Retrieval-Augmented Chatbots for Scalable Educational Support in Higher Education

Hassan Soliman[1], Hitesh Kotte[1], Miloš Kravčík[1], Norbert Pengel[2] and Nghia Duong-Trung[1,3]

[1]German Research Center for Artificial Intelligence (DFKI), Alt-Moabit 91C, 10559 Berlin, Germany.
[2]Leipzig University, Dittrichring 5-7, 04109 Leipzig, Germany.
[3]IU International University of Applied Sciences, Frankfurter Allee 73A, 10247 Berlin, Germany.

Abstract

Students of educational sciences participate in learning activities, where appropriate support and timely feedback are crucial. However, providing scalable, personalized, and timely support becomes a major challenge. This work focuses on developing a didactic chatbot based on a Large Language Model (LLM) and enhancing its potential with existing learning materials. Retrieval Augmented Generation (RAG) allows the system to provide comprehensive, context-aware answers to specific course questions. Previous results suggested that it is possible to distinguish between different contexts in which students work and provide them with prompt responses that consider the relevant material. This paper presents insights from the technical implementation and the first results on the quality of LLM-based chatbot responses to content and organizational questions in an educational science module for student teachers. We compare previous automated evaluations using GPT-4 with newly conducted human evaluations of chatbot-generated results. Our experimentation demonstrated that the chatbot could achieve the highest correct response rate of 87%. Furthermore, human evaluations conducted by five expert annotators assessed the chatbot's responses. The agreement between the majority vote of these human judges and the GPT-4 evaluation showed substantial alignment. This study helps to demonstrate the potential of generative AI in the delivery of digitally supported courses.

Keywords

Large Language Model, Chatbot, Retrieval-Augmented Generation, Scalable Mentoring, Higher Education

1. Introduction

Providing individualised assistance and prompt feedback through scalable mentoring is a major educational challenge. However, new opportunities in digital higher education are being made possible by the rapid development of computing technologies, especially Artificial Intelligence (AI). Our goal is to enhance the student learning experience by designing a chatbot that allows for more flexible and adaptable responses. While previous rule-based systems struggled with adaptability and were limited to template-driven interactions, LLMs offer the opportunity to generate more nuanced and contextually aware responses, addressing the dynamic needs of students and accommodating a wide range of inquiries, from course content to organizational matters.

In educational science modules and teacher training programs, students benefit from receiving context-aware responses that address their specific learning needs. LLMs can potentially analyze existing learning and information materials and process descriptions (e.g., mentoring structure or feedback systems) to generate responses that go beyond static, predefined answers. This leads to our central research question: How can an LLM-enhanced chatbot be designed, implemented, and evaluated to support scalable educational support in higher education? Our focus is on applying LLMs in a didactically meaningful way to provide students with personalized and contextualized responses via a web-based interface, thereby promoting self-regulated learning and facilitating mentoring experiences.

Our paper presents the conceptual foundation, design, and implementation status of this iterative process, with a particular focus on the technological aspects, such as chatbot design and LLM integration.

Second International Workshop on Generative AI for Learning Analytics, 2025
hassan.soliman@dfki.de (H. Soliman); hitesh.kotte@dfki.de (H. Kotte); milos.kravcik@dfki.de (M. Kravčík); norbert.pengel@uni-leipzig.de (N. Pengel); nghia_trung.duong@dfki.de (N. Duong-Trung)
0009-0003-4574-9074 (H. Soliman); 0009-0005-8885-889X (H. Kotte); 0000-0003-1224-1250 (M. Kravčík); 0000-0002-3263-6877 (N. Pengel); 0000-0002-7402-4166 (N. Duong-Trung)

In the subsequent sections, we first discuss related work and explain the pedagogical context. The main section presents the technical background, including designing and implementing the LLM-based chatbot prototype. The paper then moves on to the experimental results, discussing the chatbot's performance based on human evaluations and automated assessments. Finally, we conclude with insights from these outcomes and propose future directions to enhance the chatbot's capabilities further.

2. Related Work

AI has become integral to education, offering solutions for students, teachers, and administrators. By analyzing extensive data from these groups, AI enhances personalized learning, optimizes administrative processes, and provides insightful feedback [1]. Among generative AI technologies, LLMs are particularly impactful, enabling human-like text generation and interactive educational tools [2]. These models underpin sophisticated educational chatbots that engage in meaningful conversations as teachers, learners, guides, or mentors [3].

Early educational chatbots relied on rule-based or template-driven systems, using predefined responses and basic Natural Language Understanding (NLU) techniques to interact with users [4]. For example, chatbots built with the RASA framework[1] utilized NLU models to classify user intents and recognize entities. However, these approaches suffered from limited flexibility and contextual awareness, leading to rigid responses that struggled with dynamic or complex queries, resulting in user frustration and reduced engagement [5]. To address these limitations, recent research has leveraged LLMs to develop more adaptable and context-aware conversational agents. LLM-based chatbots generate nuanced and relevant responses by understanding user context and intent, thereby enhancing the overall user experience [3]. This shift represents a significant advancement in educational chatbots' ability to support deeper and more meaningful student interactions.

RAG approaches have recently emerged as a promising solution to enhance educational chatbot performance by combining traditional document retrieval with LLMs' generative capabilities, resulting in more informed and contextually relevant responses [6]. This hybrid method allows chatbots to utilize extensive educational content repositories, ensuring coherent and accurate information. In [7], AI-powered chatbots were used to scale mentoring support in higher education, providing 24/7 assistance, answering FAQs, and offering personalized feedback. Further research implemented chatbots in large-scale settings with over 700 students [5], showing that chatbots significantly supported self-study and alleviated traditional mentoring resource constraints. Additionally, [6] introduced a RAG approach for academic environments, demonstrating that integrating document retrieval with LLMs enhances information access efficiency and relevance, thereby creating more effective educational assistants. Similarly, [8] developed MoodleBot, an LLM-driven chatbot integrated into the Moodle Learning Management System (LMS) to support self-regulated learning. The study involving 46 students revealed an 88% accuracy rate in course-related assistance and positive student acceptance, highlighting LLM-based chatbots' potential to enhance higher education despite challenges like bias, hallucinations, and resistance to AI technologies.

Despite advancements, deploying retrieval-augmented chatbots in education faces several challenges. Organizational and pedagogical issues, such as ensuring data privacy, maintaining information quality, and aligning chatbot responses with educational objectives, are critical [7]. Additionally, scaling these systems across diverse educational environments and adapting to various instructional styles and curricula remain ongoing challenges. Moreover, as highlighted in [8], LLM-driven chatbots must ensure response accuracy, manage potential biases and hallucinations, and overcome educators' resistance to new AI technologies. The study underscores the need for robust fact-checking mechanisms and alignment of chatbot responses with course content to preserve educational integrity. Keeping indexed materials current and reflective of course content is essential for maintaining chatbot reliability and effectiveness.

[1] https://rasa.com/

The effectiveness of RAG-based chatbots fundamentally depends on the quality of their retrieval processes. In our earlier work [9] we utilized basic RAG techniques to facilitate information retrieval. Building upon this prototype, we further refined and expanded our approach [10], experimenting with a curated evaluation dataset and introducing hybrid ensemble retrievers that combine different methods (e.g., keyword-based and semantic similarity searches) to optimize the retrieval of relevant information from large datasets. This enabled more accurate retrieval of relevant content from course materials, improving the chatbot's overall performance. However, we identified limitations in response relevance and depth, which prompted us to explore more advanced methods to enhance performance.

To overcome these challenges, we incorporated reranker models into the retrieval pipeline. Rerankers analyze the initially retrieved chunks and reorder them based on their relevance to the user's query, significantly improving the precision of the retrieved context and enabling more accurate, contextually relevant responses. Additionally, we conducted extensive evaluations to assess the impact of reranker models. Moreover, we compared the chatbot's performance using automated GPT-4 evaluations with human evaluations by domain experts. These human evaluations provided insights into the agreement and discrepancies between human judgment and machine assessments, essential for understanding the potential and limitations of LLM-based chatbots in education. Overall, integrating reranker models and a comprehensive evaluation approach represents a significant advancement in developing scalable, intelligent chatbots. By enhancing retrieval precision and thoroughly assessing chatbot performance, we advance the creation of more effective and reliable educational tools.

3. Design and Implementation

From a didactic perspective, we focus on self-regulated learning, mentoring, and counseling, with mentoring being an effective way to support learning [11]. For example, supporting students in an advisory capacity helps clarify problematic situations. A dyadic mentor-mentee relationship [12] is ideal but often unattainable due to limited resources, posing the challenge of scaling mentoring processes. This requires an integrated environment with various facilities, where the chatbot serves as a permanent virtual contact [13], fulfilling dual roles as an expert and learning companion [14]. As an expert, the chatbot answers questions about course content and organization. As a learning companion and mentor, it supports the individual learning process with feedback on submitted writing tasks, encouraging students to plan, monitor, and reflect on their learning. The BiWi (course acronym) AI Tutor addresses the scalability challenge by primarily acting as an expert on course material and answering students' questions about content and organizational information. The latest version of the chatbot includes the mentoring module (psychosocial support), which was not available in our evaluations.

3.1. LLM Based Prototype

The BiWi AI Tutor, an LLM-based chatbot prototype, provides scalable learning support by retrieving knowledge from lecture slides, seminar texts, and organizational materials for a German-taught university-level Education Science course. Utilizing the GPT-3.5-turbo model from OpenAI[2] and the LangChain[3] library, the chatbot offers responsive, contextually aware dialogic interactions (see Figure 1). Based on LangChain's Function Calling Agent, it dynamically selects relevant tools or data according to contextual needs. It determines when to select tools and appropriate context materials for a query, feeding the results back to the agent to determine subsequent steps. This iterative loop enables dynamic, context-sensitive interactions and handles multi-question queries.

The chatbot's retriever mechanism is crucial for selecting the most relevant learning material for a given query. For instance, the query "What are the main points of lecture 1 and lecture 3?" is split into two sub-queries: "main points of lecture 1" and "main points of lecture 3" as shown in Figure 2 (originally in German). This allows the retrieval of the most pertinent material chunks for each sub-query. Using

[2] https://platform.openai.com/docs/models/gpt-3-5-turbo
[3] https://www.langchain.com/

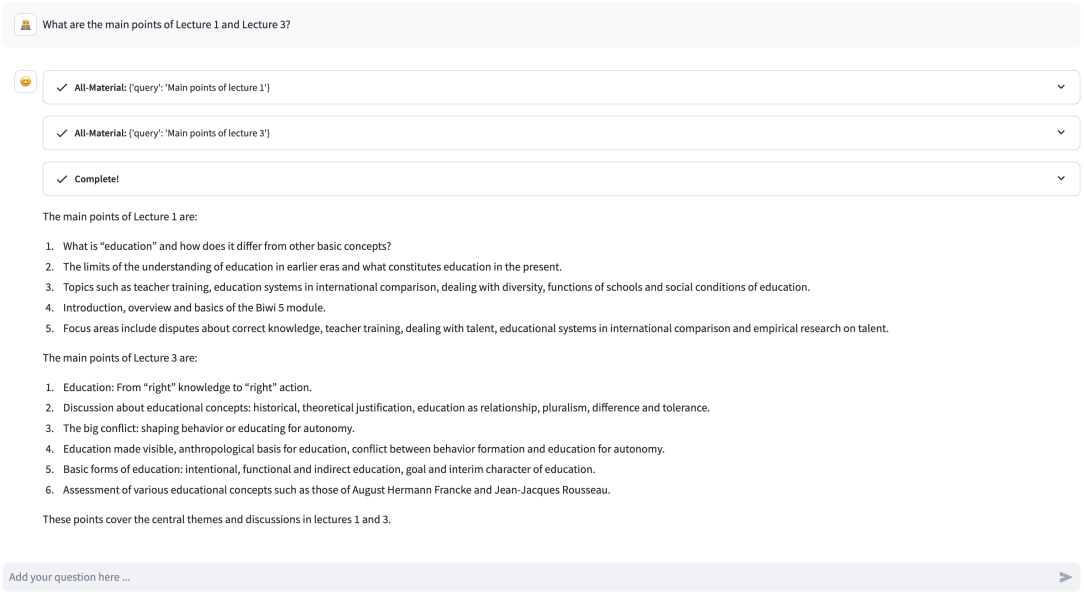

Figure 1: The BiWi AI Tutor demo chat interface (originally in German).

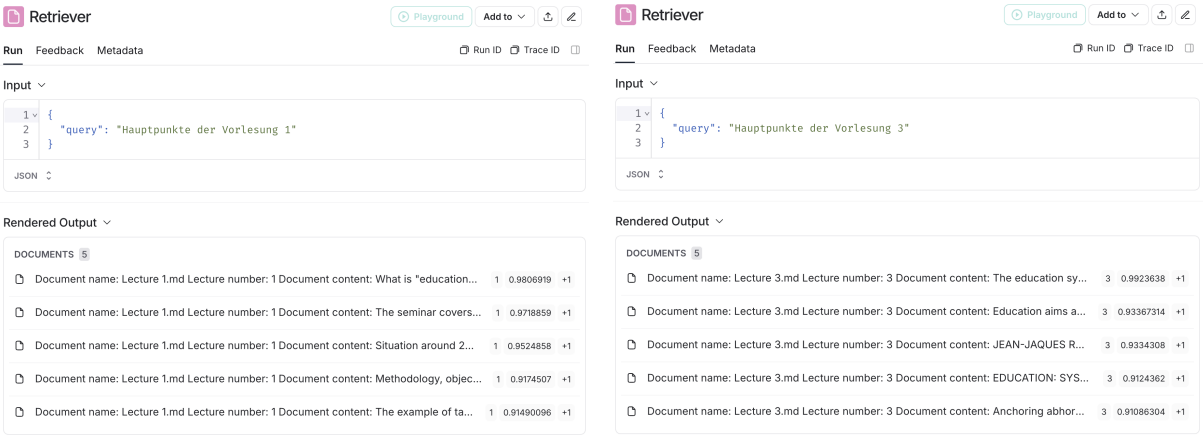

Figure 2: Retriever extracts chunks with similarity scores from both lecture 1 and lecture 3 (originally in German).

the LangSmith[4] library for observability, the retrieved chunks for each sub-query can be displayed and they are ranked by relevance. The agent then formulates a comprehensive final answer by combining the retrieved materials. The retriever employs semantic similarity and keyword matching to locate relevant content in real-time, streamlining content selection for user queries. This integration of an LLM's reasoning and generative capabilities with retrieval systems efficiently locates relevant learning content, a process known as RAG in the literature.

3.2. Learning Material Indexing and Chatbot Interaction Flow

To enable the chatbot to provide accurate and contextually relevant answers, we implemented a comprehensive indexing and retrieval process of the course materials, alongside a well-defined interaction flow for the chatbot. These processes are illustrated in Figure 3 and Figure 4. The steps in Figure 3 serve as the foundational processes that support the interaction flow in Figure 4. In addition to the indexing and interaction processes, we employed a basic system prompt. This prompt defines the chatbot's role as a tutor for the course, explains to it the course materials, and directs it to answer students'

[4] https://smith.langchain.com/

questions in German. This setup ensures that the chatbot maintains a consistent and helpful persona while interacting with students.

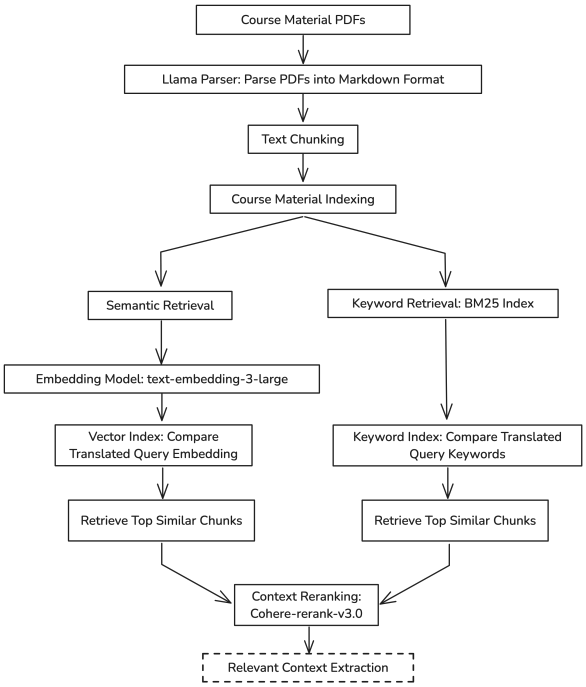

Figure 3: Indexing and Retrieval Process.

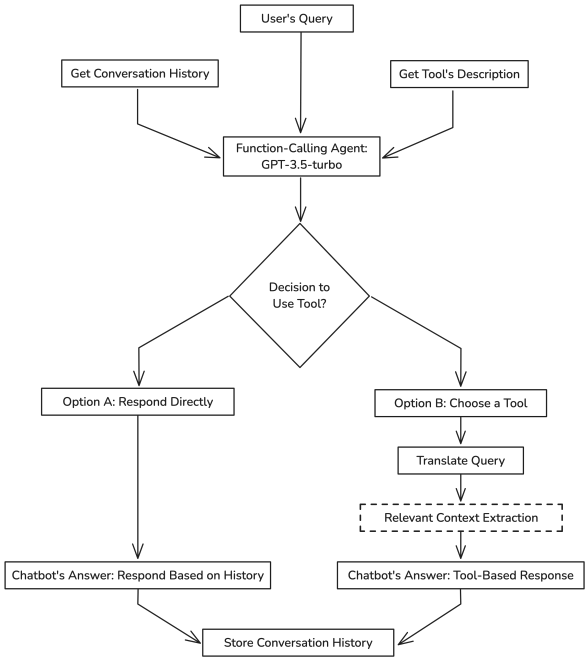

Figure 4: Chatbot Interaction Flow.

3.2.1. Learning Material Indexing and Retrieval Process

The indexing and retrieval process involves the following steps:

1. **Indexing Course Material**: Collect and prepare course materials, including lectures, seminar PDFs, and organizational PDFs, for indexing.
2. **Text Parsing**: Utilize the Llama Parser module from the LlamaIndex[5] library to parse PDF files into structured formats like Markdown, simplifying processing and enhancing compatibility with LLMs.
3. **Text Chunking**: The parsed text is divided into manageable chunks of 1024 tokens with an overlap of 20 tokens. This choice was based on preliminary experiments that demonstrated 1024 tokens provided an optimal balance between maintaining sufficient context and ensuring processing efficiency. The 20-token overlap helps preserve continuity between chunks, reducing the likelihood of losing critical contextual information that spans chunk boundaries.
4. **Course Material Indexing**: Organize chunks into:
   a) **Vector Index**: Generate embeddings using OpenAI's ("text-embedding-3-large") model[6] for semantic retrieval, storing them in a vector database.
   b) **BM25 Index**: Apply the BM25 algorithm for keyword-based retrieval based on term frequency and inverse document frequency.
5. **Query Translation**: Preprocess and translate user queries into a suitable format for retrieval systems, potentially involving language translation or keyword extraction.
6. **Semantic Retrieval**: Embed the query using the same embedding model to create a vector representation and retrieve semantically similar chunks from the Vector Index.
7. **Keyword Retrieval**: Use the BM25 Index to extract chunks containing relevant keywords from the query.

[5] https://www.llamaindex.ai/
[6] https://platform.openai.com/docs/guides/embeddings/embedding-models

Table 1
Learning Material Statistics

Material Source	Seminar	Lecture	Organizational
Number of PDF files	73	12	1
Number of Tokens	2,428,520	153,182	5,447
Number of Chunks	3,248	212	20

8. **Retrieve Top Similar Chunks**: Combine the top 50 chunks from both semantic and keyword retrievals, totaling 100 candidate chunks.
9. **Context Reranking**: To refine the retrieved context, we employ a reranker model from Cohere[7] ("cohere-rerank-v3.0"). The reranker re-evaluates the 100 candidate chunks based on their relevance to the query and selects the top 5 most relevant chunks.
10. **Relevant Context Extraction**: Utilize the top-ranked chunks as the relevant context for generating the chatbot's response.

This indexing and retrieval process ensures that the chatbot accesses the most pertinent sections of the course material, enabling accurate and context-aware answers. Statistics for the learning material are shown in Table 1.

3.2.2. Chatbot Interaction Flow

Building upon the indexing and retrieval process, the chatbot's interaction flow is designed for a seamless and contextually rich user experience, as illustrated in Figure 4. The interaction flow involves the following steps:

1. **User's Question**: The user submits a question to the chatbot, e.g., "Wann ist die Klausur?" (When is the exam?).
2. **Retrieve Conversation History**: Retrieve the past $k = 10$ messages from the conversation history to provide context.
3. **Get Tool's Description**: Consider descriptions of available tools (e.g., the course material tool) to decide their usage in the response.
4. **Decision to Use Tool**: Determine whether to answer based on conversation history or utilize the course material tool (indexed course materials):
   a) **Option A**: If sufficient information exists in the conversation history, respond directly.
      - **Example**: "Laut dem Gespräch ist der Klausurtermin am Dienstag, den 09.07.2024, um 13.00 Uhr." (Based on the conversation, the exam date is Tuesday, July 9, 2024, at 1:00 PM.)
   b) **Option B**: If additional information is needed, use the course material tool.
5. **Query Translation**: Translate the user's query or extract relevant keywords to facilitate retrieval.
   - **Example**: Extracting "Klausurtermin" (exam date) from the query.
6. **Relevant Context Extraction**: Invoke the retrieval process in the previous subsection to obtain relevant context from indexed materials.
   - Involves semantic and keyword retrieval, context reranking, and extracting top chunks.
7. **Chatbot's Answer**: Generate a response using the retrieved context.
   - **Tool-based Response**: "Der Klausurtermin ist am Dienstag, den 09.07.2024, um 13.00 Uhr." (The exam date is Tuesday, July 9, 2024, at 1:00 PM.)
   - **Direct Response**: If not using the tool, respond based on conversation history.

[7] https://cohere.com/rerank

8. **Store Conversation History**: Update the conversation history with the user's query and the chatbot's response to maintain context for future interactions.

By integrating the indexing and retrieval process with the interaction flow, the chatbot effectively serves as an expert on the course material, supporting students in their learning journey. The decision-making process allows the chatbot to handle queries efficiently, providing direct answers when possible and accessing the broader course material knowledge base when necessary.

4. Experimental Results

The evaluation of the BiWi AI Tutor chatbot utilized a dataset comprising questions derived from the course materials, corresponding true answers, and the chatbot's generated responses. These question-answer pairs were developed by the instructors of the educational science course to assess the chatbot's ability to generate accurate and relevant answers. In terms of evaluation methodology, we believe it is important to take into account the learning objectives of the course authors, especially in a domain where expert agreement is not always easy to obtain. The dataset was curated to reflect the diversity of learning materials, including lecture slides, seminar readings, and organizational information, and consisted of 60 questions evenly distributed across the three categories.

4.1. Evaluating Chatbot Responses

To assess the performance of the BiWi AI Tutor chatbot, two evaluation methods were employed: manual evaluation using human annotators and automated evaluation using GPT-4 from OpenAI[8]. This dual approach provided a comprehensive understanding of the chatbot's accuracy and reliability.

Manual Evaluation Using Human Annotators Five human raters, all domain experts and instructors of the course, independently evaluated the chatbot's responses. Each evaluator reviewed the same set of 60 questions and scored the chatbot's answers as either correct (1) or incorrect (0). The majority vote among the five raters was calculated for each response to provide a consensus judgment.

Automated Evaluation Using GPT-4 The second evaluation method utilized the GPT-4 model to assess the correctness of the chatbot's answers. We employed the Question Answer (QA) evaluation prompt from the LangChain library to judge the factual accuracy of the chatbot's responses, disregarding differences in style, wording, and format. Each response was graded as either correct (1) if factually accurate or incorrect (0) otherwise.

Addressing Potential Bias We acknowledge that having the course instructors both develop the evaluation dataset and serve as evaluators may introduce potential bias, as they are familiar with the expected answers and may have subconscious expectations about the chatbot's performance. However, involving external domain experts was not feasible due to resource constraints and the specialized nature of the course content. To mitigate bias:

- **Independent Evaluations**: Each evaluator assessed the responses independently to reduce groupthink and collective bias.
- **Clear Evaluation Criteria**: Evaluators used a binary grading system focused solely on factual accuracy, minimizing subjective interpretations.
- **Inter-Rater Reliability Analysis**: Calculated Fleiss' Kappa to assess agreement levels, highlighting variability and reducing overconfidence in the results.

[8] https://platform.openai.com/docs/models/gpt-4-and-gpt-4-turbo

Table 2
Evaluation results showing the percentage of correct answers as determined by the human majority vote and GPT-4 judgments

Category	Total Questions	Correct Responses (Human Majority Vote)	Correct Responses (GPT-4)
Lecture	20	16 (80%)	17 (85%)
Seminar	20	15 (75%)	15 (75%)
Organizational	20	17 (85%)	17 (85%)

Table 3
Cohen's Kappa values comparing the judgments of GPT-4 with human raters and the majority vote

Evaluator	Lecture	Seminar	Organizational
Evaluator 1	0.58 (moderate)	0.52 (moderate)	0.41 (moderate)
Evaluator 2	0.49 (moderate)	0.62 (substantial)	0.13 (slight)
Evaluator 3	0.21 (fair)	0.36 (fair)	0.82 (strong)
Evaluator 4	0.69 (substantial)	0.73 (substantial)	0.13 (slight)
Evaluator 5	0.82 (strong)	0.46 (moderate)	0.82 (strong)
Majority Vote	0.69 (substantial)	0.85 (strong)	1.0 (perfect)

Evaluation Results The results from both the human majority vote and GPT-4's judgments are summarized in Table 2. The table presents the percentage of correct answers as determined by both evaluations for each category.

Inter-Rater Reliability and Agreement Analysis To evaluate the consistency among human raters, we calculated Fleiss' Kappa scores for each category:

- **Lecture Questions**: Fleiss' Kappa = 0.37 (fair agreement)
- **Seminar Questions**: Fleiss' Kappa = 0.47 (moderate agreement)
- **Organizational Questions**: Fleiss' Kappa = 0.15 (slight agreement)

We also calculated Cohen's Kappa to assess the agreement between GPT-4's judgments and each human evaluator, as well as the majority vote. The results are presented in Table 3.

The varying levels of agreement reflect individual differences among evaluators. The substantial to perfect agreement between GPT-4 and the majority vote indicates that GPT-4's assessments align well with collective human judgment.

4.2. Evaluating the Effect of Using Rerankers

The previous evaluations were conducted without rerankers as the human annotations were gathered before the reranking mechanism was adopted. To investigate the impact of using rerankers, we conducted additional experiments comparing GPT-3.5-turbo's performance with and without rerankers. As seen in Table 4, the reranking mechanism provided a noticeable improvement, particularly in the organizational questions, where accuracy reached 100%. The reranker-based approach leverages semantic re-ranking to filter and improve the retrieval of the most contextually relevant text fragments, leading to better overall answer quality. Although the overall correct response rate was high, certain types of questions, especially open-ended ones, such as those related to seminars and lectures, exhibited greater variability in responses. This reflects the inherent complexity of interpreting such data, leading to lower scores compared to the organizational questions.

Table 4
Comparison of GPT-4 judgments of GPT-3.5-turbo generation with and without reranker

Category	Total Questions	GPT-3.5-turbo (Without Reranker)	GPT-3.5-turbo (With Reranker)
Seminar	20	15 (75%)	16 (80%)
Lecture	20	17 (85%)	16 (80%)
Organizational	20	17 (85%)	20 (100%)

5. Conclusion and Future Work

In prior works, educational chatbots primarily utilized template or rule-based systems to address students' questions. While effective for Frequently Asked Questions (FAQs), these systems lacked flexibility and adaptability, with pre-defined responses leading to static, context-insensitive interactions. With evolving technology, LLMs offer a dynamic alternative, enabling chatbots to generate flexible and deeply contextualized responses. This shift from rule-based to LLM-powered chatbots significantly enhances personalized and nuanced student conversations, accommodating more complex queries. The BiWi AI Tutor chatbot exemplifies an LLM-based system that efficiently retrieves information from sources like lecture slides, seminar materials, and organizational documents. Utilizing a Function Calling Agent from the LangChain library, it accesses specific tools to retrieve relevant material for any query. The system combines generative capabilities with an advanced RAG approach by processing material chunks into embeddings stored in a vector database, facilitating retrieval based on semantic similarity. Additionally, a reranker model filters and prioritizes the most relevant chunks, enhancing information precision and ensuring accurate, targeted responses. The chatbot iteratively refines its answers by dynamically selecting appropriate material chunks, guaranteeing that students receive precise and relevant information. Our experiments provide valuable insights into the chatbot's performance. Comparing GPT-4 with human evaluators, the chatbot consistently delivered correct answers, closely aligning with most human judgments. For organizational questions, there was perfect agreement between the chatbot and majority human evaluations. Introducing rerankers further enhanced accuracy, achieving 100% for organizational content, which underscores the rerankers' effectiveness in filtering retrieved material and improving overall response quality.

Future enhancements can explore multiple dimensions. From a use-case perspective, developing mentoring-style chatbots that provide factual answers while responding to students' emotional needs could involve routing questions to different models based on query nature and support type. However, addressing the psychological aspects and scalability introduces ethical considerations, recognizing that machines may sometimes require human expert involvement. For evaluation, implementing a student feedback mechanism with ratings on a 0-5 scale could assess mentoring effectiveness. Additionally, future iterations could generate personalized learning pathways based on interaction history, allowing the chatbot to adapt to individual learning preferences. From a safety standpoint, the chatbot must include privacy guardrails to filter sensitive or inappropriate data before retrieval. As LLMs become more integral to education, addressing potential biases and ensuring AI decisions are transparent and explainable to both students and educators is crucial. Lastly, scalability and openness are essential for wider adoption, enabling educators to deploy customized chatbot versions by uploading their materials and configuring custom instructions. Future work will also benchmark state-of-the-art open-source LLMs against proprietary models like those from OpenAI to determine the best fit for this use case.

Acknowledgments

The research leading to these results has received funding from the German Federal Ministry of Education and Research (BMBF) through the project "Personalisierte Kompetenzentwicklung und hybrides KI-Mentoring" (tech4compKI) (grant no. 16DHB2206, 16DHB2208)

Declaration on Generative AI

The authors declare that Generative AI tools have not been used during manuscript preparation.

References

[1] R. Kizilcec, To advance ai use in education, focus on understanding educators, International Journal of Artificial Intelligence in Education (2023) 1–8.

[2] F. H. et al., The world of generative ai: Deepfakes and large language models, arXiv preprint (2024). URL: https://ar5iv.labs.arxiv.org/html/2402.04373v1.

[3] M. Sharples, Towards social generative ai for education: theory, practices and ethics, Learning: Research and Practice 9 (2023) 159–167.

[4] A. Neumann, P. de Lange, R. Klamma, Collaborative creation and training of social bots in learning communities, in: 2019 IEEE 5th International Conference on Collaboration and Internet Computing (CIC), 2019, pp. 11–19.

[5] A. Neumann, T. Arndt, L. Köbis, R. Meissner, A. Martin, P. de Lange, N. Pengel, R. Klamma, H. Wollersheim, Chatbots as a tool to scale mentoring processes: Individually supporting self-study in higher education, Frontiers in Artificial Intelligence 4 (2021) 64–71.

[6] M. Maryamah, M. M. Irfani, E. B. T. Raharjo, N. A. Rahmi, M. Ghani, I. K. Raharjana, Chatbots in academia: a retrieval-augmented generation approach for improved efficient information access, in: 16th Int. Conference on Knowledge and Smart Technology (KST), IEEE, 2024, pp. 259–264.

[7] R. Klamma, P. de Lange, A. T. Neumann, B. Hensen, M. Kravcik, X. Wang, J. Kuzilek, Scaling mentoring support with distributed artificial intelligence, in: International Conference on Intelligent Tutoring Systems, Springer, 2020, pp. 38–44.

[8] A. Neumann, Y. Yin, S. Sowe, S. Decker, M. Jarke, An llm-driven chatbot in higher education for databases and information systems, IEEE Transactions on Education (2024).

[9] H. Soliman, M. Kravcik, A. T. Neumann, Y. Yin, N. Pengel, M. Haag, Scalable mentoring support with a large language model chatbot, in: Technology Enhanced Learning for Inclusive and Equitable Quality Education, Springer, 2024, pp. 260–266.

[10] H. Soliman, M. Kravcik, A. T. Neumann, Y. Yin, N. Pengel, M. Haag, H.-W. Wollersheim, Generative ki zur lernenbegleitung in den bildungswissenschaften: Implementierung eines llm-basierten chatbots im lehramtsstudium, in: Proceedings of DELFI 2024, Gesellschaft für Informatik eV, 2024, pp. 171–177.

[11] L. Eby, E. Dolan, Mentoring in postsecondary education and organizational settings (2015).

[12] A. Ziegler, Mentoring: konzeptuelle grundlagen und wirksamkeitsanalyse, in: Mentoring: Theoretische hintergründe, empirische befunde und praktische anwendungen, 2009, pp. 7–29.

[13] M. Dibitonto, K. Leszczynska, F. Tazzi, C. Medaglia, Chatbot in a campus environment: Design of lisa, a virtual assistant to help students in their university life, in: 2018 International Conference on Human-Computer Interaction, volume 10903 of *Lecture Notes in Computer Science*, Springer, 2018, pp. 103–116. doi:10.1007/978-3-319-91250-9_9.

[14] J. Dyrna, J. Riedel, S. Schulze-Achatz, Wann ist lernen mit digitalen medien (wirklich) selbstgesteuert? ansätze zur ermöglichung und förderung von selbststeuerung in technologieunterstützten lernprozessen (2018).