Knowledge Scaffolding Recommendation System for Supervising Term Papers

Nghia Duong-Trung[1,2(✉)], Xia Wang[1], Rahul Rajkumar Bhoyar[1], Angelin Mary Jose[1], Silke Elisabeth Wrede[3], Lars van Rijn[3], Theresa Panse[3], Claudia de Witt[3], and Niels Pinkwart[1,4]

[1] German Research Center for Artificial Intelligence (DFKI), Alt-Moabit 91 C, 10559 Berlin, Germany
{nghia_trung.duong,xia.wang,rahul_rajkumar.bhoyar,angelin_mary.jose, niels.pinkwart}@dfki.de
[2] IU International University of Applied Sciences, Frankfurter Allee 73A, 10247 Berlin, Germany
[3] FernUniversität in Hagen, Universitätsstraße 11, 58097 Hagen, Germany
{silke.wrede,lars.vanrijn,claudia.dewitt}@fernuni-hagen.de
[4] Humboldt-Universität zu Berlin, Unter den Linden 6, 10117 Berlin, Germany

Abstract. Generative AI is increasingly used to enhance educational support in higher education, particularly for distance learning. This paper introduces the Term Paper Recommendation System (TPRS), an AI-powered scaffolding tool designed to assist students in selecting and refining research topics for academic writing. TPRS integrates generative language models with knowledge-based and expert-driven recommendation strategies, dynamically adapting feedback based on the student's confidence level. The system leverages structured validation, multi-shot prompting with historical supervision data, and semantic similarity for literature recommendations. Deployed within a Bachelor of Arts program in Culture and Social Sciences at FernUniversität in Hagen, TPRS was evaluated using the CRS-Que framework and expert grading of student submissions. Results show statistically significant improvements in topic formulation quality, enhanced engagement, and reduced instructor workload. However, user feedback highlighted the need for improved transparency and control. TPRS offers a novel hybrid architecture that positions AI as a learning scaffold rather than an automation tool. This work contributes to responsible AI integration in higher education by demonstrating how generative systems can support inquiry-based learning while preserving student agency.

Keywords: AI Recommendation · Intelligent Tutoring · Generative AI · Distance Learning · Knowledge Scaffolding · CRS-Que

1 Introduction

Integrating AI technologies into teaching and learning has become a key strategy for addressing pedagogical challenges and improving instructional efficiency

K. Tammets et al. (Eds.): EC-TEL 2025, LNCS 16063, pp. 138–152, 2026.
https://doi.org/10.1007/978-3-032-03870-8_10

in higher education [2]. AI-driven tools have been widely explored in various domains, including personalized education [14], intelligent tutoring [17], and automated feedback [30]. This study focuses on a specific use case: developing and evaluating a TPRS designed to support German-speaking students in a Bachelor of Arts distance learning program in Culture and Social Sciences. A core challenge in our distance learning programs is the cognitive complexity of research topic selection and question formulation. Students are expected to independently propose research topics and refine their questions, usually submitting them through the Moodle platform for instructor feedback. However, this process is time-intensive, prone to delays, and results in inconsistent guidance across students. From the cognitive perspective, distilling complex academic materials into a focused research idea can be overwhelming, leading to cognitive overload and fragmented workflows. Delayed feedback further demotivates students, reducing engagement and academic persistence.

To mitigate these issues, we propose AI-powered knowledge scaffolding as a pedagogical framework for enhancing student support. Scaffolding, a well-established concept in educational psychology, refers to providing structured, adaptive guidance gradually reduced as learners develop competence [9,37]. Unlike AI systems that provide fully automated answers, TPRS is a scaffolded learning assistant [15,23], supporting students in refining research topics while preserving independent inquiry and critical thinking. The system employs large language models (LLMs) alongside traditional knowledge-based and expert-based recommendation mechanisms to (i) *Analyze student inputs and suggest structured and relevant research topics*; (ii) *Provide instant, formative feedback on the clarity and feasibility of research questions*; and (iii) *Recommend domain-specific literature to guide further academic exploration.* Unlike purely generative AI approaches, which may encourage overreliance on automated outputs, TPRS is explicitly designed to foster student autonomy by offering personalized, inquiry-driven support. Its recommendations act as cognitive scaffolds, helping students construct their ideas without replacing the need for academic reasoning and decision-making. To evaluate the effectiveness of TPRS, we apply the CRS-Que framework [13] to assess system usability and pedagogical impact. This framework examines key dimensions such as accuracy, explainability, user control, trust, and transparency. Our findings highlight that TPRS recommendations: (i) Enhance student engagement by reducing cognitive barriers in topic selection; (ii) Improve efficiency in feedback loops, allowing students to refine ideas more quickly; and (iii) Alleviate instructor workload, enabling educators to focus on higher-order feedback.

2 Related Work

The integration of AI into education has been extensively studied, focusing on enhancing student learning, personalizing feedback, and streamlining administrative processes. Pérez-Ortiz *et al.* summarized the concept of AI-based personal learning companions, emphasizing their potential to provide adaptive support

tailored to the personalization of learning [27]. Many systematic reviews of AI applications in higher education have been conducted, highlighting the growing role of AI-powered tools in facilitating self-directed learning and academic support [6,26,36]. The UNESCO AI Competency Frameworks for Teachers and Students highlight the urgent need to equip educators and learners with AI-related knowledge, skills, and ethical awareness, emphasizing AI's transformative impact on education. The teacher framework underscores AI pedagogy, ethical considerations, and professional learning, ensuring educators can effectively integrate AI into teaching [20]. Meanwhile, the student framework focuses on AI literacy, ethical responsibility, and system design, preparing learners for an AI-driven world [21]. Together, these frameworks motivate research on AI-driven educational systems, advocating for adaptive learning, personalized AI tutors, and responsible AI development to enhance educational equity, inclusivity, and sustainability.

In the context of higher education, the integration of a literature recommender system has the potential to enhance the learning experience and to support students in their academic journey, such as course selection and planning [34], provision of personalized feedback and guidance [19] in an online learning environment [1,25]. Recommender systems are crucial in supporting students and researchers in selecting relevant literature, refining research topics, and structuring their academic writing [10,11,24]. Beel *et al.* reviewed research-paper recommender systems, identifying challenges in optimizing relevance, novelty, and personalization [4]. Similarly, Knoth & Herrmannova explored recommendation mechanisms that assist researchers in identifying pertinent literature by measuring the semantic similarity of publications connected in a citation network [16]. Haruna *et al.* addressed the challenge of efficiently guiding researchers to pertinent publications amid an overwhelming volume of information [12]. In that paper, the authors used a public dataset to demonstrate personalized recommendations regardless of the research field and regardless of the user's expertise. Similarly, Chen & Lee tackled the challenge of information overload in academia by developing a recommender system tailored to assist scholars in discovering pertinent research papers. The system leverages big scholarly data to enhance the relevance and accuracy of its recommendations, thereby facilitating more efficient literature exploration for researchers [7]. Recent surveys on scientific paper recommender systems highlight significant challenges, including cold start, trust, transparency, privacy, and the lack of unified scholarly data standards [3,29].

3 TPRS System Design and Development

3.1 Challenges in the Research-Oriented Term Paper Process

In the final semester of a Bachelor of Arts program, students must write a term paper synthesizing knowledge from previous courses, particularly in Media Education and Media Pedagogy. This research-oriented semester follows a structured three-phase process but presents several challenges for students and teachers. The Term Paper Preparation phase requires students to engage with course materials independently, reflect on their thematic interests, and formulate a

research topic with corresponding questions. However, this process can be overwhelming. The intention to distill broad and complex learning materials into a specific research focus often leads to cognitive overload. Many students struggle to articulate straightforward, well-defined research topics and questions without substantial guidance, resulting in frustration and delays.

The Student-Teacher Interaction phase, conducted via a Moodle discussion forum, adds another layer of difficulty. Students must submit their preliminary research ideas for review, but feedback is not immediate. The long wait times for responses disrupt their workflow and motivation. Additionally, the quality of initial proposals varies significantly, often requiring multiple rounds of revision. While necessary for refining research ideas, this iterative process extends the timeline and can make students feel stuck in a cycle of revisions rather than progressing toward their final paper. Once the research topic is approved, students move into the Term Paper Writing phase, where they develop their final submission. However, delays from the earlier phases can compress the time available for actual writing, potentially affecting the quality of their work.

From the teachers' perspective, the process is equally challenging. The individual nature of the feedback required in the Moodle forum creates a substantial workload. Reviewing and refining numerous proposals within tight academic timelines is labor-intensive, especially for large cohorts. Additionally, ensuring consistency and depth of feedback across students with varying abilities and research topics is difficult. Some students require extensive guidance, while others need minimal intervention, leading to an uneven distribution of effort. The administrative burden of tracking revisions across multiple iterations further consumes valuable time that could otherwise be spent on enhancing the learning experience.

3.2 System Architecture and Workflow

TPRS aims to encourage students to reflect more deeply and effectively on their term paper topics and research questions through AI-generated feedback and recommendations. This eventually helps them identify the subject and relevant content for their term papers. Our TPRS system combines generative AI with traditional recommendation techniques, including knowledge-based [32] and expert-based methods [5], to provide reliable and contextually relevant suggestions. The overall workflow of TPRS is illustrated in Fig. 1.

The process begins when a student submits a short initial term paper proposal, which consists of a research topic (RT) and a set of research questions (RQ_i, where $i \geq 0$). A sentiment analysis (SA) module assesses the student's confidence level in their submission. Based on this evaluation, TPRS selects either the knowledge-based engine (KE) or the expert-based engine (EE) to generate feedback. If the student's confidence is high, the knowledge-based engine is activated, offering literature-based recommendations that provide further relevant academic reading resources. However, if the student's confidence is low, which means that he or she might need more professional help in specifying a research topic, the expert-based engine provides more fundamental guidance, such as

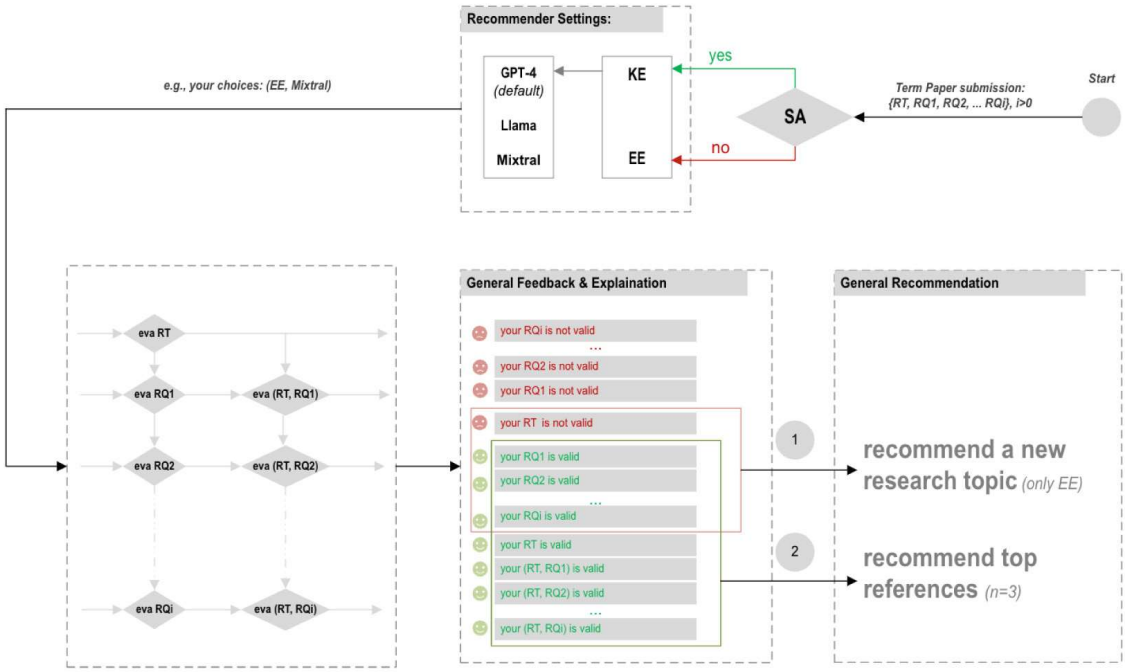

Fig. 1. TPRS v2.0 System Workflow.

refining the research topic or restructuring the research questions. Students also have the flexibility to override the system's choice and manually select a recommender engine and LLMs, such as ChatGPT-4o, Mixtral-8x7B-Instruct-v0.1, and Llama-3-8B.

Following the initial analysis, the system evaluates the research topic and questions through structured validation. A set of evaluation modules (Eva_X) have been designed to assess the validity of these submitted research statements. More specifically, these validators analyze coherence and feasibility by evaluating the research topic (Eva_{RT}), each research question (Eva_{RQ1}, Eva_{RQ2}, ..., Eva_{RQ_i}), and the pair of the research topic and a research question ($\text{Eva}_{\text{RT, RQ}_i}$). Currently, the validation module is simply implemented through structured prompting techniques using ChatGPT-4o.

Once the validation process is complete, the general feedback module can provide positive or negative feedback based on the results with a brief explanation (for system transparency and trustworthiness). Furthermore, the current TPRS only generates general recommendation information for two situations: 1) when the research topic is invalid but the research questions are valid, the expert-based engine suggests a new research topic aligned with the valid research questions. 2) When the research topic and questions are all valid, the knowledge-based engine recommends up to three most relevant academic literature for students' further reading. Such a general recommendation module ensures that students receive contextually appropriate and academically rigorous suggestions that help them refine their submissions effectively.

TPRS employs an expert-based multi-shot prompting approach to enhance accuracy, leveraging historical interactions from previous Moodle-based supervi-

sion. It has been trained on a dataset of approximately 70 students, each engaging in an average of 13∼15 interactions with teachers, allowing it to simulate high-fidelity teacher feedback. The dataset and those interactions were collected from the same bachelor program during 2023 and the winter semester of 2023/2024. The multi-shot prompting technique incorporates historical exchanges into AI-generated responses, ensuring the system provides detailed and pedagogically relevant recommendations. Those pairs of interactions were fed into ChatGPT-4o to create a master prompt [18], acting as a teacher or an expert. Additionally, TPRS uses Chain-of-Thought (CoT) reasoning to decompose complex queries into logical steps, improving interpretability and the overall coherence of feedback.

Generally, the feedback generation process follows a structured algorithm (see Algorithm 1). First, the system constructs a feedback prompt by integrating historical interactions into a structured format. A prompt template is generated, instructing the model to produce feedback that closely mimics teacher responses. This template combines the student's initial submission to form the final input prompt. The selected LLM, whether ChatGPT- 4o, Mixtral-8x7B-Instruct-v0.1, or Llama-3-8B, processes the input and generates structured feedback. The feedback output aims to provide actionable insights that guide students in refining their research topics and formulating more precise research questions.

Algorithm 1. Feedback Generation via Multi-shot Prompting

Require: Student submission $\mathbf{T}$, historical interactions $\mathcal{C} = \{C_1, C_2, \ldots, C_n\}$, number of top relevant interactions k, LLMs $\text{LLM}_1, \text{LLM}_2, \text{LLM}_3$

1: Generate a feedback prompt template P: $P \leftarrow \text{Generate\_Prompt}(\mathcal{C})$
2: Integrate student submission Q into the prompt: $P' \leftarrow [P, \mathbf{T}]$
3: Select $\text{LLM}_j \in \{\text{LLM}_1, \text{LLM}_2, \text{LLM}_3\}$
4: Generate feedback F_j using the LLM_j: $F_j \leftarrow \text{LLM}_j(P')$
5: **return** Feedback F_j

4 Evaluation

4.1 Demographic Insights

In the evaluation, we asked for voluntary in one cohort in the Bachelor of Arts program. Eighteen students participated, covering six age groups. Figure 2 presents a detailed analysis of participant demographics and behavioral characteristics across different age groups. The study included 18 students, spanning six age groups. Female participants accounted for the majority, representing 77.78% of the total, with 14 participants. Male and miscellaneous gender participants made up 16.67% and 5.55%, with 4 and 1 participants, respectively. The largest age group was 35–44 years, which included 7 participants, followed by the 25–34 and 45–54 years groups with 3 participants.

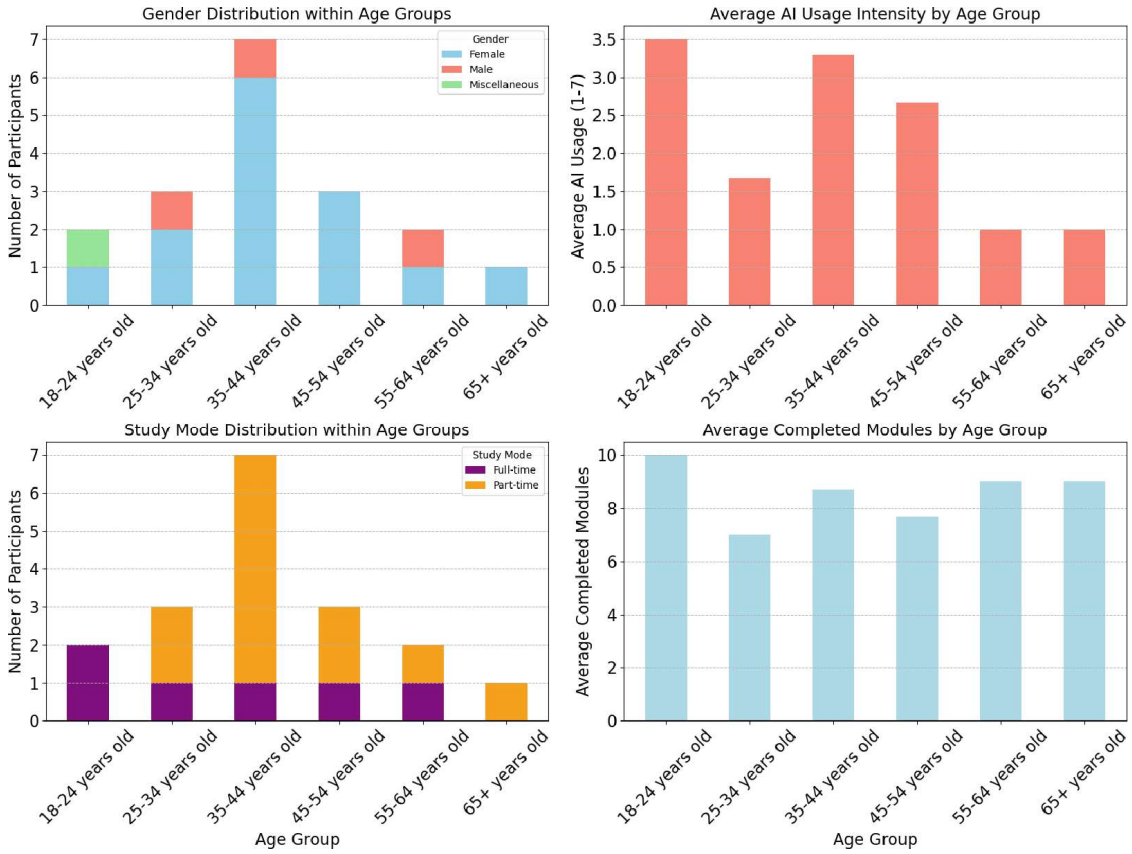

Fig. 2. Summary of Demographic Data.

AI usage intensity varied significantly across age groups. The highest engagement was reported by younger participants in the 18–24 years group, with a mean score of 3.5, and mid-career participants in the 35–44 years group, with a mean score of 3.3. In contrast, older participants in the 55–64 years and 65+ years groups reported lower AI adoption, with mean scores of 0.8 and 1.0, respectively. The overall average AI usage across all participants was 2.1. Study mode preferences also differed by age. Among all participants, 33.33%, corresponding to 6 individuals, were full-time students, while 66.67%, corresponding to 12 individuals, studied part-time. Despite these differences, academic progress remained relatively stable across age groups. On average, participants completed 8.5 modules, with the highest completion rate observed among participants in the 18–24 years group, who completed an average of 9.5 modules. The lowest completion rate was recorded in the 25–34 years group, with an average of 6.8 modules. These findings highlight the need for flexible and accessible educational technologies to support students with diverse learning needs. The higher AI engagement among younger participants suggests opportunities to integrate AI-driven tools into academic workflows, potentially enhancing learning experiences and improving study efficiency.

4.2 System Assessment Using the CRS-Que Framework

The authors conducted the pilot evaluation for the Bachelor of Arts program in Culture and Social Sciences at FernUniversität in Hagen from 01.11.2024 to 20.12.2024. We apply the CRS-Que: A User-centric Evaluation Framework for Conversational Recommender Systems [13] to evaluate our system. We believe this framework could help researchers conduct standardized user-centric research for conversational recommender systems and provide practitioners with insights into designing and evaluating a CRS from users' perspectives. The CRS-Que framework assessed multiple dimensions of user experience, including **Perceived Qualities**, **User Beliefs**, **User Attitudes**, and **Behavioral Intentions**. Below, we summarize the key findings from the evaluation results.

The **Perceived Qualities** dimension included metrics such as *Accuracy*, *Novelty*, and *Explainability*. The participants rated the system for:

- *Accuracy*: Users found the recommendations to be well-chosen, relevant, and interesting.
- *Novelty*: Participants noted that the system helped them discover new aspects or ideas. Unexpected and positive recommendations were frequently highlighted as a key strength.
- *Explainability*: The system effectively communicated the reasons behind its recommendations, enhancing transparency and trust.

The **User Beliefs** dimension measured constructs such as *Perceived Ease of Use*, *Perceived Usefulness*, and *User Control*. The participants rated the system for:

- *Ease of Use*: Participants reported that the system was intuitive and easy to use for obtaining recommendations related to their academic tasks.
- *Usefulness*: TPRS was highly effective in supporting users to complete core tasks, including identifying topics and relevant resources.
- *Control*: Users appreciated the ability to tailor recommendations and felt in control of modifying outputs based on their preferences.

The **User Attitudes** dimension evaluates *Trust and Confidence* and *Satisfaction*; specifically:

- Users trusted the system's recommendations and expressed confidence in its ability to assist with planning homework.
- Satisfaction scores reflected overall happiness with the system's performance, highlighting its role in improving task outcomes.

The **Behavioral Intentions**, including *Intention to Use*, revealed strong user loyalty and advocacy. Participants expressed their willingness to continue using the system for academic planning and recommend it to peers.

Besides the CRS-Que framework, the authors added **Judgment on Quality of Work** dimension by asking the students: How would you rate the quality of your work on the three main tasks for finding a topic for your homework in the module? Please rate individually in the case study, theoretical references, selected topic, and central question. We present the details of those dimensions in Table 1.

Table 1. Evaluation Metrics are defined in the CRS-Que and our department's evaluation frameworks. 7-point Likert scale: Do not agree at all, Disagree, Rather disagree, Neither agree nor disagree, Agree more, Agree, I completely agree.

Dimension	Code	Item
Perceived Qualities		
Accuracy	Acc1	The recommendations were well chosen
	Acc2	The recommendations were relevant
	Acc3	The recommendations were interesting
Novelty	Nov1	TPRS helped me discover new aspects/ideas
	Nov2	I received unexpected recommendations, pointing out new aspects that I would not have thought of
	Nov3	TPRS gave me recommendations that I would not have considered at first, but which turned out to be a positive and unexpected discovery
	Nov4	They were welcome recommendations, some aspects of which I would not have discovered elsewhere
Explainability	Expl1	RS explained to me why I received the recommendations and related information
	Expl2	The logic of the information received in the recommendations was explained to me by TPRS
	Expl3	The reason why I received the recommendations was given to me by TPRS
User Beliefs		
Perceived Ease of Use	PEU1	I was able to easily use TPRS to get recommendations related to my research interest
	PEU2	With TPRS, it was easy to find the recommendations that would help me plan my homework
	PEU4	It was easy to get recommendations that I found helpful
Perceived Usefulness	PU1	TPRS helped me to find the ideal recommendations for my solutions to the three leading tasks from the module
	PU2	It is easy to use TPRS to help me complete the three tasks
	PU3	TPRS gave me helpful recommendations for solving the three main tasks
User Control	UC1	I felt like I had control over changing the recommendation output in TPRS
	UC2	I was able to control the issue of recommendations well
	UC3	I had control over tailoring the recommendations to my interests
User Attitudes		
Trust & Confidence	TC1	TPRS's recommendations can be trusted
	TC2	I can rely on TPRS when I want to receive recommendations on how to plan my homework
	TC4	The recommendations convinced me
	TC5	I was confident that the recommendations would help me in planning my homework
Satisfaction	Sat1	I was satisfied with the recommendations
	Sat4	I was happy with TPRS
Behavioral Intentions		
Intention of Use	IU1	I would use TPRS again
	IU2	I would use TPRS more often while planning my homework
	IU3	I will tell fellow students about TPRS
Quality of Work (Extracted from our department's evaluation framework)		
Case study	QF	Evaluative criteria
Theoretical References	QTB	Evaluative criteria
Selected Topic	QGT	Evaluative criteria
Central question	QZF	Evaluative criteria

4.3 Measuring System Effectiveness

We analyzed user ratings across multiple evaluation dimensions to evaluate the effectiveness of the TPRS. Figure 3 presents the average scores and standard deviations. The system performed well in several key dimensions, though some areas require improvement. As shown in Fig. 3, the highest-scoring criteria were *Accuracy* (*Acc1*, *Acc2*, *Acc3*) and *Perceived Usefulness* (*PU1*, *PU2*, *PU3*), all of which had mean scores above 5. It indicates that participants generally found TPRS useful and effective in providing relevant recommendations. However, *Satisfaction* (*Sat1*, *Sat4*) received mean scores below 5, suggesting that while students acknowledged the system's utility, they were not entirely satisfied with its overall performance. Similarly, *Trust and Confidence* (*TC1*, *TC2*) received moderate ratings, indicating that users found the system somewhat reliable but still had reservations. The criteria related to *Explainability* (*Expl1*, *Expl2*, *Expl3*) received moderate ratings, indicating that while users generally understood the rationale behind the recommendations, there is room for improvement in making the system's decision-making process more transparent. Given that trust (*TC1*, *TC2*) was also rated moderately, enhancing explainability could help strengthen user confidence in the system.

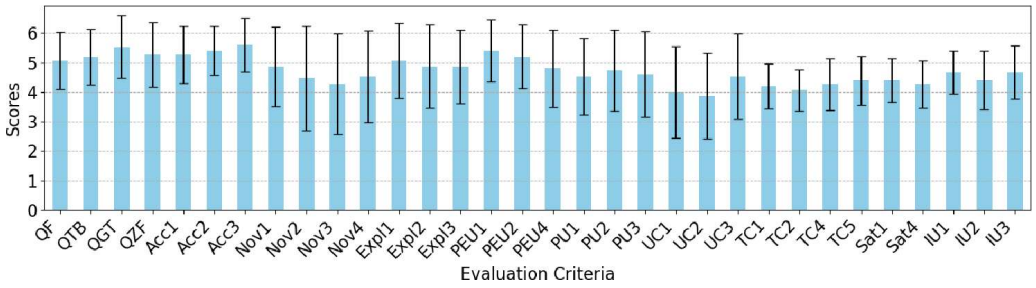

Fig. 3. Average Scores with Standard Deviations Across Evaluation Dimensions.

4.4 Impact Assessment on Student Submissions

To systematically evaluate the impact of TPRS on students' academic writing, domain experts conducted a structured assessment of submission quality. Drawing on the benefits of personalized recommendations in supporting the scientific writing process, we hypothesized that the system would contribute to an improvement in submission quality over time. To test this hypothesis, the research team collaborated with course lecturers to develop a coding scheme that ensures consistency and objectivity in evaluation. The scheme was applied to assess both the initial and final submissions recorded in the system log, as well as submissions made in the tutoring forum. A two-stage analysis was conducted using an 11-point scale, measuring four key academic criteria: use case description (3 points), theoretical background (3 points), topic of term paper (2 points), and research question formulation (3 points).

Two domain experts independently rated each submission, ensuring inter-rater reliability. To quantify agreement among raters, we calculated quadratic weighted Cohen's k, which revealed substantial agreement ($k = .675$, $p < .001$). Discrepancies in coding were resolved through discussion to establish consensus. Additionally, to account for potential covariate effects, we collected demographic information, including age, gender, study mode (full-time or part-time), and academic progress (measured through completed credit points). However, subsequent analysis showed no significant correlations between these variables and submission quality, confirming that covariates did not influence the observed trends.

The results indicate a modest but statistically significant improvement in submission quality following the use of TPRS ($\mathbf{M}_{t1} = 6.79$, $\mathbf{sd}_{t1} = 2.07$; $\mathbf{M}_{t2} = 7.53$, $\mathbf{sd}_{t2} = 1.74$). Given the absence of significant covariate effects, a directional paired t-test was deemed appropriate to assess the significance of this change. A Shapiro-Wilk test revealed that the $t1$ data was not normally distributed ($p < .05$). Further examination identified a single outlier; however, upon reviewing the corresponding submission, it was determined to be a valid data point and retained in the analysis. As directional t-tests are generally robust to minor deviations from normality and the presence of outliers, we proceeded with the analysis. Hedge's g was used to estimate effect size, yielding a small to moderate effect ($g = .44$, $p < .05$), indicating that TPRS significantly enhanced student submissions' quality.

5 Lessons Learned and Future Work

5.1 Addressing the Challenge of Missing Labeled Data

In scenarios like ours, where structured feedback data was unavailable, we faced significant hurdles in developing a retrieval system capable of generating high-quality prompts, denoted as $P = \text{Generate\_Prompt}(\mathcal{C})$. To address this challenge, we directly involved teachers in annotating their reasoning and judgment criteria for providing feedback on students' term papers. Teachers were asked to reflect on their past interactions and respond to strucHowed questionnaires such as: "Does the term paper contain practical examples?", "How many topics are covered in the term paper?", "Is the problem well formulated?", "Are appropriate theories or theoretical concepts provided?", "Do the listed theories align with the Bachelor Module Handbook?", "Does the term paper contextualize its findings within an educational framework?" This approach revealed a critical insight: teachers inherently apply their judgment criteria, honed through years of experience, to evaluate and provide feedback on term papers [22]. Although consistent, these criteria are typically applied implicitly rather than explicitly. Therefore, the central question became: how can we fine-tune LLMs to replicate these well-established, experience-driven judgment criteria? [8]. We believe this challenge is not unique to our project. Many educators and developers face similar difficulties in creating robust AI models without access to comprehensive, labeled datasets. To address this issue, we devised a forward-looking strategy:

leveraging the outputs generated by LLMs during the pilot phase and engaging teachers to refine and annotate these outputs. By asking educators to correct and expand upon the model's suggestions, we aim to build a high-quality labeled dataset iteratively. This dataset will not only support the future development of TPRS but also serve as a valuable resource for other academic AI systems.

5.2 Effectiveness of Small Language Models (SLMs)

The experimental results revealed an intriguing trend: students engaged more frequently with Mixtral-8x7B-Instruct-v0.1, regarding the number of interactions and the diversity of term paper content explored. This increased interaction suggests that Mixtral-8x7B-Instruct-v0.1, with its tailored instruction tuning, aligned more closely with the student's needs and expectations. Moreover, the outputs of Mixtral-8x7B-Instruct-v0.1 demonstrated greater consistency, as evidenced by the smallest variation in output token counts across all models. This consistency hints at its potential to deliver predictable, reliable feedback under varied conditions, a critical factor in building trust and usability for educational AI tools. These findings underscore a pivotal opportunity: With carefully designed prompts and well-labeled datasets, SLMs can serve as viable alternatives to larger models. They offer scalable and cost-effective solutions without compromising on quality, which is particularly relevant for educational institutions and researchers operating under limited computational resources or budget constraints [28,31].

5.3 Applying AI to Support Learning, Not Replace It

From the outset, we recognized the potential risks associated with generative AI in academic contexts [33,35]. If students were to rely on AI to perform the core tasks of researching, writing, and analyzing, the educational value of the learning process would be lost. Instead of fostering deep learning and critical thinking, such misuse could lead to superficial engagement with the subject matter. To mitigate this risk, we deliberately chose to hard-code safeguards into TPRS. For example, when prompted with questions such as "Can you write the term paper for me?", TPRS is programmed to respond in ways that redirect the student toward active participation in the learning process, offering guidance, feedback, and suggestions rather than performing the task. This design philosophy ensures that TPRS is a supportive tool that encourages inquiry-based learning. By guiding students through the iterative process of refining their ideas and addressing weaknesses in their term papers, TPRS maintains the integrity of the learning experience while leveraging AI's strengths to enhance accessibility and efficiency.

6 Conclusion

This study introduced the TPRS, a hybrid AI-powered platform designed to scaffold students through the cognitively demanding process of academic inquiry in

distance learning. By integrating generative language models with knowledge- and expert-based recommender strategies, TPRS supports learners in formulating, refining, and validating research topics while alleviating instructor workload and feedback bottlenecks.

A key technical contribution of this work is the hybrid recommendation architecture, which integrates LLMs with structured knowledge repositories to ensure pedagogically sound and domain-relevant recommendations. Unlike traditional AI-assisted writing tools, TPRS dynamically selects recommendation strategies based on user confidence levels, employing knowledge-based retrieval for experienced users and expert-driven scaffolding for those needing additional guidance. The multi-shot prompting mechanism, informed by historical teacher interactions, enhances feedback accuracy while preserving adaptability. Furthermore, the evaluation framework and logging infrastructure provide a scalable foundation for future iterations and deployment across diverse educational settings. Through a systematic evaluation, we analyzed the effectiveness of TPRS using system logs and expert assessments, comparing students' initial and final submissions. The findings indicate a statistically significant improvement in submission quality, suggesting that AI-driven formative feedback enhances students' ability to structure their academic work. Additionally, the CRS-Que framework confirmed that students perceived the system as accurate, transparent, and pedagogically useful, though challenges related to user control and explainability remain areas for improvement. More broadly, this study contributes to the ongoing development of AI-driven educational tools, offering a scalable, modular framework that balances automation with domain expertise. By demonstrating the feasibility of hybrid recommendation architectures in academic writing support, TPRS serves as a blueprint for future AI-assisted research guidance systems, paving the way for larger-scale deployments in distance learning and beyond.

Acknowledgment. The authors kindly appreciate the support of CATALPA—Center of Advanced Technology for Assisted Learning and Predictive Analytics, FernUniversität in Hagen—through the Project "AI.EDU Research Lab 2.0".

References

1. Abel, M., Germain, W., Mahatody, T.: Pedagogical alignment of large language models (LLM) for personalized learning: a survey, trends and challenges (2024)
2. Al-Zahrani, A.M., Alasmari, T.M.: Exploring the impact of artificial intelligence on higher education: the dynamics of ethical, social, and educational implications. Humanit. Soc. Sci. Commun. **11**(1), 1–12 (2024)
3. Bai, X., Wang, M., Lee, I., Yang, Z., Kong, X., Xia, F.: Scientific paper recommendation: a survey. IEEE Access **7**, 9324–9339 (2019)
4. Beel, J., Gipp, B., Langer, S., Breitinger, C.: Paper recommender systems: a literature survey. Int. J. Digit. Libr. **17**, 305–338 (2016)
5. Bhoyar, R.R., Wang, X., Duong-Trung, N.: KaggleGPT: prompt-based recommender system for efficient dataset discovery. In: Proceedings of DELFI Workshops 2024, pp. 10–18420. Gesellschaft für Informatik eV (2024)

6. Castillo-Martínez, I.M., Flores-Bueno, D., Gómez-Puente, S.M., Vite-León, V.O.: AI in higher education: a systematic literature review. In: Frontiers in Education, vol. 9, p. 1391485. Frontiers Media SA (2024)
7. Chen, T.T., Lee, M.: Research paper recommender systems on big scholarly data. In: Yoshida, K., Lee, M. (eds.) PKAW 2018. LNCS (LNAI), vol. 11016, pp. 251–260. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-97289-3_20
8. Duong-Trung, N., Wang, X., Kravčík, M.: BloomLLM: large language models based question generation combining supervised fine-tuning and bloom's taxonomy. In: Ferreira Mello, R., Rummel, N., Jivet, I., Pishtari, G., Ruipérez Valiente, J.A. (eds.) EC-TEL 2024. LNCS, vol. 15160, pp. 93–98. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-72312-4_11
9. Ertugruloglu, E., Mearns, T., Admiraal, W.: Scaffolding what, why and how? A critical thematic review study of descriptions, goals, and means of language scaffolding in bilingual education contexts. Educ. Res. Rev. 100550 (2023)
10. Fathi, J., Rahimi, M.: Utilising artificial intelligence-enhanced writing mediation to develop academic writing skills in EFL learners: a qualitative study. Comput. Assist. Lang. Learn. (2024)
11. Fleckenstein, J., Liebenow, L.W., Meyer, J.: Automated feedback and writing: a multi-level meta-analysis of effects on students' performance. J. Educ. Manag. Instr. **6**, 1162454 (2023)
12. Haruna, K., Akmar Ismail, M., Damiasih, D., Sutopo, J., Herawan, T.: A collaborative approach for research paper recommender system. PLoS One **12**(10), e0184516 (2017)
13. Jin, Y., Chen, L., Cai, W., Zhao, X.: CRS-QUE: a user-centric evaluation framework for conversational recommender systems. ACM Trans. Recommender Syst. **2**(1), 1–34 (2024)
14. Kaswan, K.S., Dhatterwal, J.S., Ojha, R.P.: AI in personalized learning. In: Advances in Technological Innovations in Higher Education, pp. 103–117. CRC Press (2024)
15. Kim, J., Lee, H., Cho, Y.H.: Learning design to support student-AI collaboration: perspectives of leading teachers for AI in education. Educ. Inf. Technol. **27**(5), 6069–6104 (2022)
16. Knoth, P., Herrmannova, D.: Towards semantometrics: a new semantic similarity based measure for assessing a research publication's contribution. D-Lib Mag. **20**(11), 8 (2014)
17. Lin, C.C., Huang, A.Y., Lu, O.H.: Artificial intelligence in intelligent tutoring systems toward sustainable education: a systematic review. Smart Learn. Environ. **10**(1), 41 (2023)
18. Lo, L.S.: The art and science of prompt engineering: a new literacy in the information age. Internet Ref. Serv. Q. **27**(4), 203–210 (2023)
19. Maier, U., Klotz, C.: Personalized feedback in digital learning environments: classification framework and literature review. Comput. Educ.: Artif. Intell. **3**, 100080 (2022)
20. Miao, F., Cukurova, M.: AI competency framework for teachers (2024)
21. Miao, F., Shiohira, K.: AI competency framework for students (2024)
22. Monarch, R.M.: Human-in-the-Loop Machine Learning: Active learning and annotation for human-centered AI. Simon and Schuster (2021)
23. Munshi, A., Biswas, G., Baker, R., Ocumpaugh, J., Hutt, S., Paquette, L.: Analysing adaptive scaffolds that help students develop self-regulated learning behaviours. J. Comput. Assist. Learn. **39**(2), 351–368 (2023)

24. Muslimin, I.: The use of chatgpt to improve scientific article productivity of postgraduate students. J. Educ. Manag. Instr. **3**, 63–71 (2023)
25. Ng, C., Fung, Y.: Educational personalized learning path planning with large language models (2024)
26. Ouyang, F., Zheng, L., Jiao, P.: Artificial intelligence in online higher education: a systematic review of empirical research from 2011 to 2020. Educ. Inf. Technol. **27**(6), 7893–7925 (2022)
27. Perez-Ortiz, M., Novak, E., Bulathwela, S., Shawe-Taylor, J.: An AI-based learning companion promoting lifelong learning opportunities for all. arXiv preprint arXiv:2112.01242 (2021)
28. Shan, R.: Language artificial intelligence at a crossroads: deciphering the future of small and large language models. Computer **57**(8), 26–35 (2024)
29. Sharma, R., Gopalani, D., Meena, Y.: An anatomization of research paper recommender system: overview, approaches and challenges. Eng. Appl. Artif. Intell. **118**, 105641 (2023)
30. Shi, H., Aryadoust, V.: A systematic review of AI-based automated written feedback research. ReCALL, pp. 1–23 (2024)
31. Sterbini, A., Temperini, M.: An exploration of open source small language models for automated assessment. In: 2024 28th International Conference Information Visualisation (IV), pp. 1–6. IEEE (2024)
32. Uta, M., et al.: Knowledge-based recommender systems: overview and research directions. Front. Big Data **7** (2024)
33. Vázquez-Madrigal, C., García-Rubio, N., Triguero, Á.: Generative artificial intelligence in education: risks and opportunities. In: Valls Martínez, M.D.C., Montero, J. (eds.) Teaching Innovations in Economics, pp. 233–254. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-72549-4_11
34. Yu, J., et al.: From MOOC to MAIC: reshaping online teaching and learning through LLM-driven agents (2024)
35. Yusuf, A., Pervin, N., Román-González, M.: Generative AI and the future of higher education: a threat to academic integrity or reformation? Evidence from multicultural perspectives. Int. J. Educ. Technol. High. Educ. **21**(1), 21 (2024)
36. Zawacki-Richter, O., Marín, V.I., Bond, M., Gouverneur, F.: Systematic review of research on artificial intelligence applications in higher education-where are the educators? Int. J. Educ. Technol. High. Educ. **16**(1), 1–27 (2019)
37. Zeng, J., Zhang, P., Zhou, J., Shang, J., Black, J.B.: The impact of embodied scaffolding sequences on stem conceptual learning. Educ. Technol. Res. Dev. 1–26 (2024)