LLM-based Literature Recommender System in Higher Education – A Case Study of Supervising Students' Term Papers

Xia Wang[1,*], Nghia Duong-Trung[1,2], Rahul R. Bhoyar[1] and Angelin Mary Jose[1]

[1]German Research Center for Artificial Intelligence (DFKI), Alt-Moabit 91 C, 10559, Berlin, Germany.
[2]IU International University of Applied Sciences, Frankfurter Allee 73A, 10247 Berlin, Germany.

Abstract

This paper presents the design and implementation of a Large Language Model (LLM)-based Literature Recommender System (LRS) to support students in higher education during the early stages of their term paper preparation. The system, named LRS4TP, provides personalized feedback and literature recommendations to help students formulate research topics and questions, thereby enhancing their critical thinking and research skills. Unlike existing AI-driven tools, LRS4TP focuses on inspiring students to explore diverse resources and refine their ideas through iterative feedback rather than automating the writing process. The paper outlines a case study conducted in a Bachelor of Arts program, where the recommender system assists students in developing term papers through a combination of natural language processing, sentiment analysis, and expert-based recommendations. Key challenges such as handling creative variations in student submissions, providing explainable AI recommendations, and ensuring system transparency are addressed. Initial evaluations suggest that LRS4TP reduces teacher workload while maintaining high-quality feedback, freeing up educators to provide more meaningful support. The paper concludes with insights into future developments for combining traditional recommendation techniques with LLM-based approaches to enhance learning in higher education contexts.

Keywords

Literature Recommender System, Large Language Models, Higher education, Term paper

1. Introduction

Exploring how to apply artificial intelligence (AI) technologies in daily teaching and learning in higher education [1, 2], this study focuses on a challenging and representative application case. It considers using a recommender system (RS) as an intelligent assistant to both students and teachers. The use case of this research project is to provide instructive, inspiring, and personalized feedback on initial ideas for the term papers (TPs) to be submitted by students. *Are the existing well-researched recommendation techniques [3] capable of meeting the needs of our use case?* And *what are the specific requirements of our use case that challenge current AI techniques?* These are the two aspects to be discussed first in this paper.

A detailed discussion of the term paper use case is presented in Section 3. In a nutshell, at the end of their last semesters, students begin preparing term papers by first submitting their initial ideas in the form of short texts. Such a text comprises one specified research topic (RT) and several related research questions (RQs). Then, there are multiple rounds of 1:1 discussions between a teacher and the student until a consensus is reached. During the discussions, the teacher evaluates the students' ideas and gives some inspiring feedback and recommendations to stimulate the students to think independently and deeply to develop the final ideas for the term paper.

GenAI-LA'25: International Workshop on Generative AI and Learning Analytics, 15th International Learning Analytics and Knowledge Conference (LAK'25), March 03–07, 2025, Dublin, Ireland.
*Corresponding author.
xia.wang@dfki.de (X. Wang); nghia_trung.duong@dfki.de (N. Duong-Trung); rahul_rajkumar.bhoyar@dfki.de (R. R. Bhoyar); angelin_mary.jose@dfki.de (A. M. Jose)
0000-0003-0849-5645 (X. Wang); 0000-0002-7402-4166 (N. Duong-Trung); 0009-0005-7249-7305 (R. R. Bhoyar); 0009-0005-5026-8990 (A. M. Jose)

The workload, time consumption, and instruction difficulties are obvious for teachers, and the 1:1 supervision through forum posts or emails is also very inefficient. A RS in the educational domain is defined as a context-bound combination of AI technologies and didactic design to provide recommendations to educational stakeholders [4]. Thus, part of our research investigates which recommenders are suitable to support students in generating individual term paper topics and research questions and to what extent. This is also a central challenge in higher education and a widespread issue in teaching.

Unlike some current applications that are purely based on large language models, which can directly generate long texts or entire papers, our research does not aim to assist students in any writing of their term papers but to motivate and inspire them to delve deeply and to read extensively to enhance their own learning and research abilities finally. Therefore, any recommendations provided at the end should not be definitive conclusions but pointers to additional resources for further reflection and contemplation. Moreover, a specific knowledge competency model in inquiry-based learning will be considered to explore and evaluate suitable AI methods for assisting students in finding topics and generating research questions for their term papers. After elaborating on our use case at the beginning of this paper, we use a term paper as an example to walk through the proposed LRS framework and explain the generated recommendations with a chain of thought.

2. Related Work

LLMs have revolutionized the field of Natural Language Processing (NLP) and have demonstrated their feasibility in a wide range of tasks such as dialogue generation, question answering, and text summarization [5, 6, 7], rendering them ideally suited to participate in the development of RS by the use of human-like dialogue [8, 9]. In the context of higher education, the integration of LRS has the potential to enhance the learning experience and to support students in their academic journey, such as course selection and planning [10], provision of personalized feedback and guidance in an online learning environment [11, 12]. Although focusing on different aspects, most existing RSs demonstrate the positive benefits of incorporating natural language dialogue into the recommendation process. However, factors such as integrating educational data from specific domains, personalizing recommendations based on learning profiles, and the ethical-related considerations with using AI-powered systems in educational settings still require further exploration.

A few challenges and limitations encountered when an LRS is integrated with LLMs are also addressed in our study. For instance, the first one is the phenomenon of 'hallucination', where language models produce outputs that sound plausible but are factually incorrect or not based on the input data [13, 14]. Next, we will address how to safeguard the output produced by an LLM. Moreover, according to [9] data-driven LLMs used for an RS may also pose severe threats to users and society [15, 16] due to unreliable decision-making, various biases, lack of transparency [17] and explainability [18], and privacy issues stemming from the extensive use of personal data for customization, among other concerns. Providing users with some transparency and explainability, similarly to [19], at both the data and algorithmic levels is also part of our work in this research (see 4.2).

More importantly, the research challenge in our use case goes beyond simple feedback; it necessitates expert recommendations or discussions that encourage deeper student thinking. Consequently, we reviewed the development of recommender systems in education [20, 21, 3]. For example, [20] analyzed 52 papers from 2019 to 2024, focusing on their techniques, models, datasets, and metrics. They found that generative AI models, such as generative adversarial networks (GANs), variational autoencoders (VAEs), and autoencoders, are widely used and outperform traditional AI methods. [21] examined 272 articles published between 2007 and 2021 in the Scopus database, identifying sixteen research themes, with a primary focus on e-learning, followed by classroom activities and course selection. [3] categorized various recommendation techniques, datasets, algorithms, similarity measurement methods, and evaluation metrics, which serve as key references for this work.

3. Use Case Description

In the final semester of a Bachelor of Arts program in Culture and Social Sciences, students must write their respective term papers based on what they have learned in several previous Media Education and Media Pedagogy courses. This semester is research-oriented and divided into three phases. The first is the preparation phase (see the left side of Fig. 1), in which students independently work on the course modules' learning material. Students are suggested to reflect on their learning with a short text summary, including answering questions about the learning content, their thematic interests, possible real-life cases, any confusion or contradiction, etc. This process can inspire students to form initial ideas for their term papers, especially on a research topic and related research questions. The result of this phase is a short text that defines and describes their choices of topics and research questions and that they are ready for discussion with teachers. Second, in the interaction phase (see the right side of Fig. 1), students intensively discuss their ideas with teachers and revise the ideas with evolutionary and iterative feedback until agreement is reached. The interactions between teachers and students are 1:1 tutoring via the Moodle forum. Finally, students can start the writing process independently (3rd phase), constituting their examination performance.

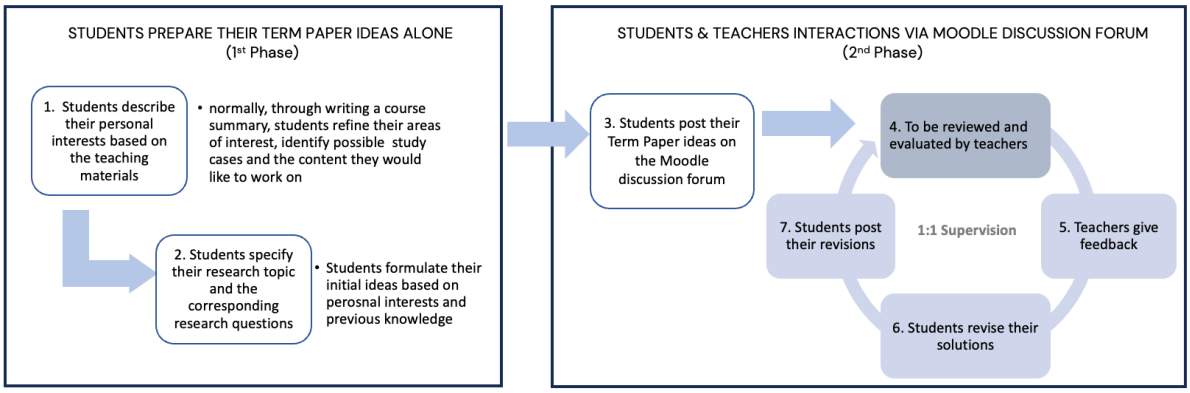


Figure 1: Use case illustration of preparing a Term Paper (TP).

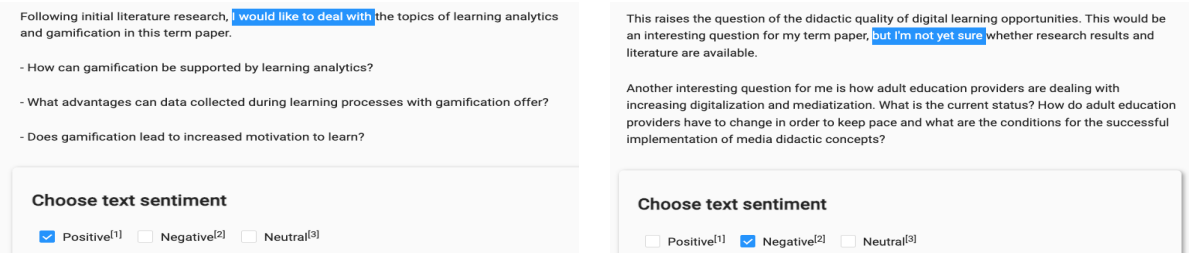
In such a use case, students have an exceptionally high need for personal support in formulating questions and finding/selecting topics for their term papers. Currently, the aforementioned interaction process consumes a great deal of the tutor's time to maintain high satisfaction with the student's 1:1 support. Besides, for teachers, the constant answering of recurring questions and constant feedback on repetitive mistakes made by students are detrimental and heavy burdens. This also leaves less time for research-promoting, stimulating interaction in 1:1 supervision.

The Moodle forum data previously collected from the two semesters of 2021 allowed some pre-analysis of 1:1 supervision: for approximately 70 students, there are, on average, some $13 \sim 15$ interactions with each teacher. Moreover, the feedback and recommendations collected from three teachers are also available for further analysis. Therefore, for this use case, we propose an RS to achieve the following goals,

- to provide high-quality instant and personal feedback and recommendations to students' term paper proposals, and to inspire them to work on their term papers more diligently.
- to address recurring questions and errors and to support students in their term paper preparation process.
- to free up instructor time and resources for more in-depth and substantial supervisory support of the students.

3.1. Use Case Example

Although not a single example can cover all the scenarios of the use case, here are two concrete examples from students (see Fig. 2) and a teacher's first feedback to the *Example A* with manual annotations (by using the labeling tool, named *Label Studio*, see Fig. 3). As shown, the student has proposed a research topic on "learning analytics and gamification" and planned to address three related questions inside the term paper later, e.g., "*How can gamification be supported by learning analytics?*", or "*Does gamification lead to increased motivation to learn?*".



(a) $Example_A$ on Task 3 (b) $Example_B$ on Task 3

Figure 2: Two term paper examples.

As shown in Fig. 3, except for the usual greetings and endings, we can also discover the following annotated inside the teacher's first feedback: i) the teacher approved that the student's research topic is exciting and well-founded; ii) the teacher specifically pointed out that the creative part lies in the plan "*to build a bridge between learning analytics and gamification*"; iii) a concrete recommendation to "*focus on one of the two focal points (here, for example, on the promotion of learning motivation)*". Usually, in some other cases (not shown in this example), teachers also give literature references as suggestions for further reading to inspire students to think deeply.

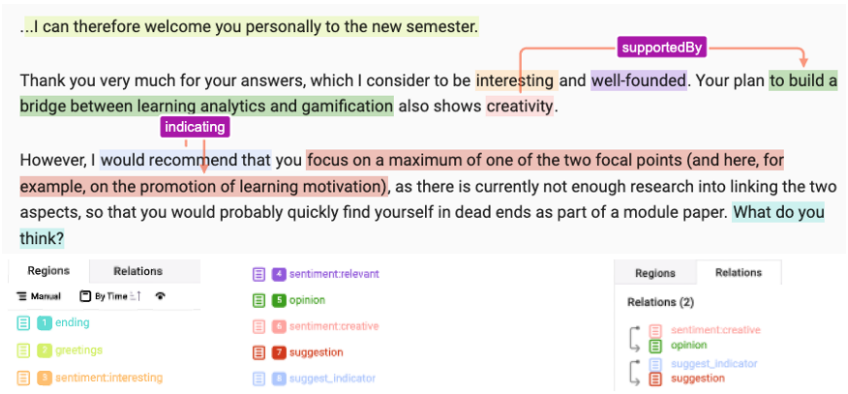


Figure 3: Annotated Teacher's 1st feedback to a term paper proposal.

3.2. Use Case Research Challenges

We selected this use case for our research project due to its complex questions and the innovative significance it represents. For instance, deep natural language understanding (NLU) in this case is essential for various high-level NLP tasks, including topic modeling, information retrieval, relation extraction, sentiment analysis, and argument mining. Given that student-teacher interactions occur through natural conversations, natural language generation (NLG) must be utilized extensively. Since late 2022, LLMs have showcased their capabilities in NLU and NLG, giving us confidence to address the challenges of this use case. Specifically, we extracted the following three research challenges,

RC1: Open-ended Recommendation. No specific and uniform item corpus is available for recommendation for this use case. Unlike recommending movies from the IMDB or rotten tomatoes database [22] or commodities from the Amazon Product dataset [23], individual recommendations to students' term papers are different and not uniform; basically, it is case-by-case. More than that, students' term papers are all different. Historical recommendations are difficult to use directly. Nonetheless, topic-based literature recommendation is our first step, as demonstrated in this paper.

RC2: Evaluating Human Creativity. Machine learning (ML), which is trained and learned from big and balanced datasets, typically works well [24, 25], while it becomes a challenge when it has to work with a small dataset, and the data contained do not have much in common. Generally speaking, students' term papers are different from each other over the years. Even if there are occasional submissions on similar research topics, their content or research questions should differ. How to evaluate a creative idea [26] is a critical issue. Although LLMs enhanced with transfer learning [27] or knowledge graphs with semantic inference [28] maybe two attempts to achieve solutions, this part of the work is not covered in this paper. Apart from this, other academic discussions on creativity, such as how ML affects human creativity [29] and human-machine creativity [30, 31], especially concerning writing, art, and music, have become increasingly valued academic research directions.

RC3: Explainability and Transparency. As an RS is to be widely used by students in universities, there is a specific demand for such a system's explainability and transparency. Despite the great success of Deep Neural Networks (DNNs) and many LLMs as black boxes, there is still no comprehensive theoretical understanding of their learning or inner organization [32, 33, 34]. Our study aims to reveal and visualize the RS to some extent regarding data and decision-making information, thereby increasing student acceptance of the advice generated. Students can learn what data are used to make recommendations and for what reason. Specifically, when integrating LLMs (e.g., GPT-4) into our recommender engine, we try to explain the generated information with a chain-of-thought (see Section. 4.2).

4. Literature Recommender System for Term Papers (*LRS4TP*)

4.1. Recommender System Framework

As shown in Fig. 4, the proposed LLM-based recommendation system inputs a student's term paper proposal. Then, it goes through several steps (more in Section 4.2) to generate personalized recommendations on the initial idea for the term paper, which is output in natural language. For various scenarios, it was considered to realize our system with a knowledge-based, expert-based, multi-criteria, profile-based, or hybrid recommendation engine. However, only the knowledge-based recommendation engine has currently been implemented and superficially evaluated.

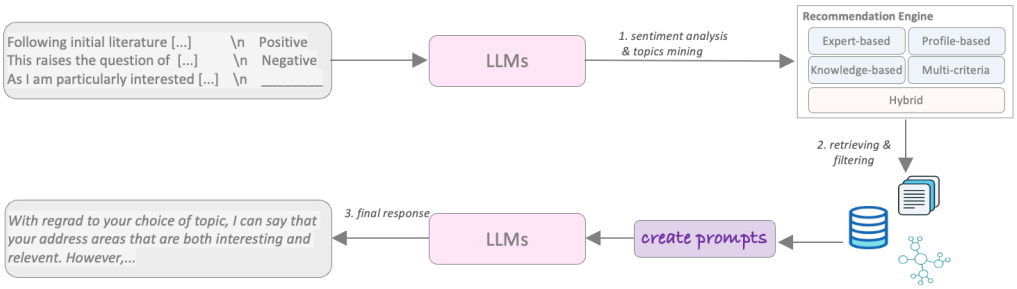
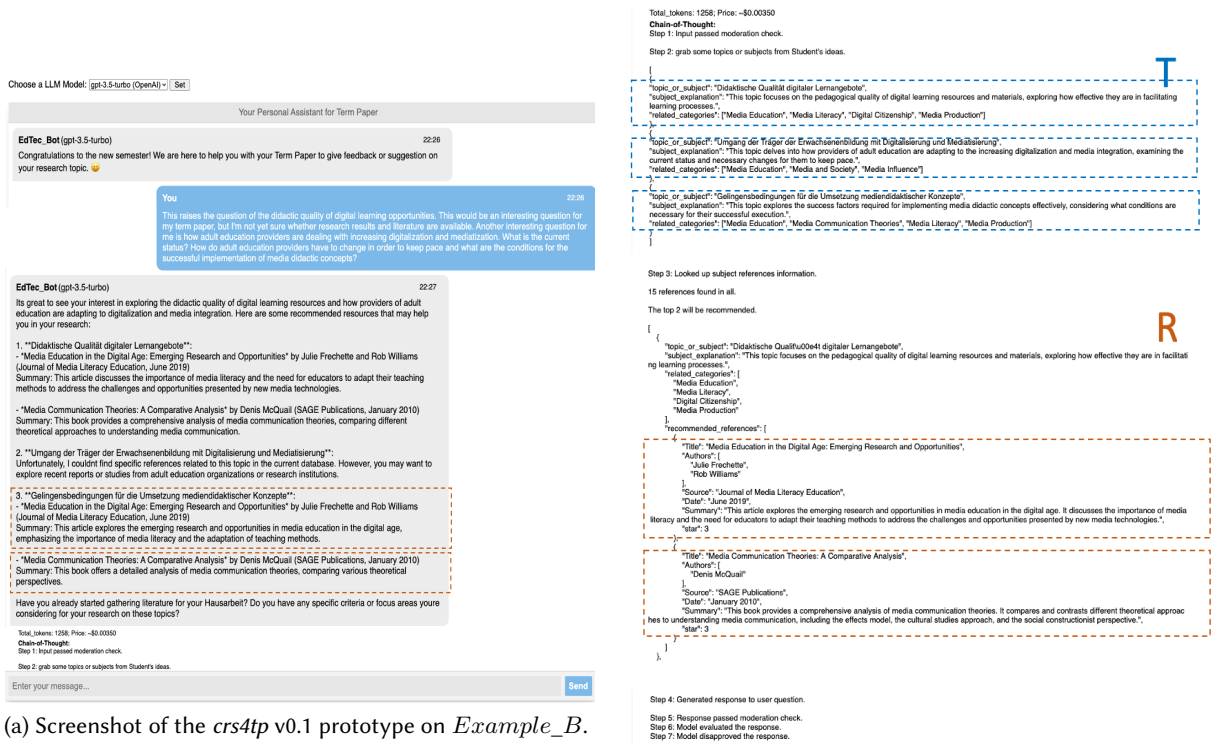


Figure 4: LLMs-enabled recommender system with In-Context-Learning.

The system involves three intermediate processes (Fig. 4). First, a sentiment analysis (SA) [35, 36] is conducted to determine the student's level of confidence level with the term paper proposal (similar as in Fig. 2). Distinctly, positive results are seen as the student being very confident in completing the topic. At this point, the system likely triggers the knowledge-based recommendation engine. On the contrary, with a negative result showing the student's lack of confidence and certainty, it is assumed

that the student pursues expert advice, and the expert-based recommendation engine tends to be triggered. Of course, students can specify their choices of recommendation engines, which are given the highest priority within the system. Second, the LLMs are tasked with summarizing and extracting the first k topics or subjects from the student's term paper proposal, which is how to understand the student's text content. For instance, the extracted topic set T is notated as $T = \{T_1, T_2, ..., T_k\}, k \in \mathbb{N}$ with $T_i = \{\text{topic}_i, \text{explanation}_i, C_i\}$ and $C_i = \{\text{category}_1, \text{category}_2, \ldots, \text{category}_j\}, j \in \mathbb{N}$, is the associated category list defined by the course modules. Moreover, $C_i \subseteq C$ and C is the whole list of concepts, theories, or knowledge areas retrieved from the course textbooks.



(a) Screenshot of the *crs4tp* v0.1 prototype on $Example\_B$.

(b) Chain-of-thoughts of *crs4tp* v0.1 in steps (clipped due to length)

Figure 5: Prototype of an LLM-based Conversational Recommender System for Term Papers.

Next, the In-Context Learning (ICL) approach and the triggered recommendation engine (e.g., knowledge-based) are applied to the topic set T to filter further topics, notated as T'. Then, based on the given categories and the linked topics in T', $crs4tp$ starts to search in the pre-prepared external resource corpus (RC), which is the recommendation item corpus, to retrieve a list of literature references as results, notated as $R = \{(T_1, R_1), (T_1, R_2), (T_i, R_i), \ldots, (T_u, R_v)\}$, where one reference R_m can correspond to multiple topics T_n. The resource corpus is supposed to be generated from the course textbooks and the domain knowledge base, which contains domain concepts, categories, and related literature. Finally, the above results are used as information to create prompts for the LLMs, which generate the final response for the student as feedback. This step utilizes the LLMs to transform the retrieved literature list into a text in natural language for students.

4.2. Chain-of-Thought (CoT)

To understand or evaluate the recommendations generated, $LRS4TP$ also provides students with certain explanations by giving the chain of thought, making the back-end operating mechanism of this recommendation system more transparent. For instance, the left figure of Fig. 5 shows the first demonstration of *LRS4TP* v0.1 with the GPT-4 model by a chatbot named $EdTec\_bot$. It converses naturally and is much more human-like than the traditional rule-based or scripted-based chatbots. The

right figure of Fig. 5 presents the CoT of $Example_B$ behind the scenes. Specifically, upon receiving this term paper proposal, the first step, $Step$ 1, is to have a moderation check with the OpenAI endpoint to filter any potentially harmful or inappropriate requests. In $Step$2, the LLM extracts the main topics from the term paper, resulting in a topic set T comprising three topics, each with a brief explanation and the specific categories it belongs to. For instance, the topic *Didaktische Qualität digitaler Lernangebote* focuses on the pedagogical quality of digital learning resources and is categorized under "Media Education", "Media Literacy", "Digital Citizenship", and "Media Production". In $Step$3, an external static resource corpus is searched to identify each topic's top two literature references, forming the result set R (refer to the orange box). In this instance, 15 literature references were discovered. The LLM produces the final response in natural language in response to a prompt created using the result set R. Then, in $Step$5, the LLM performs another moderation check before providing feedback to the student. By deploying CoT, we are able to check how LLMs behaviors to migrate any possible issues of ethical risks.

5. Experiments Setup and Proof of Concepts

5.1. Reference Collection and Sorting Algorithm

This section provides an overview of the key statistics related to the concepts and their corresponding references. As discussed in section 3, the idea is to provide students with suitable reading materials for their term papers. The left table of the Fig. 6 shows the distribution of references across different educational concepts. The right side algorithm inside the Fig. 6 presents the core idea of star-based references to decide the relevant literature.

Concept	# of References	Percentage
Media Education	10	7.75%
Media Communication Theories	10	7.75%
Mass Communication	10	7.75%
Interpersonal Communication	13	10.07%
Media Literacy	10	7.75%
Digital Citizenship	10	7.75%
Media and Society	11	8.52%
Media Production	11	8.52%
Media Ethics	12	9.30%
Media Law	15	11.62%
Media Regulation	17	13.22%
Total	**129**	**100%**

Algorithm 1 Find References for the proposal of a Term Paper *TP*

Require: a set of topics or concepts of a *TP* as $TP_C = \{C_1, C_2, ..., C_i\}$, and the references corpus consisting of a set of concepts and their semantic linked references, as $CR = (C_j, \{R_{j1}, R_{j2}, \ldots, R_{jk}\})$, $i, j, k \in \mathbb{N}$

Ensure: a sorted list of references $\{R_1, R_2, ..., R_m\}$, $m \in \mathbb{N}$

```
Initialize an empty list reference_list, star = 1
for every star number of concepts of TP_C do
    reference_list = search for the references that link to all of the star numbers of the concepts
    for each item in reference_list do
        if item['Title'] is not in [re['Title'] for re in reference_list] then
            Set item['star'] to star
            Append item to reference_list
        end if
    end for
    star ++
end for
Sort reference_list by star value in descending order
return sorted_list, length of reference_list
```

Figure 6: Distribution of References by Concept (left) and Literature Selection Algorithm (right).

5.2. Expert Evaluation

To evaluate our system's initial performance, we conducted a pilot test using 56 term paper examples from the previous semester. The system generated a list of literature recommendations that were available in the university's library. Two independent experts assessed the quality of these recommendations for four hours. The experts evaluated the system on four critical criteria: (i) Relevance to Content: The degree to which the recommended references were relevant to the term paper's main topics; (ii) Accuracy of Citations: Whether the citations were correctly formatted and included all necessary information; (iii) Consistency in Citation Style: Whether the citations followed a consistent citation style throughout; and (iv) Overall Usefulness: How useful the recommended references were in providing a solid starting point for further research on the student's topic. Each expert rated the system on a scale of 1 (Strongly Disagree) to 5 (Strongly Agree) for each criterion. These ratings were then compared to assess the system's performance and the consistency between the two evaluators. Figure 7 illustrates the experts' ratings across the four criteria.

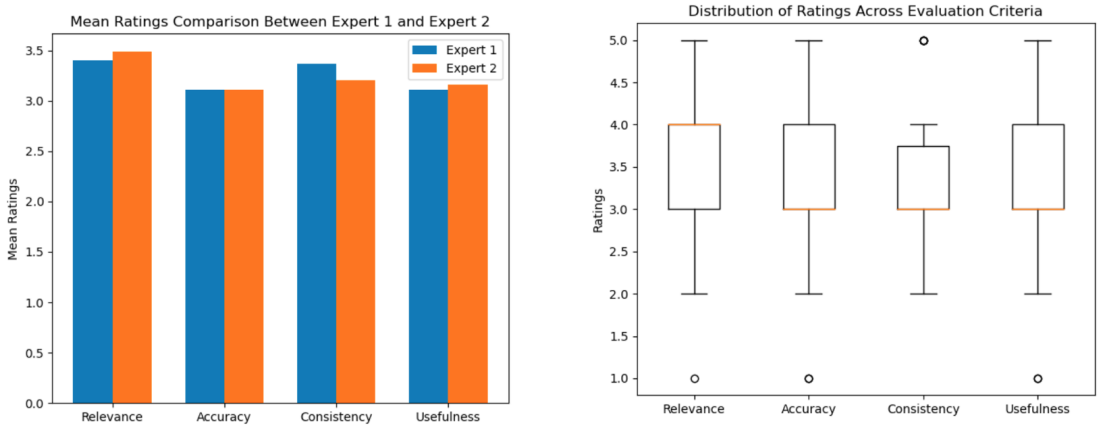


Figure 7: Rating comparison between two experts (left), and distribution of ratings across criteria (right).

Additionally, we used Cohen's Kappa [37] to quantify the inter-rater reliability for each criterion. The results were as follows: (i) Relevance to Content: 0.5105 (substantial agreement); (ii) Accuracy of Citations: 0.4746 (moderate agreement); (iii) Consistency in Citation Style: 0.3805 (fair agreement); and (iv) Overall Usefulness: 0.4683 (moderate agreement). These results indicate substantial agreement on the relevance of content but only moderate to fair agreement on the other criteria, with the lowest agreement on citation style consistency.

5.3. Remarks and Discussion

The initial inter-rater reliability scores and average expert ratings across the four evaluation criteria suggest that our system has the potential for effective integration and further development. The average relevance score is approximately 3.5 indicates that the recommendation algorithm using the "star" method is functional. However, the average accuracy rating of 3.0 highlights a significant limitation: the "star" method alone is insufficient to leverage the full power of LLMs, particularly when comparing the semantic meanings between student proposals and literature content. Another observation is that we found several pieces of literature recommended with very high frequency for term papers of different students on various topics. The reason for this, most likely, is that the semantic links between the available literature and the term paper topics are not very specialized; of course, it is not rule out the fact of the presence of popularity bias in the data set. Although the current experiments, from the proof of concept aspect, demonstrate the feasibility of our work, the larger as much as possible dataset is critical, and more domain experts in the field of teaching are needed to validate the semantic links. Moreover, to address this in future work, we propose converting term papers and literature references into latent semantic embedding and using semantic mining to refine the recommendation process.

6. Conclusions

This paper presented a detailed case study of dealing with students preparing term papers in higher education. It explores using LLMs to automatically provide students with generative natural language feedback and recommendations by an LRS. Additionally, this paper demonstrated an RS prototype integrated with LLMs (i.e., GPT-4), showing the specific application of LLMs in various aspects, including sentiment analysis, topic mining, natural language understanding, and answer generation. Through the experiments, we believe integrating large language models into traditional recommendation systems is essential and has significant positive implications. Our future work aims to combine adapted traditional recommendation technologies with large language models to develop a conversational recommendation system as an intelligent assistant for students and teachers. The new release of $LRS4TP$ v0.2 is ready for demonstration and deployment in our university's library services in the coming semesters.

Acknowledgments

The authors kindly appreciate the support of CATALPA, FernUniversität in Hagen by the "AI.EDU Research Lab 2.0" Project.

Declaration on Generative AI

Generative AI tools were not used when preparing the manuscript.

References

[1] W. Xu, F. Ouyang, The application of ai technologies in stem education: a systematic review from 2011 to 2021, International Journal of STEM Education 9 (2022) 59.
[2] H. Crompton, D. Burke, Artificial intelligence in higher education: the state of the field, International Journal of Educational Technology in Higher Education 20 (2023) 22.
[3] J. Joy, R. V. G. Pillai, Review and classification of content recommenders in e-learning environment, Journal of King Saud University - Computer and Information Sciences 34 (2022) 7670–7685. doi:https://doi.org/10.1016/j.jksuci.2021.06.009.
[4] H. Drachsler, K. Verbert, O. C. Santos, N. Manouselis, Panorama of Recommender Systems to Support Learning, Springer US, Boston, MA, 2015, pp. 421–451.
[5] OpenAI, Gpt-4 technical report, 2024. arXiv:2303.08774.
[6] L. Qin, Q. Chen, X. Feng, Y. Wu, Y. Zhang, Y. Li, M. Li, W. Che, P. Yu, Large language models meet nlp: A survey, 2024. doi:10.48550/arXiv.2405.12819.
[7] S. Vatsal, H. Dubey, A survey of prompt engineering methods in large language models for different nlp tasks, 2024. doi:10.48550/arXiv.2407.12994.
[8] A. Vats, V. Jain, R. Raja, A. Chadha, Exploring the impact of large language models on recommender systems: An extensive review, arXiv preprint arXiv:2402.18590 (2024).
[9] Z. Zhao, W. Fan, J. Li, Y. Liu, X. Mei, Y. Wang, Z. Wen, F. Wang, X. Zhao, J. Tang, et al., Recommender systems in the era of large language models (llms), IEEE Transactions on Knowledge and Data Engineering (2024).
[10] J. Yu, Z. Zhang, D. Zhang-li, S. Tu, Z. Hao, R. Li, H. Li, Y. Wang, H. Li, L. Gong, J. Cao, J. Lin, J. Zhou, F. Qin, H. Wang, J. Jiang, L. Deng, Y. Zhan, C. Xiao, M. Sun, From mooc to maic: Reshaping online teaching and learning through llm-driven agents, 2024. doi:10.48550/arXiv.2409.03512.
[11] C. Ng, Y. Fung, Educational personalized learning path planning with large language models, 2024. doi:10.48550/arXiv.2407.11773.
[12] M. Abel, W. Germain, T. Mahatody, Pedagogical alignment of large language models (llm) for personalized learning : A survey, trends and challenges (2024).
[13] P. Manakul, A. Liusie, M. J. F. Gales, Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models, 2023. arXiv:2303.08896.
[14] N. McKenna, T. Li, L. Cheng, M. J. Hosseini, M. Johnson, M. Steedman, Sources of hallucination by large language models on inference tasks, 2023. arXiv:2305.14552.
[15] W. Fan, X. Zhao, X. Chen, J. Su, J. Gao, L. Wang, Q. Liu, Y. Wang, H. Xu, L. Chen, Q. Li, A comprehensive survey on trustworthy recommender systems, 2022. arXiv:2209.10117.
[16] T. Y. Zhuo, Y. Huang, C. Chen, Z. Xing, Exploring ai ethics of chatgpt: A diagnostic analysis, ArXiv abs/2301.12867 (2023). URL: https://api.semanticscholar.org/CorpusID:256390238.
[17] T. South, R. Mahari, A. Pentland, Transparency by design for large language models, Computational Legal Futures, Network Law Review. (2023).
[18] Y. Hu, N. Giacaman, C. Donald, Enhancing trust in generative ai: Investigating explainability of llms to analyse confusion in mooc discussions (2023).
[19] H. Abu-Rasheed, M. H. Abdulsalam, C. Weber, M. Fathi, Supporting student decisions on learning

recommendations: An llm-based chatbot with knowledge graph contextualization for conversational explainability and mentoring, arXiv preprint arXiv:2401.08517 (2024).
[20] M. O. Ayemowa, R. Ibrahim, M. M. Khan, Analysis of recommender system using generative artificial intelligence: A systematic literature review, IEEE Access PP (2024) 1–1. doi:10.1109/ACCESS.2024.3416962.
[21] M. H. M. NOR, Educational recommender systems: A bibliometric analysis for the period 2002–2022, Journal of Quality Measurement and Analysis JQMA 20 (2024) 197–215.
[22] B. M. G. Al Awienoor, E. B. Setiawan, Movie recommendation system based on tweets using switching hybrid filtering with recurrent neural network., International Journal of Intelligent Engineering & Systems 17 (2024).
[23] S. Rajput, N. Mehta, A. Singh, R. Hulikal Keshavan, T. Vu, L. Heldt, L. Hong, Y. Tay, V. Tran, J. Samost, M. Kula, E. Chi, M. Sathiamoorthy, Recommender systems with generative retrieval, in: A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, S. Levine (Eds.), Advances in Neural Information Processing Systems, volume 36, Curran Associates, Inc., 2023, pp. 10299–10315.
[24] J. Rodrigues, G. Vasconcelos, Big data machine learning benchmark on spark, 2019. doi:10.21227/t8bg-yc46.
[25] A. K. Badhan, A. Bhattacherjee, R. Roy, Deep Learning Techniques in Big Data Analytics, Springer Nature Singapore, Singapore, 2024, pp. 171–193. doi:10.1007/978-981-97-0448-4_9.
[26] O. M. Kleinmintz, T. Ivancovsky, S. G. Shamay-Tsoory, The two-fold model of creativity: the neural underpinnings of the generation and evaluation of creative ideas, Current Opinion in Behavioral Sciences 27 (2019) 131–138.
[27] M. Patidar, A. Singh, R. Sawhney, I. Bhattacharya, et al., Combining transfer learning with in-context learning using blackbox llms for zero-shot knowledge base question answering, arXiv preprint arXiv:2311.08894 (2023).
[28] X. Liu, T. Mao, Y. Shi, Y. Ren, Overview of knowledge reasoning for knowledge graph, Neurocomputing (2024) 127571.
[29] M. Farina, A. Lavazza, G. Sartori, W. Pedrycz, Machine learning in human creativity: status and perspectives, AI & SOCIETY (2024). doi:10.1007/s00146-023-01836-5.
[30] D. Dwivedi, G. Mahanty, Human creativity vs. machine creativity: Innovations and challenges, in: Multidisciplinary Approaches in AI, Creativity, Innovation, and Green Collaboration, IGI Global, 2023, pp. 19–28.
[31] M. D. Mumford, D. C. Lonergan, G. Scott, Evaluating creative ideas: Processes, standards, and context, Inquiry: Critical thinking across the disciplines 22 (2002) 21–30.
[32] R. Shwartz-Ziv, N. Tishby, Opening the black box of deep neural networks via information, arXiv preprint arXiv:1703.00810 (2017).
[33] Z. Chen, J. Chen, M. Gaidhani, A. Singh, M. Sra, Xplainllm: A qa explanation dataset for understanding llm decision-making, arXiv preprint arXiv:2311.08614 (2023).
[34] H. Zhao, H. Chen, F. Yang, N. Liu, H. Deng, H. Cai, S. Wang, D. Yin, M. Du, Explainability for large language models: A survey, 2023. arXiv:2309.01029.
[35] J. Chun, K. Elkins, explainable ai with gpt4 for story analysis and generation: A novel framework for diachronic sentiment analysis, International Journal of Digital Humanities (2023). doi:10.1007/s42803-023-00069-8.
[36] K. Kheiri, H. Karimi, Sentimentgpt: Exploiting gpt for advanced sentiment analysis and its departure from current machine learning, 2023. arXiv:2307.10234.
[37] K. L. Gwet, Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters, Advanced Analytics, LLC, 2014.