B. Proposed Model Architecture

Three Recurrent Neural Network (RNN) architectures, classical RNN, Long Short-Term Memory (LSTM), and Gated Recurrent Unit (GRU), were evaluated to improve prediction performance. Due to its ability to capture long-term dependencies with a relatively low parameter amount, the GRU was selected. The final model combines a many-to-one GRU with a time-distributed CNN for feature extraction. It integrates weather data and features from 12 radar images to model temporal dependencies, with predictions generated by a fully connected layer.

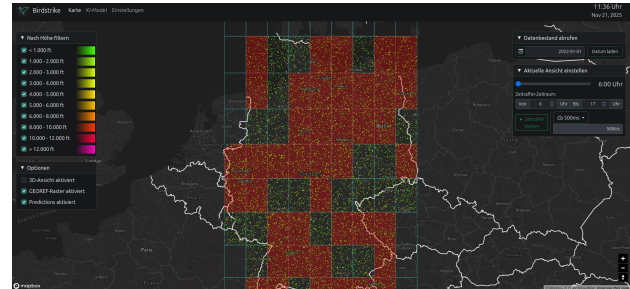


Fig. 1. Current state of the GUI with implemented changes.

IV. PROJECT GOALS

- Implementation and analysis of Recurrent Neural Networks for radar data analysis.
- Automatic classification of bird migration events into alert levels for individual grid cells.
- Microservice architecture (Docker / Podman, Go, Angular).
- AI: CNN-GRU, training on systems of the HPC-Lab DHBW-Mosbach
- Further development of the prototype implementation of the birdstrike alerting system
- Implementation of missing features (backend & frontend)

V. PROSPECTS FOR FURTHER RESEARCH

- Implementation and benchmarking of the proposed AI model, and improvements to the data integration will be carried out in subsequent student projects.

REFERENCES

- [1] W. Ruhe, "Verhütung von Vogelschlägen mit Hilfe meteorologischer Informationen", promet, 1-2 vol. 39, pp. 4-10, 2014.
- [2] Webseite Studienzentrum WION, <https://mosbach.dhbw.de/wion>
- [3] A. Witzens, "Neuronale Netze und Vogelzugwarnungen in der Luftfahrt", Masterarbeit, FernUniversität Hagen, 2021
- [4] T. Lin et al., "MistNet: Measuring historical bird migration in the US using archived weather radar data and convolutional neural networks", Methods Ecol Evol. 2019;00:1-15. <https://doi.org/10.1111/2041-210X.13280>
- [5] Webseite FIM / HPC-Labor, <https://mosbach.dhbw.de/fim>

Enriching Multi-Agent Datasets: Automated Pseudo-Label Generation for LiDAR Data

1st Jannik Loth
Faculty of Technology
DHBW Stuttgart

Stuttgart, Germany
inf22050@lehre.dhbw-stuttgart.de

2nd Leon Kühn
Faculty of Technology
DHBW Stuttgart

Stuttgart, Germany
inf22059@lehre.dhbw-stuttgart.de

3rd Marcel Vosschans
Volkman Strassen- und
Verkehrstechnik

Montabaur, Germany
m.vosschans@volkmann-sv.de

4th Alexander Baumann
Faculty of Computer Science and
Engineering

University of Applied Sciences
Esslingen
Esslingen, Germany
alexander.baumann@hs-esslingen.de

5th Matthias Drüppel
Faculty of Technology
Center for Artificial Intelligence (ZfKI)

DHBW Stuttgart
Stuttgart, Germany
matthias.drueppel@dhbw-stuttgart.de

EXTENDED ABSTRACT

Motivation: Deep Learning perception on LiDAR point clouds heavily relies on large amounts of high-quality annotated training data, whose creation is time-consuming and costly [1]. Image-based sensors on the other hand benefit from information-dense pixel values, large scale public datasets and thus often show significantly stronger performance in computer vision tasks. The core idea of this work is to use state-of-the-art image-based models to transfer semantic information from camera data to LiDAR data to reduce manual annotation effort on LiDAR point cloud data.

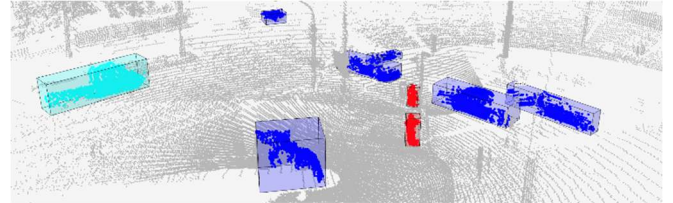
CoopScenes Dataset: For this, we use the CoopScenes dataset [2] that provides synchronized multimodal sensor data from a moving ego vehicle and a stationary infrastructure unit, comprising temporally and spatially aligned frames recorded across nine urban and rural scenes. Each agent is equipped with calibrated RGB cameras, LiDARs and localization sensors, with high-precision temporal synchronization that enable robust multimodal fusion.

Contribution: We develop a pipeline to annotate LiDAR data points with semantic class labels extracted from camera data. This pipeline includes (1) data pre-processing, (2) semantic segmentation on the camera data, (3) projecting of the camera labels onto the LiDAR point clouds followed by (4) filtering and post-processing. First, the 3D LiDAR data is transformed from the sensors into a global coordinate system and all LiDAR reflections that stem from the ground are removed. The latter is achieved using RANSAC (Random Sample Consensus). The image segmentation stage utilizes a state-of-the-art (YOLO11) semantic segmentation model to produce pixel-wise semantic classifications on the 2D-image. These predictions are then transformed into a custom 3-channel representation with information about the class, the instance and the confidence score. In the projection stage, these classified 2D-image pixels are then projected onto the 3D LiDAR data using matrix calculations. This establishes a direct association between 2D semantic labels and 3D point cloud data and results in a 6-attribute global point cloud containing information about the 3D position, the associated object class, the instance ID and a confidence score of each point. Finally, filtering and post-processing are applied to remove inconsistent, ambiguous, or noisy label assignments.

Essential parts of this are the clustering using a DBSCAN (Density-Based Spatial Clustering of Applications with Noise) to group points that are likely to belong to the same object and the calculation of ground-aligned bounding boxes to enable filtering by object. The outcome is a semantically enriched LiDAR point cloud with bounding boxes.

Results and Evaluation: Qualitative results show that the proposed cross-modal annotation pipeline produces visually consistent semantic enrichment of LiDAR point clouds (see example image below with cars in blue, pedestrians in red and trucks in turquoise). For a quantitative evaluation we hand-annotated a 3D bounding box validation set, consisting of 147 instances. Quantitative evaluation remains challenging, considering the small amount of available LiDAR ground truth data and the remaining methodological challenges in the projection and filtering stages. Initial F1 scores reach up to 7%. Overall, the results highlight clear potential as well as future improvements in robustness and label accuracy.

ACKNOWLEDGMENT



The authors would like to thank the University of Applied Sciences Esslingen for providing access to the CoopScenes dataset. We further acknowledge the Cooperative State University Baden-Württemberg (DHBW) for granting access to the Nvidia-DGX computing system.

REFERENCES

- [1] Q. Meng, W. Wang, T. Zhou, et al., ‘Towards a Weakly Supervised Framework for 3D Point Cloud Object Detection and Annotation,’ *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 8, pp. 4454–4468, Aug. 2022, doi: 10.1109/TPAMI.2021.3063611.
- [2] M. Vosschans, A. Baumann, M. Drueppel, O. Ait-Aider, Y. Mezouar, T. Dang, and M. Enzweiler, ‘CoopScenes: Multi-Scene Infrastructure and Vehicle Data for Advancing Collective Perception in Autonomous Driving,’ in *Proc. 2025 IEEE Intelligent Vehicles Symposium (IV)*, 2025, pp. 1040–1047, doi: 10.1109/IV64158.20

SELBI: A Responsible Generative AI Tool for Educator-Centred Curriculum Planning in Early Childhood Education

Shirley Ann Gauci
*Applied Research and Innovation
Centre
Malta College of Arts, Science and
Technology
Paola, Malta*
ORCID No: <https://orcid.org/0009-0006-3170-6876>

Francis Delicata
*Applied Research and Innovation
Centre
Malta College of Arts, Science and
Technology
Paola, Malta*
ORCID No: <https://orcid.org/0000-0002-5339-5275>

Alan Gatt
*Institute of Information and
Communication Technology
Malta College of Arts, Science and
Technology
Paola, Malta*
ORCID No: <https://orcid.org/0009-0005-6715-6071>

Frankie Inguanez
*Institute of Information &
Communication Technology
Malta College of Arts, Science and
Technology
Paola, Malta*
ORCID No: <https://orcid.org/0000-0001-8396-4443>

Heathcliff Schembri
*Institute of Community Services
Malta College of Arts, Science and
Technology
Paola, Malta*
ORCID No: <https://orcid.org/0000-0003-4607-0795>

Kimberly Terribile
*Institute of Applied Sciences
Malta College of Arts, Science and
Technology
Paola, Malta*
ORCID No: <https://orcid.org/0000-0002-5472-3795>

EXTENDED ABSTRACT

Applied artificial intelligence is increasingly deployed to augment knowledge work, yet educational use cases require particular attention to responsible design, transparency, and human agency. In curriculum planning, educators must translate policy aims and skill frameworks into coherent learning experiences under significant time constraints. Generic generative AI tools can accelerate drafting, but they often lack pedagogical grounding and may weaken professional accountability unless they are designed to support human evaluation through constrained, explainable interaction patterns.

SELBI (Supporting Early Childhood Education in Blue Skills with Generative Artificial Intelligence) is a TRL4 prototype of a generative AI-supported curriculum planning platform designed for educator use. The system targets an applied AI use case: producing curriculum planning suggestions that are aligned with pedagogical intent and skill targets, while maintaining educators' interpretive authority. SELBI operationalizes human-AI cooperation by combining structured prompt frameworks with curated, curriculum-aligned resources, enabling educators to iteratively query, refine, and contextualize AI outputs rather than accept them uncritically [1].

The development follows a design-based research approach with iterative cycles of prototyping and refinement [2]. The technical architecture comprises: (i) a repository of pedagogical resources and blue skills descriptors aligned with Maltese and European curriculum expectations; (ii) prompt procedures that constrain output structure to support traceability and reduce hallucinated or decontextualized recommendations; (iii) internal testing focused on usability and planning fit; and (iv) a responsible AI orientation that prioritizes transparency of assumptions, educator control over revisions, and avoidance of sensitive learner data. In this way, SELBI is positioned as an educator-facing tool that supports accountable decision-making and preserves professional judgement.

Transferability is a central design consideration. Although SELBI has been developed for use in Early Childhood Education, the underlying framework could be adapted for other education and training contexts through modular substitution of repositories and prompt constraints, including teacher education, vocational education, and continuing professional development. This approach aligns with evidence that large language models in education offer material opportunities alongside substantial risks, reinforcing the need for application-specific governance, human oversight, and responsible deployment patterns [3], [4]. The TRL4 prototype has been developed and internally tested, and educator-based usability testing has been completed. Next steps focus on consolidating these findings and undertaking external validation of output quality, after which feasible routes towards TRL5–6 will be evaluated in line with stakeholder priorities and implementation pathways.

ACKNOWLEDGMENT

The SELBI project is supported by Xjenza Malta through the Research Excellence Programme (REP-2024-005). The authors also acknowledge institutional support from the Malta College of Arts, Science and Technology (MCAST).

REFERENCES

- [1] P. Lewis *et al.*, "Retrieval-augmented generation for knowledge-intensive NLP tasks," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 9459–9474.
- [2] The Design-Based Research Collective, "Design-based research: An emerging paradigm for educational inquiry," *Educ. Res.*, vol. 32, no. 1, pp. 5–8, 2003.
- [3] E. Kasneci *et al.*, "ChatGPT for good? On opportunities and challenges of large language models for education," *Learn. Individ. Differ.*, vol. 103, Art. no. 102274, 2023.
- [4] O. Zawacki-Richter, V. I. Marin, M. Bond, and F. Gouverneur, "Systematic review of research on artificial intelligence applications in higher education—Where are the educators?" *Int. J. Educ. Technol. High. Educ.*, vol. 16, Art. no. 39, 2019.

Artificial intelligence in dual study programs: Challenges, hurdles, and potential from the perspective of teachers and students at the DHBW

Ernst Deuer
Faculty of Economics
Baden-Württemberg Cooperative State
University
Ravensburg, Deutschland
ORCID ID: 0009-0002-2053-3114

EXTENDED ABSTRACT

The growing importance of artificial intelligence (AI) also poses new challenges for dual study programmes (as in [1]). Surveys of students, professors and lecturers at the Baden-Württemberg Cooperative State University (DHBW) (as in [2]) conducted 2025 show how these stakeholders assess the relevance, use and impact of AI in everyday student life. In addition, their attitudes towards the institutional use of AI tools at the university were surveyed – in the context of the tension between regulation and empowerment.

A total of 3,569 students, 184 professors and 1,504 lecturers from all faculties and campuses of the DHBW took part in the studies. The surveys were conducted in spring and summer 2025, both online and as comprehensive surveys.

The empirical findings show both similarities and clear differences in the perspectives and practices of the groups of stakeholders involved (as in [3]): students use AI tools more frequently in their studies than teachers and also feel more confident in doing so. Scepticism about the accuracy of AI tool responses is greatest among professors. All groups reject a ban on AI tools in dual study programmes. On the other hand, all groups would like students to be taught how to use AI tools.

Based on these multi-perspective findings, starting points can be derived for how universities can actively shape and support AI-related change in dual study programmes. The integration of this content into the curriculum and the further training of teachers play a key role in this.

KEY FINDINGS TO BE ILLUSTRATED IN THE POSTER:

The use of AI tools in the context of dual study programmes is steadily increasing among dual students. Lecturers use AI tools in this context noticeably less frequently than students.

The majority of dual students increasingly perceive themselves as confident in using AI tools within the context of dual study programmes. Lecturers perceive themselves as significantly less confident in using AI tools in their teaching.

A majority of both students and lecturers trust the accuracy of AI tools' responses only 'to some extent'.

A clear majority of both students and lecturers reject a ban on AI tools within the context of dual study programmes.

Instead, universities should teach students how to use AI tools responsibly.

REFERENCES

- [1] E. Deuer, Using AI tools in dual study programs—a question of attitude, digital competence, and/or gender? Nutzung von KI-Tools im Rahmen des dualen Studiums – eine Frage der Einstellung, der Digitalkompetenz und/oder des Geschlechts? In: Handbuch E-Learning, Wolters Kluwer, Ergänzungslieferung 2025.
- [2] E. Deuer, T. Meyer, From starting your studies to starting your career. Motives, challenges, and future prospects in dual study programs. Vom Studienstart bis zum Berufseinstieg. Motive, Herausforderungen und Zukunftsperspektiven im dualen Studium, wbv 2024.
- [3] V. Gerstung-Jungherr, E. Deuer, Artificial intelligence in dual study programs: A comparative analysis of the perspectives and practices of stakeholders in dual study programs. Künstliche Intelligenz im dualen Studium: Eine vergleichende Analyse der Perspektiven und Praktiken von Stakeholdern des dualen Studiums. In: Aktuelle Erkenntnisse aus dem Studienverlaufspanel 25/2025. Stuttgart 2025.

AI Competence Development as a Foundation for Transfer in Dual Higher Education: The Center for Artificial Intelligence (ZfKI) at DHBW Stuttgart

1st Thomas Kessel
DHBW Stuttgart
Stuttgart, Germany
thomas.kessel@dhbw-stuttgart.de

2nd Dirk Reichardt
DHBW Stuttgart
Stuttgart, Germany
dirk.reichardt@dhbw-stuttgart.de

Abstract—The successful transfer of artificial intelligence (AI) into industry and society depends on qualified professionals and institutional structures linking research, teaching, and practice. Dual higher education systems offer a strong framework by integrating academic learning with industrial environments.

The Center for Artificial Intelligence (ZfKI) at DHBW Stuttgart embeds AI competencies across teaching, research, and continuing education to strengthen knowledge transfer. By combining interdisciplinary expertise, shared infrastructure, and collaborative industry projects, the center supports AI-driven innovation and competence development.

I. INTRODUCTION

Artificial intelligence increasingly shapes industrial processes and decision-making, yet its practical adoption remains challenging, especially for small and medium-sized enterprises lacking expertise. Higher education institutions therefore play a central role in competence development and knowledge diffusion.

The dual study model at DHBW Stuttgart provides an ideal environment for transfer-oriented AI education by linking academic teaching with corporate practice. The Center for Artificial Intelligence (ZfKI) coordinates interdisciplinary activities and strengthens transfer structures across faculties and partner organizations.

II. STRUCTURE AND EDUCATIONAL APPROACH

The ZfKI is an interdisciplinary center uniting faculty from business, engineering, social sciences, and health disciplines. It provides shared computational infrastructure, including high-performance computing, and integrates expertise in generative AI, computer vision, cybersecurity, and quantum-supported methods.

A core objective is to make AI tangible in teaching. Collaborative working groups exchange materials, discuss trends, and develop didactic concepts, supported by a shared repository that improves course quality while reducing preparation effort.

Students engage with practice-oriented case studies, group-based problem solving, and modern AI tools. Integrating real industry challenges fosters motivation and application-oriented competencies. Dual partners contribute practical problems and

technological feedback, enabling joint development of relevant AI solutions.

III. TRANSFER-ORIENTED PROJECTS

The center's transfer mission is reflected in numerous collaborations. The *KISMi* (Künstliche Intelligenz Support Plattform Mittelstand) initiative supports SMEs through workshops, consulting, and training formats that facilitate the identification and implementation of AI applications.

Industry-integrated projects demonstrate applied impact, such as the development of a large-language-model-based service chatbot within a teaching context that addressed both technical implementation and regulatory requirements. Additional research projects in quantum-supported optimization, infrastructure monitoring via computer vision, and sensor-based mobility datasets illustrate the scientific breadth of the center while maintaining clear transfer relevance.

ZfKI activities enable transfer on three levels: (1) competence transfer, as students apply AI skills in partner companies; (2) method transfer, since teaching concepts and materials can be reused in corporate training; and (3) innovation transfer through joint research projects that lead to practical AI applications.

This integrated approach positions the ZfKI as a bridge between academic research, higher education, and industrial innovation ecosystems.

IV. CONCLUSION

Sustainable AI transfer requires more than technological research; it depends on integrated educational structures, interdisciplinary collaboration, and strong industry partnerships. The dual higher education model provides a powerful framework for such integration and can serve as a blueprint for transfer-oriented AI initiatives in other contexts.

REFERENCES

- [1] Holmes, W., Bialik, M., & Fadel, C. (2019). *Artificial Intelligence in Education*. Center for Curriculum Redesign.
- [2] Luckin, R. et al. (2016). *Intelligence Unleashed: An Argument for AI in Education*. Pearson.
- [3] Ifenthaler, D. (2017). Learning analytics in higher education. *Educational Technology Research and Development*.

Advanced Cobot Accuracy Through Artificial Intelligence: A Two-Stage Calibration Framework with Optical Tracking Integration

1st Florian Stöckl
florian.stoeckl@dhbw-karlsruhe.de

2nd Silvan Müller
silvan.mueller@dhbw-karlsruhe.de

3rd Oliver Rettig
oliver.rettig@dhbw-karlsruhe.de

4th Marcus Strand
Robot and Human Motion (RaHM)-Lab
Duale Hochschule Baden-Württemberg
Karlsruhe, Deutschland
marcus.strand@dhbw-karlsruhe.de

I. INTRODUCTION AND MOTIVATION

With over 4.3 million industrial robots deployed globally [1], the demand for high-precision automation is critical. However, cost-effective cobots like the UR5e often lack absolute accuracy, exhibiting millimeter-range deviations. Despite high repeatability, their absolute positioning error can reach 1.28 mm [2], caused by geometric Denavit-Hartenberg (DH) deviations and non-geometric factors like joint stiffness or thermal fluctuations [3].

Conventional calibration is often too complex and time-consuming for small and medium-sized enterprises (SMEs). To address this, we propose a hybrid two-stage framework combining marker-based optical measurements with Artificial Neural Networks (ANN). Integrating kinematic modeling with AI-based compensation reduces mean absolute error by over 50 % while addressing long-term effects like wear. This provides a scalable solution for high-accuracy tasks in flexible environments.

II. METHODOLOGY: THE TWO-STAGE HYBRID APPROACH

We employ a unique two-stage procedure:

- 1) **Kinematic Calibration:** A multi-camera optical tracking system (e.g., Vicon) synchronizes the movement of retroreflective markers attached to all robot segments [4]. This allows for the simultaneous determination of DH parameters (θ, d, a, α), providing a more comprehensive dataset than traditional end-effector measurements.
- 2) **AI-Based Error Compensation:** To account for non-linear, non-geometric errors, an ANN is trained using the captured measurement data. This model specifically targets deformations caused by end-effector payloads and simulated joint wear [2].

III. PREVIOUS RESULTS AND EVALUATION

Experimental validation on a UR5e system yielded significant performance gains:

Funded by the Bundesministerium für Bildung und Forschung (BMBF)

- **General Accuracy:** The mean absolute accuracy was improved from 1.28 mm to 0.54 mm, representing a 59.3 % improvement in positioning performance [2].
- **Peak Error Reduction:** The maximum positional error was reduced from 2.42 mm to 0.98 mm [2].
- **Wear Compensation:** Simulations based on joint wear models showed that an uncompensated, additional error of up to 6.0 mm could be mitigated to less than 0.25 mm using the proposed compensation algorithm [2].

IV. FUTURE RESEARCH

Subsequent research efforts aim to generalize the hybrid calibration approach and facilitate its industrial transfer. This includes:

- **Safety-Integrated Learning:** Integration of safety layers and constraint learning during the ANN training process to ensure that AI-driven compensations adhere to industrial safety norms and operational boundaries.
- **System Generalization:** Comprehensive validation of the methodology across several different robotic arms and kinematic configurations to demonstrate the cross-platform applicability of the AI-based models.
- **Industrial Accessibility:** Development of user-friendly software solutions to lower the barriers for automated calibration in production environments, specifically targeting the needs of small and medium-sized enterprises (SMEs).

REFERENCES

- [1] International Federation of Robotics (IFR), "World Robotics 2024 - Industrial Robots," Technical Report, 2024.
- [2] J. Raible et al., "Artificial Neural Network Guided Compensation of Nonlinear Payload and Wear Effects for Industrial Robots," in *IEEE 19th International Conference on Automation Science and Engineering (CASE)*, 2023.
- [3] R. Kluz, A. Kubit, and T. Trzepieciniski, "Investigations of temperature-induced errors in positioning of an industrial robot arm," *Journal of Mechanical Science and Technology*, 2018.
- [4] O. Rettig, S. Müller, and M. Strand, "A marker based optical measurement procedure to analyse robot arm movements," in *IAS-16*, 2021.

Automating Sustainability Reporting Using AI-Assisted Data Pipelines

Zhang-Liang Xie 

DHBW Mosbach

Mosbach, Germany

Zhangliang.Xie@mosbach.dhbw.de

Alexander F. Auch

DHBW Mosbach

Mosbach, Germany

Alexander.Auch@mosbach.dhbw.de

Markus Zimmer

DHBW Mosbach

Mosbach, Germany

Markus.Zimmer@mosbach.dhbw.de

EXTENDED ABSTRACT

Digitalisation is widely recognised as a strategic driver of competitiveness and a key enabler of the green transition [3], [4]. At the same time, regulatory developments at the European level increasingly require companies to disclose structured Environmental, Social and Governance (ESG) information in order to enhance transparency, comparability, and accountability in sustainability performance [1], [2]. Despite these ambitions, small and medium-sized enterprises (SMEs) often lack the technical capacity to efficiently comply with such ESG disclosure frameworks. ESG-relevant data are predominantly collected manually from invoices, receipts, and internal records, resulting in high administrative burden, limited reproducibility, and weak traceability.

We present a software system that automates sustainability reporting while preserving deterministic regulatory compliance. The system separates AI-based document understanding from compliance-critical computation: AI-assisted optical character recognition (OCR) using Tesseract [5] and semantic layout analysis via LayoutLMv3 [8] extract structured values from heterogeneous documents. These extracted values are validated, normalized, and mapped to predefined official disclosure fields according to the official taxonomy [9]. All calculations follow predefined rule-based logic to avoid non-deterministic behavior and to reduce the risk of hallucinations associated with generative models [7].

Data processing follows a strict pipeline: import, validate, compute, persist, export. Each imported value is stored together with provenance metadata, including source reference, extraction method, reporting period, and unit normalization. A data quality score [10] is assigned to every entry to distinguish verified data from estimated or inferred values. In addition, anomaly detection compares new entries against historical values to flag implausible deviations before report generation. The final output is automatically written into the official Excel template, ensuring structural conformity while allowing manual review.

The system also introduces an optional reference dataset. This dataset stores normalized values without raw documents and enables cross-company benchmarking, anomaly detection, and statistical estimation of missing values. Predictive methods such as k-nearest neighbors or Bayesian inference can suggest

plausible proxy values for incomplete or low-confidence inputs while explicitly labeling its uncertainty. Participation in this shared pool is opt-in and designed to comply with General Data Protection Regulation (GDPR) requirements for data privacy [6]. This reference layer transforms reporting from a static compliance task into a data-driven quality improvement process.

As for the software design, the prototype is implemented with an event-driven architecture. Independent services handle document import, validation, deterministic energy calculations, persistent storage, and report generation. An internal event bus coordinates processes while preserving modularity and auditability, enabling each module to scale independently and ensuring that changes or increased load in one component do not interfere with others.

The contribution of this work is to combine AI-assisted document extraction with a compliance-first goal and an integrated data quality framework tailored to automate SME sustainability reporting. By constraining AI to data extraction and validity check stages, the system reduces reporting burden significantly while maintaining regulatory reliability and transparency.

ACKNOWLEDGMENT

The authors gratefully acknowledge the support provided by Pro DHBW Mosbach (Stiftung Pro DHBW Mosbach, Duale Hochschule Baden-Württemberg Mosbach).

The authors used GPT-5.1 Pro and GPT-5.1 Instant & Thinking (ChatGPT/OpenAI) to assist initial literature search, to assist with reference formatting in the LaTeX code and to proofread and correct the language and grammar. We did not turn on the sharing function in ChatGPT to keep the chat in private. All text was reviewed by the authors for accuracy and tone. Prompts are available from the contact author on reasonable request.

REFERENCES

- [1] EFRAG, *Voluntary Sustainability Reporting Standard for Non-Listed Micro-, Small- and Medium-Sized Undertakings (VSME)*, 2024.
- [2] European Commission, *Directive (EU) 2022/2464 on corporate sustainability reporting*, 2022.
- [3] European Commission, *An SME Strategy for a Sustainable and Digital Europe*, COM(2020) 103 final, 2020.
- [4] European Commission, *Long-Term Competitiveness of the EU: Looking Beyond 2030*, COM(2023) 168 final, 2023.

- [5] R. Smith, “An Overview of the Tesseract OCR Engine,” ICDAR, 2007.
- [6] European Parliament and Council, *Regulation (EU) 2016/679 (GDPR)*, 2016.
- [7] Z. Ji et al., “Survey of Hallucination in Natural Language Generation,” ACM Computing Surveys, 2022.
- [8] Y. Huang et al., “LayoutLMv3,” ACM MM, 2022.
- [9] EFRAG, *VSME Digital Template and XBRL Taxonomy*, 2026.
- [10] PCAF, *The Global GHG Accounting and Reporting Standard for the Financial Industry*, 2022.