Scalable Mentoring Support with a Large Language Model Chatbot

Hassan Soliman[1][0009−0003−4574−9074], Milos Kravcik[1][0000−0003−1224−1250], Alexander Tobias Neumann[2][0000−0002−9210−5226], Yue Yin[2][0009−0006−8369−8396], Norbert Pengel[3][0000−0002−3263−6877], and Maike Haag[3][0009−0001−6161−2022]

[1] DFKI, Berlin, Germany
{firstname.lastname}@dfki.de
[2] RWTH Aachen University, Aachen, Germany
{lastname}@dbis.rwth-aachen.de
[3] Leipzig University, Leipzig, Germany
{firstname.lastname}@uni-leipzig.de

Abstract. Education students engage in diverse learning activities requiring appropriate assistance and timely feedback. As their numbers grow, providing them with scalable support is an important challenge. Here, we focus on the development of a didactic chatbot based on a Large Language Model (LLM). The potential of LLMs is enhanced by existing materials and pedagogical course descriptions. Using Retrieval Augmented Generation (RAG), the bot can retrieve and analyse course materials, in order to provide comprehensive answers to specific questions. Preliminary results indicate that it is possible to distinguish between different student contexts and to generate a prompt answer, taking into account the relevant materials. The evaluation results achieved 84.78% accuracy in providing correct answers for seminar materials.

Keywords: Large Language Model · Chatbot · Scalable Mentoring.

1 Introduction

Scalable mentoring, including individualised support and timely feedback, is a major challenge in education. The rapid development of technologies opens new perspectives in the context of digital higher education. Our aim is to develop a chatbot helping students in educational science modules of teacher training programmes. Existing learning and information materials as well as process descriptions of the courses can enhance text comprehension and generation potential. This leads us to our research question: How can a chatbot, enhanced with LLMs, be designed and implemented to support scalable mentoring in higher education? The focus on LLMs used in our work is specific and tailored to the field of educational science. The didactic purpose is to provide personalised and contextualised responses to students within a web-based interface called Mentoring Workbench (MWB), facilitating their self-directed learning and mentoring experiences. We wanted to improve an already implemented and tested chatbot [5].

2 Related Work

AI can play a central role in education with its ability to analyse large amounts of data related to the learning process [4]. Generative AI includes technologies for various modalities, while LLMs are a subset of it focused on human-like text [3]. Building social generative AI for education will require developing powerful AI systems that can converse with humans [10].

Researchers have experimented with generative AI models to determine what tasks they can support. Examples include automatically generated online courses for language learning [8], assessing the correctness of students' answers and generating feedback in a digital learning game [7], and evaluating essays to strengthen students' writing skills [9]. Such results will show which tasks can be delegated to AI and which will continue to require genuine human skills.

3 Design and Implementation

Didactically, our scenarios focus on self-directed learning, mentoring, and counselling. Mentoring is considered an effective way to support learning processes [2]. In courses with numerous students, implementing a dyadic relationship between mentor and mentee [12] is not possible due to the limited resources. The question is how mentoring processes can be scaled. It requires an integrated environment with a spectrum of facilities. As a virtual contact available permanently, the chatbot is the central component. It should fulfil two functions: expert and learning companion [1]. As an expert, the chatbot answers questions about the content and organisation of the course. As a learning companion and mentor, the chatbot supports the individual learning process with feedback on submitted writing tasks and encourages students to plan, monitor and reflect on their own learning process. This article deals mainly with the part of answering students' questions about content and organisation (expert).

Our original chatbot [5] is integrated into the MWB [11], which bundles various tools to support students in acquiring skills through self-directed learning, counselling, and mentoring. Until the end of 2022 our development focused on open-source solutions such as RASA[4], a machine learning framework for creating conversational AI. We utilised parts of it, deviating from the traditional RASA Core, to integrate the Social Bot Framework (SBF). This was motivated by its capability for web-based collaborative development, as it provides a visually intuitive platform for defining user intentions and bot responses through drag-and-drop functionalities [6]. We integrated RASA NLU (Natural Language Understanding), allowing to maintain a solid foundation for conversational AI while improving the development experience. Our domain experts in education science crafted a set of examples for each identified intent and entity. For the NLU component, we employed models such as the Dual Intent and Entity Transformer (DIET), a state-of-the-art architecture for simultaneous intent classification and entity recognition. This process ensured the chatbot could parse and

[4] https://rasa.com/

extract relevant information from user interactions. Despite these advancements, our approach shared some limitations and challenges, particularly in employing template-driven Natural Language Generation (NLG): limited flexibility (queries beyond the pre-defined rules), complex conversation management (lack of contextual awareness) and user frustration (less engaging user experience).

3.1 LLM Based Prototype

The LLM-based chatbot prototype "BiWi AI Tutor" facilitates scalable learning support by retrieving knowledge from lecture slides, seminar text, and organisational materials. Harnessing the advanced capabilities of the `gpt-3.5-turbo` model from OpenAI[5] and the LangChain[6] library, the chatbot functions seamlessly within the MWB, offering responsive and contextually aware dialogic interaction (Fig. 1).

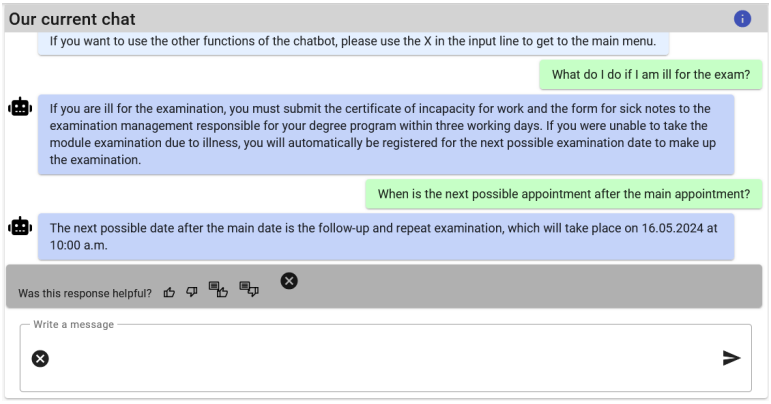

Fig. 1. The BiWi AI Tutor chat interface in the MWB (originally in German).

To this end, the learning materials were categorized and processed into three distinct types: lecture slides, seminar materials and organizational information (in PDF format). These materials underwent parsing, chunking, and indexing using the LlamaParse module from the LlamaIndex[7] library, creating separate indices for each category. This indexing process involved dividing the PDF files into chunks of 1024 tokens with an overlap of 20 tokens, converting these chunks into embeddings via the "text-embedding-3-small" model from OpenAI[8], and storing the embeddings in a vector database for retrieval (Tab. 1).

By combining the generative power of an LLM with retrieval systems, it efficiently locates relevant sections of the learning content from which to generate

[5] https://platform.openai.com/docs/models/gpt-3-5-turbo
[6] https://www.langchain.com/
[7] https://www.llamaindex.ai/
[8] https://platform.openai.com/docs/guides/embeddings/embedding-models

Table 1. Learning material statistics.

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

responses. This is known as Retrieval Augmented Generation (RAG). The chatbot is based on LangChain's ConversationalChatAgent module, which employs the ReAct (Reasoning and Acting) paradigm. This approach leverages LLMs as reasoning engines to determine the actions (tool selection) an agent should take and the appropriate inputs for those actions. After executing these actions, the results are fed back to the LLM to determine the next steps, enabling dynamic, context-sensitive interactions and the handling of multi-question queries.

The chatbot's interaction architecture is meticulously constructed to ensure a user-friendly experience. It follows a logical sequence, starting with the user's message, understanding and processing the query, and selecting the correct tool (material index) based on the nature of the question. The tools employed are textually described and are indices to the most pertinent learning material, enabling the chatbot to retrieve relevant and precise information. After receiving a query, the chatbot generates a semantically rich vector representation of the question, which is used to retrieve contextually relevant text from the content indices through semantic similarity measures. The top 7 semantically similar chunks are retrieved to serve as the context and knowledge for the LLM. Upon retrieval of the pertinent context, the chatbot constructs an informed response to the user's query. This represents the culmination of a comprehensive interaction flow that begins with the user's input and ends with a response answering the query and reflecting a deep understanding of the course material.

3.2 Results

The evaluation of the BiWi AI Tutor chatbot leveraged a dataset comprising questions based on the course, corresponding true answers, and the chatbot's generated responses. This dataset was curated to reflect the diversity of learning materials encompassing lecture slides, seminar texts, and organizational information. The evaluation dataset consisted of 90 questions based on the learning material categories (Tab. 2). These questions were designed by teachers based on BiWi course material to assess the chatbot's ability to generate responses. The dataset was evaluated using the QA evaluation prompt from the LangChain library, where gpt-4 model from OpenAI[9] was employed to judge the correctness of the chatbot's responses. The grading process followed a structured prompt that emphasized factual accuracy, disregarding stylistic and phrasing differences between the chatbot's output and the true answer. The scoring mechanism was guided by the following prompt provided to gpt-4:

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

> "You are a teacher grading a quiz. Grade the student answers based ONLY on their factual accuracy. Ignore differences in punctuation and phrasing between the student's answer and the true answer. It is OK if the student's answer contains more information than the true answer as long as it contains no conflicting statements. Begin!"

The grading of the LLM responses was binary: "Correct" if the chatbot's response was factually correct based on the true answer, and "Incorrect" otherwise. The performance was assessed separately for each material category, with two sets of results: one assuming the correct learning material was provided and another evaluating the chatbot's autonomous tool selection (Tab. 2). The evaluation highlights a stronger performance in answering questions from seminar texts, attributed to the comprehensive nature of textbook content compared to the more concise lecture slides. Additionally, the low number of lecture questions increases the variance of the results compared with the higher number of seminar and organisational questions. The chatbot's ability to autonomously select the correct learning material emerged as a challenge, especially notable in the lecture category. It's challenging even for humans to correctly decide whether the question should be answered from the lecture or seminar material. This difficulty in tool selection is a critical area for future enhancements to improve the chatbot's contextual understanding and decision-making capabilities. This assessment underscores the chatbot's potential as an academic aid while also pinpointing areas for further refinement, particularly in autonomous tool selection.

Table 2. Evaluation results with number of correct responses using two tool selection strategies.

Category	**Total Questions**	**Provided Tool**	**Auto-Selected Tool**
Seminar	46	39 (84.78%)	37 (80.43%)
Lecture	10	8 (80%)	4 (40%)
Organisational	34	26 (76.47%)	21 (61.76%)

4 Conclusion

In our previous trials, the Social Bot Framework and RASA enabled rule-based responses. However, responses to FAQs or course content were fixed. With the move to LLM, it is now possible to benefit from much more flexible and adaptive response generation, enabling deeper contextualisation. This results in a more personalised user experience and allows students to ask more complex questions.

The prototype shows how an LLM-based chatbot can be designed to provide students with answers to questions about the content and organisation of an educational science university course. In order to be more in line with the concepts of mentoring and counselling, it will be necessary to adapt the interaction of the

LLM-based chatbot so that it responds to the current state of the students and also helps them when they do not know exactly what they are looking for.

To further improve the chatbot's performance, we are conducting human evaluations of the chatbot-generated results and comparing them with automated evaluations using gpt-4. We are also experimenting with advanced RAG methods, such as ensemble hybrid search and reranker models. These aim to improve the relevance and accuracy of chatbot responses.

Acknowledgements. 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, 16DHB2213).

References

1. Dyrna, J., Riedel, J., Schulze-Achatz, S.: Wann ist lernen mit digitalen medien (wirklich) selbstgesteuert? ansätze zur ermöglichung und förderung von selbststeuerung in technologieunterstützten lernprozessen (2018)
2. Eby, L.T., Dolan, E.L.: Mentoring in postsecondary education and organizational settings. (2015)
3. Huszar, F., et al.: The world of generative ai: Deepfakes and large language models. arXiv preprint arXiv:2402.04373 (2024), https://ar5iv.labs.arxiv.org/html/2402.04373v1
4. Kizilcec, R.F.: To advance ai use in education, focus on understanding educators. International Journal of Artificial Intelligence in Education pp. 1–8 (2023)
5. Neumann, A.T., Arndt, T., Köbis, L., Meissner, R., Martin, A., de Lange, P., Pengel, N., Klamma, R., Wollersheim, H.W.: Chatbots as a Tool to Scale Mentoring Processes: Individually Supporting Self-Study in Higher Education. Frontiers in Artificial Intelligence **4**, 64–71 (2021)
6. Neumann, A.T., de Lange, P., Klamma, R.: Collaborative Creation and Training of Social Bots in Learning Communities. In: 2019 IEEE 5th International Conference on Collaboration and Internet Computing (CIC). pp. 11–19 (2019)
7. Nguyen, H.A., Stec, H., Hou, X., Di, S., McLaren, B.M.: Evaluating chatgpt's decimal skills and feedback generation in a digital learning game. In: European Conference on Technology Enhanced Learning. pp. 278–293. Springer (2023)
8. Rüdian, S., Pinkwart, N.: Auto-generated language learning online courses using generative ai models like chatgpt (2023)
9. Seßler, K., Xiang, T., Bogenrieder, L., Kasneci, E.: Peer: Empowering writing with large language models. In: European Conference on Technology Enhanced Learning. pp. 755–761. Springer (2023)
10. Sharples, M.: Towards social generative ai for education: theory, practices and ethics. Learning: Research and Practice **9**(2), 159–167 (2023)
11. Zawidzki, J., Bez, M., Jalilov, O.: Needs and requirements of teachers, learners and administrative staff for ai based mentoring tools in higher education. In: INTED2023 Proceedings. pp. 4674–4679. IATED (2023)
12. Ziegler, A.: Mentoring: konzeptuelle grundlagen und wirksamkeitsanalyse. Mentoring: Theoretische hintergründe, empirische befunde und praktische anwendungen pp. 7–29 (2009)