International Journal of Artificial Intelligence in Education
https://doi.org/10.1007/s40593-024-00424-y

ARTICLE

Enlarged Education – Exploring the Use of Generative AI to Support Lecturing in Higher Education

Darius Hennekeuser[1] · **Daryoush Daniel Vaziri**[1] · **David Golchinfar**[1] · **Dirk Schreiber**[1] · **Gunnar Stevens**[2]

Accepted: 28 July 2024

Abstract

Large Language Models (LLMs) are rapidly gaining attention across the open-source and commercial fields, bolstered by their constantly growing capabilities. While such models have a vast array of applications, their integration into higher education—as supportive tools for lecturers—has been largely unexplored. Exploring this area entails understanding the specific requirements and viewpoints of higher education lecturers. We developed an LLM-based assistant with retrieval augmented generation (RAG) capabilities and lecturing materials as its data foundation. For the design of the system, we followed a user-centered design approach. Subsequently, we conducted user studies and qualitative interviews with university lecturers. Our findings suggest that lecturers are ready to use LLMs with RAG in higher education under the condition that such systems are reliable, explainable, controllable, and trustworthy. We discuss design implications that designers of LLM-based systems should consider when developing such tools for higher education. This paper adds to the scarce existing studies on the usage of LLMs in educational contexts.

Keywords Natural Language Processing · Semantic Search · Large Language Models · Human-centered Design · AI · Education

✉ Darius Hennekeuser
darius.hennekeuser@h-brs.de

[1] University of Applied Sciences Bonn-Rhein-Sieg, Sankt Augustin, Germany

[2] University of Siegen, Siegen, Germany

Published online: 06 August 2024

Introduction

The rapid advancements in the field of artificial intelligence have led to the development of sophisticated large-scale language models (LLMs) capable of comprehending and generating human-like language. Among these models, the Generative Pre-trained Transformer (GPT) series, developed by OpenAI, has gained considerable attention for its creative abilities and unparalleled performance. The GPT models have exhibited exceptional capacities to generate coherent and contextually relevant text, performing at or near human levels across various language benchmarks (Brown et al., 2020; Radford et al., 2018, 2019). Similarly, the English model series LLAMA 2, developed by Meta and released under an open-source license, garnered significant attention (Touvron et al., 2023). In most academic tasks, successive improvements in the GPT series have led to the development of models with increased numbers of parameters, leading to the latest GPT-4 model's superior performance (OpenAI, 2023a).

As LLMs gain prominence, numerous opportunities and challenges arise for various sectors, particularly in education. Studies by Kasneci et al. and Milano et al. have highlighted potential educational use cases and challenges for LLMs. These models can analyze student performance, provide tailored feedback, support lesson planning, generate questions and prompts, facilitate language learning, and assist in research and writing tasks. Furthermore, LLMs can enhance teachers' professional development by offering resources and insights into new methodologies, tools, and even semi-automate grading and plagiarism checking (Kasneci et al., 2023). However, challenges such as distinguishing between model-generated and student-generated content (Milano et al., 2023), data privacy concerns, overreliance on models, and the difficulty of customizing the models must be addressed (Kasneci et al., 2023). While the research conducted by Kasneci et al. and Milano et al. provides a preliminary understanding regarding the potential challenges and prospects of LLMs in educational environments, they did not incorporate a collaborative user study involving educational peers such as lecturers or students. Recently, Smolansky et al. (2023) published a concise report detailing the results of a survey where both students and lecturers expressed their concerns regarding the integration of LLMs into educational settings. The authors emphasize that to successfully implement educational reforms; it is crucial to actively involve both these key stakeholders (2023). The scarce empirical findings about the integration of LLMs in educational contexts and the need to involve key stakeholders in the educational integration of LLM-based systems highlight an ongoing research gap. Our research aims to bridge this gap by probing into possible usage scenarios and formulating design recommendations for systems based on LLMs, particularly from the perspective of university lecturers.

To ensure a customized interaction between the lecturers and the LLM-based system, we integrate lecturing material into the interaction mechanism by using the technique of retrieval-augmented generation (RAG), which enables pretrained LLMs to access context-sensitive knowledge (Lewis et al., 2021). Therefore, we develop semantic embeddings for instructional materials, which serve as the foundation of an embedding-based retrieval system for ranking document texts according to their similarity to a lecturer's search query (Huang et al., 2020). This enables the integration

of highly similar learning materials from relevant lecturer's resources into the LLM's processing of user queries while considering the specific teaching content. Then, we conduct controlled user studies coupled with qualitative interviews to assess the application of LLM-based assistants in higher education teaching.

This paper explores the potential contributions of an assistant based on LLMs and RAG to support higher education instructions by addressing two primary research questions:

RQ1: How can LLMs with RAG support practices of university lecturers?
RQ2: What are implications for the design of applications using LLMs with RAG in higher education?

This study contributes to the limited existing research on the application of advanced language models in educational settings. More importantly, it pioneers the integration of individual educational materials into the interactive process, an aspect not yet explored in previous empirical studies at the time of this writing.

Related Work

Natural Language Processing in Educational Contexts

Numerous studies have explored the application of Natural Language Processing (NLP) techniques in the educational field, tackling various aspects of teaching and learning. This section presents a review of several notable examples relevant to the application of NLP in education. Different applications focus on using NLP to address the needs of both teachers and students (Litman, 2016). NLP methods can be used to support the personalization of curriculum materials by automatically identifying and evaluating materials from digital sources specifically suited to a student's reading proficiency and preferred subjects (Miltsakaki & Troutt, 2008; Petersen & Ostendorf, 2009; Pitler & Nenkova, 2008). E.g., Petersen and Ostendorf (2009) introduced a machine learning approach as a method of measuring the reading level of students in order to find topical texts at an appropriate reading level for foreign and second language learners. Moreover, NLP has been studied as a method for enabling the reuse of existing materials across different student proficiency levels via text simplification. Candido Jr et al. (2009) introduced a system called Simplifica, which adapts texts to cater to readers with differing literacy levels by employing a range of simplification operations.

NLP-based methods for automatically generating multiple-choice, wordbanks, and other types of test questions by processing texts in the subject domain were also explored (Litman, 2016). Here, Heilman and Smith (2010) introduced a ranking system for automatically generated questions from reading materials by a rule-based natural language generation system. The ranking system on top of a natural language generating system almost doubled the acceptance rate of automatically generated questions from 27 to 52% when only selecting the top-ranked 20% questions (2010).

Another area that has been gaining momentum is the implementation of NLP algorithms in educational web-based eBooks. A study conducted in Slovenia aimed to investigate the integration of NLP algorithms into a low-labor model for educational web-based eBooks and assess their effectiveness in grading open-ended questions. The researchers worked with computer science students to develop language-agnostic NLP algorithms capable of grading open-ended questions. In the study, NLP algorithms were integrated into the IMapBook software, which was subsequently deployed in a series of fourth-grade language arts classes in Slovenia. The best algorithm demonstrated an accuracy rate of approximately 85% in grading real-world answers as correct, partly correct, or wrong (Smith et al., 2020).

Similarly, NLP has been applied to the task of automatically grading and providing feedback on learners' written responses. Here, Allen et al. (2015) demonstrate how NLP techniques can be employed to develop stealth assessments of students' reading comprehension abilities.

As a next step in technological progression in NLP, LLMs have become a focus of research in educational applications, bringing promising advancements in various areas of education. LLMs stand apart from the methods mentioned above, as they use a different neural network architecture, namely transformers, and thus may be considered as generative AI (Lv, 2023). This enables them to create new content by predicting subsequent words based on input sequences' attention scores instead of their conventional predictive mechanics. One example can be found in a paper by Sarsa et al. (2022) which investigates the use of OpenAI Codex, an LLM, for creating programming exercises (including sample solutions and test cases) and code explanations for programming courses. The study shows that most of the automatically generated content is novel, sensible, and, in some cases, ready for immediate use. LLMs, such as OpenAI Codex, can serve as a valuable tool for instructors, although some supervision is needed to ensure the quality of the generated content (2022).

Elkins et al. (2023) studied the quality and usefulness of generated questions for use in the classroom with a sample of teachers. They applied controllable text generation, which combines a LLM with a prefix such as a keyword that guides the model to the desired output (e.g., "Generate easy questions"). The authors conclude that LLMs can aid in the generation of high-quality, diverse educational questions, reducing the burden on teachers and improving the quality of educational content. Teachers have found the generated questions to be of high quality and sufficiently useful for classroom settings. Moreover, LLMs have demonstrated efficiency in generating different types of questions across various domains, showing their promise for widespread use in education (2023).

Additionally, the potential of LLMs in medical education is explored by Kung et al. (2023). The authors evaluated ChatGPT on the United States Medical Licensing Exam and displayed promising results. The study found that ChatGPT performed at or near the passing threshold of 60% accuracy without specialized input from human trainers, demonstrating its ability to potentially assist human learners in a medical education setting (2023).

Altogether, the early findings on applications of LLMs in educational settings suggest the potential for further integration of LLMs in education.

Human-Centered Design Approaches in AI Applications for Education

Implementing human-centered design (HCD) approaches and human-centered AI (HCAI) in the field of education and AI applications not only results in improved user experience but also addresses various ethical and moral concerns, such as the difficulty of preserving non-automated aspects in education and the lack of transparency about the ownership of AI-generated content (Holmes et al., 2022). In recent years, multiple research papers have advocated for the adoption of HCD and HCAI methodologies to create more responsible and people-friendly educational technologies. Human-centered AI emphasizes the need to involve stakeholders and end-users in AI development (Auernhammer, 2020).

One proposed approach is the value-sensitive design, which incorporates the moral values of stakeholders during the design, research, and development processes. It aims to provide different perspectives on societal, diversity, interaction, and human needs in the design of computer systems, such as AI. Further, its application involves identifying social values, deciding on a moral deliberation approach, and linking values to formal system requirements and concrete functionalities. Overall, working toward human-centered approaches while developing educational technologies will benefit both providers and consumers of AI solutions for educational purposes. HCAI can help achieve better user experiences and create a balance between enhancing machine capabilities and human capabilities (Renz & Krishnaraja, 2020). In the context of AI in Education, human-centered design is an essential approach to acknowledge the unique needs of learners, teachers, and other stakeholders. Designing educational systems using HCD methodologies has been shown to improve the connection between home and school learning, parents and teachers, and contextual evaluation of system prototypes (Luckin et al., 2006).

Additionally, Renz and Vladova (2021) highlight the relevance of applying HCAI approaches in the field of education. By focusing on human-centered dimensions such as trust and transparency, the development of AI in Education could potentially lead to increased acceptance and usefulness of AI technologies in education. The authors stress that raising awareness and educating stakeholders about the benefits of HCAI would help prevent unfounded skepticism and facilitate its adoption (2021). Overall, the importance of human-centered design approaches in educational and AI applications is well founded, as they prioritize user needs, provide better user experiences, and incorporate ethical and moral considerations. HCD and HCAI thus play a crucial role in the development and acceptance of AI-driven educational technologies.

Methodology

Study Setting

In order to address RQ1, we employed the GPT-4 model, as it has demonstrated superior performance across various benchmarks at the time of our study. In addition, the multilingual capabilities and allowance for a maximum of 8,192 tokens (to the time the study was conducted) made our system capable of processing German language

and larger texts from the lecture material (OpenAI, 2023a). Offering double the input token limit compared to other multilingual models like GPT-3.5, which only permits a maximum of 4,096 tokens (as of the time our study was conducted), ensured more effective and comprehensive content processing (OpenAI, 2023b). Previous research by Kojima et al. (2023) suggests that zero-shot learning is an effective approach for LLMs, and Espejel et al.'s (2023) preliminary findings indicate that LLMs' reasoning capabilities might improve using zero-shot learning as the model size increases. Consequently, we applied zero-shot learning with an instruction to guide the LLM towards finding solutions to prompts from lecturers in educational settings. The exact prompts and system architecture can be found in Fig. 1. The human-centered design process of defining the prompt for our user study will be further described below.

The following paragraph describes the system's architecture. A detailed description of the process, from incorporating lecturing materials up to the user interaction, is described at the end of this section. We used the open-source vector database Milvus for a semantic text search of lecturer materials to avoid exceeding the maximum input and output token limit of GPT-4. A vector database is a specialized data

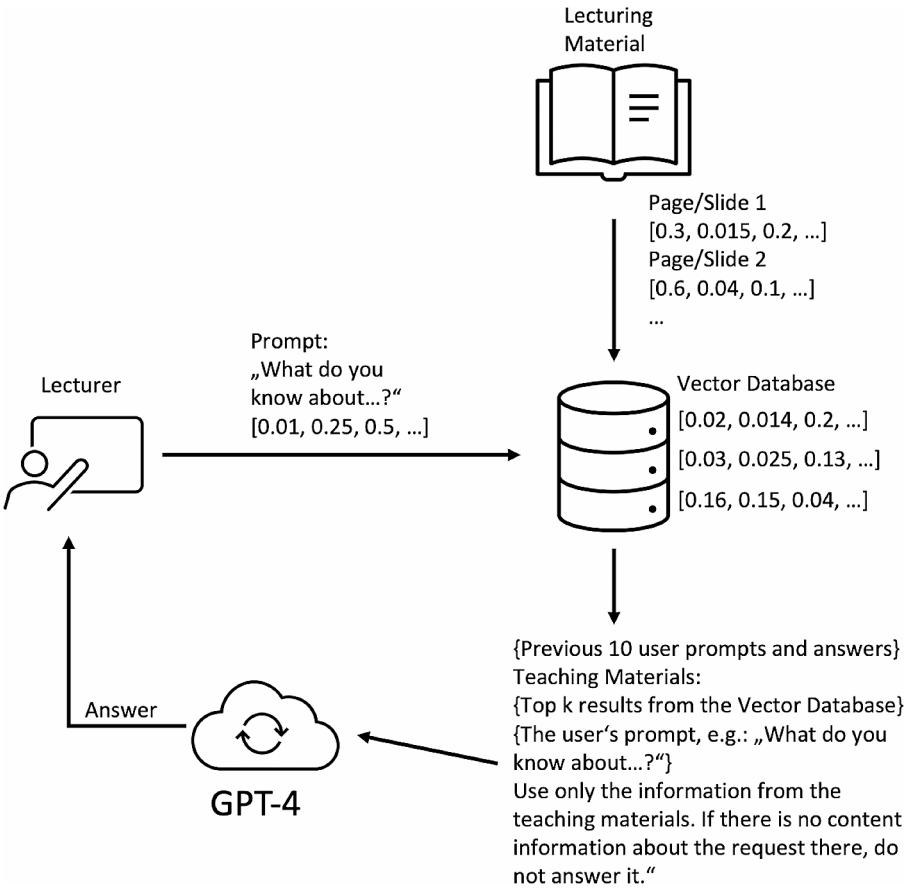

Fig. 1 System Architecture of the LLM using RAG

management system designed to efficiently store and search high-dimensional vector data, which is particularly common in data science and AI applications. These databases are optimized to handle large-scale and dynamic vector data that can be generated from machine learning models when transforming unstructured data into feature vectors for various types of data analytics. Milvus is a vector database that has been developed as a purpose-built data management system to manage large-scale vector data efficiently. It supports various similarity functions and is therefore well suited for our user study (Wang et al., 2021). The vector database is used to store written words from lecturer materials in text embeddings (basically numerical values), representing segmented educational content. By converting user queries into text embeddings, Milvus facilitates efficient identification of the k-nearest of comparable educational materials corresponding to the user's query. Subsequently, this set of k educational content items can be integrated with the user's prompt, providing contextual information derived from the lecturer's materials to the LLM. This approach is known as Retrieval Augmented Generation (RAG) and enables the output generation based on external knowledge databases (Lewis et al., 2021).

We engaged in an iterative, human-centered design workshop with two university lecturers to define the additional text extensions for the user's prompt. These text extensions are added to the user's actual prompt to force the model to generate responses based only on the lecture content, and thus, no hallucinations occur. The lecturers provided a PDF script of one of their courses, and we stored each page of the script in the vector database. Detailed documentation of the workshop and its results can be found in the supplementary material. After several iterations of redefining the prompt extension based on the lecturers' feedback, both lecturers confirmed that the teaching content provided by the system was based on their own lecture materials. The construction of the prompt, which incorporates lecture materials, the user's input, and additional text developed in the prompt design workshop, can be located at the bottom right of Fig. 1 and at the end of Chap. 4.1.

For the embedding model, we used OpenAI's (2023c) "text-embedding-ada-002" model, which is known to outperform other OpenAI models in text search tasks. These are designed to identify the most relevant documents by comparing the embedding vectors of the query and each document (OpenAI, 2023d). The chat interface used in the experiments bears similarities to the ChatGPT interface in order to ease the usage for participants with prior experiences with ChatGPT (Mastery, 2023).

To provide a more detailed understanding of our system's functionality used in this research, we will outline each stage in the process, from acquiring the lecture material to the system's response to the user's prompts.

Prior to user engagement, the following steps take place:

1. The lecturer provides the teaching materials (book or script) in a.pdf format.
2. The text from each pdf page is individually extracted.
3. The extracted text is converted into text embedding vectors using the text-embedding-ada-002-model.
4. The vector database is set up and populated with each text and its associated text embedding vector.

During user interaction, the following stages occur:

1. The user triggers the system with a prompt.
2. This prompt is transformed into a text embedding vector using the text-embedding-ada-002-model.
3. The system calculates the distance between the text embedding vector of the given prompt and all the text embedding vectors for the lecturing material. The system identifies and retrieves the top-k most similar vectors and their corresponding texts.
4. Optional: If the conversation is ongoing, the prior ten sets of user prompts and system responses can be used as a contextual reference for the conversation.
5. The GPT-4 model further receives the conversational context, if available, the top-k most analogous texts from the lecturing material, and a user prompt together with a pre-set prompt addition via the API:

"{Previous ten user prompts and answers}.
Teaching Materials:
{Top k results from the vector database}
{user prompt}

Use only the information from the teaching materials. If there is no content information about the request there, do not answer it."

Data Collection

Following the workshops on prompt design and the decisions regarding the embedding model, we conducted several experiments involving lecturers and subsequent expert interviews. Before initiating the experiment, each participant supplied us with lecture materials, such as scripts (6) and their own lecture books (3), which were integrated into the vector database for semantic text search. The scripts, which were provided by six lecturers, were intended for the presentation of content in the lectures. In contrast to the textbooks, which were provided by three lecturers, the lecture notes consisted mainly of bullet points and less of full text. Each of the teaching materials was considered the main course material for the students of the courses. All lecture materials were in German language, except for one in English. The experiment commenced with a brief questionnaire to gather demographic information and data on participants' usage of LLMs in general and, more specifically, within learning environments. Detailed demographic information about the participants can be found in Sect. 4.3. Furthermore, participants were required to provide consent for audio recording as a prerequisite for their involvement in the experiment.

The experiment began with a concise description and an overview provided by the research team. The overview included a brief explanation of how the system works. All participants knew that the teaching material they provided was converted into vectorial representations and entered into a database. They also knew that their prompt was then enriched with similar and therefore potentially relevant teaching material. We did not show the exact structure of the final prompt to the LLM, which

can be seen in the bottom right corner of Fig. 1, as it has been irrelevant for the participants to assess the output of the frontend system. Participants were instructed to interact with the chatbot, which was connected to the vector database and contained individual lecture materials. They were requested to pose open-ended questions to the chatbot in areas where they believed it could provide assistance. Participants were asked to interact with the system for at least 15 min while using the think-aloud method. While participants engaged with the chatbot, the research team asked questions relevant to the research inquiries. These questions were exploratory and aimed at gaining additional insights into how participants perceived the chatbot's responses. Primarily, such questions were posed when participants unintentionally omitted to use the think-aloud method and failed to give feedback on the responses. In some cases, feedback from participants about the (desired) functionality of the chatbot led to open questions to gain deeper knowledge about the implications for the design of RAG-based chatbots in education (e.g. in the case of P1, we asked: "What do you think about the response of the system?").

Upon completion of the experiment, semi-structured interviews were conducted with the participants, along with open-ended inquiries, to gain deeper insights into their responses. The experiments and interviews took 30–60 min for each participant. We collected the data from the 25th of April, 2023, to the 10th of May, 2023. One participant (P5, 27 years, male; see Sect. 3.3) attended the experiment online. Everyone else participated in person.

Demographic Data

In this section, we describe descriptive data about the participants gathered from the questionnaire prior to the experiment.

The age of our participants ranged from 27 to 61 years, with a mean age of 43.8 years. Gender representation among participants was skewed, with seven males and two females (see Table 1).

Table 2 delineates the distinct subjects and furnished reference materials utilized in the experiment, primarily from the perspective of Business Faculties. Three participants relied on books to deliver their lectures, while scripts were the preferred choice for six participants. Additionally, we divided our participants into senior lecturers (at least ten years of lecturing experience) and junior lecturers (less than ten years of

Table 1 Demographic data

Participants	Age	Gender
P1	58	Female
P2	29	Male
P3	58	Male
P4	37	Male
P5	27	Male
P6	34	Male
P7	61	Male
P8	44	Male
P9	46	Female

Table 2 Lecturing characteristics

Participants	Lecturing Material	Subject	Lecturing Experience
P1	Script	Business Statistics	Senior Lecturer
P2	Script	Business Informatics	Junior Lecturer
P3	Book	Internet Economics	Senior Lecturer
P4	Book	IT-Innovation-Management	Senior Lecturer
P5	Script	Introduction to Scientific Work	Junior Lecturer
P6	Script	Data Management	Junior Lecturer
P7	Book	Process Management	Senior Lecturer
P8	Script	Innovation-Management	Senior Lecturer
P9	Script	Private and Business Law	Senior Lecturer

Table 3 Experience with LLMs

Participants	Use Frequency of LLMs	Use of LLMs for Lecturing Purposes
P1	Never	Not at all
P2	Very often	To a small extent
P3	Sometimes	To a small extent
P4	Very often	To a small extent
P5	Very often	Not at all
P6	Very often	Not at all
P7	Sometimes	To a small extend
P8	Sometimes	To a small extend
P9	Never	Not at all

lecturing experience). Three of our participants can be considered junior lecturers, while the others are senior lecturers.

Lastly, our participants stated whether they had prior experience with LLMs and whether they had used LLMs for lecturing purposes prior to the experiment (see Table 3). P1 (58 years, female) and P9 (46 years, female) both had no prior experience with LLMs. All other participants had used LLMs prior to the experiment and, except for P5 (27 years, male) and P6 (34 years, male), even used them for lecturing purposes to a small extent.

Data Analysis

The Whisper model (Radford et al., 2022) was employed to transcribe each interview automatically. Afterwards, all transcripts were proofread manually and augmented with details regarding the speakers and their respective contributions to the discus-

sion. Furthermore, observation notes made during the experiments were structured for the analysis. The collaborative qualitative data analysis methodology proposed by Richards and Hemphill (2017) was implemented to examine the transcripts, as it is specifically tailored for qualitative data analysis conducted by teams, ensuring both validity and reliability. Our analysis procedure is shown in Fig. 2. Three of the authors conducted the analysis as a team.

In the initial phase, we delineated the scope of the study and identified suitable journals and conferences aligned with the research theme. Concurrently, research questions and data collection methodologies were examined, culminating in a comprehensive plan for data analysis. In the subsequent phase, a team of three researchers applied open and axial coding (Strauss & Corbin, 1998) techniques to independently identify patterns and potential codes within case studies (open coding) and to establish correlations among the observed patterns and codes across multiple cases. As a result, we established a preliminary code book, which is accessible in the supplementary material of the collaborative qualitative data analysis documentation.

In adherence to the prescribed methodology, we executed a pilot test utilizing the codebook on two uncoded datasets, wherein all team members coherently coded the transcripts, ultimately addressing discrepancies and coding conventions during team meetings. Upon making minor modifications to the codebook, the final coding phase commenced. In accordance with Richards and Hemphill's (2017) recom-

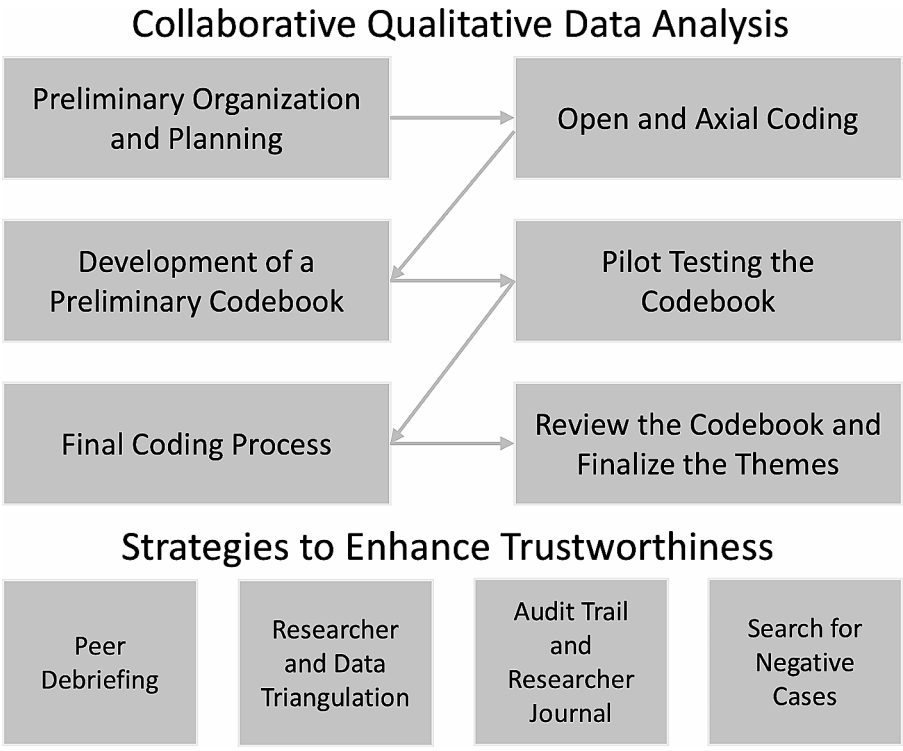

Fig. 2 Collaborative Qualitative Data Analysis (own representation)

mendations, split coding was employed for data analysis, considering the smaller size of our research team, while consensus coding was suggested for larger teams to address concerns regarding coding consistency. Lastly, the data analysis team convened to review the final codebook, which involved identifying negative cases where the coding system was inapplicable. An impartial peer, uninvolved in the research but knowledgeable about the subject matter, reviewed and endorsed the codebook and its process.

The methodology adopted in this study encompassed multiple researchers and various data sources, including a descriptive data questionnaire, experiment transcripts, interview transcripts, and observational notes acquired during the experimental phase (facilitating Researcher and Data Triangulation). The documentation of the data analysis can be found in the supplementary materials, thereby maintaining an audit trail and researcher journal strategy.

Results

Our collaborative, qualitative data analysis has led to the development of seven distinct categories: User Experience, Potential Use Cases, Challenges of Implementing within Educational Environments, Suggestions for Enhancement, Trust, and Perceived Usefulness. Each category will be elaborated upon in its own subsequent section.

User Experience

In the context of this section, we refer to user experience as the participants' total interaction with the LLM. We assess its usability based on factors such as output quality, effectiveness of information retrieval, ease of communication, processing speed, knowledge base, and user adaptation needs.

Within User Experience, we found factors influencing the usability of our system. First of all, the participants were majorly satisfied with the system. Many comments focused on the impressive and correct output quality of the LLM. P1 (58 years, female) was captivated by it, noting, "Amazing. I find this already fascinating". This sentiment was reinforced by their admiration of the system's ability to cite various lecture materials, including PowerPoint and PDF files while maintaining accuracy and consistency. One user stated, "Great! It's completely correct. It has created a summary from multiple PDF slides" (P2, 29 years, male). Users also reflected on the sophistication of the LLM's language. They found it linguistically of high quality compared to students' responses in examinations, calling it "high-quality in comparison to answers that I usually get in exams for these questions" (P3, 58 years, male). The vast majority of participants have appreciated that they can interact with the system like with a knowledgeable tutor: "It's a system that one can really simply speak to, just like a tutor who knows the entire script by heart" (P2, 29 years, male) which indicates a high degree of user satisfaction and engagement. One of the users has mentioned that the system's capability to pull together information from different sources, even if they were not directly connected in the original material, was

particularly impressive: “So what amazes me is, that’s probably through this vector storage, these texts are not all directly behind each other. (…) So he actually put that into context” (P7, 61 years, male). The consensus was that the usage of the system was satisfying. As one participant put it, “It is definitely a great system, and I can definitely envision something like this being used more frequently in the future” (P6, 34 years, male).

While the overall user satisfaction with the system was high, in some cases, dissatisfaction was caused by malfunctioning or missing functionalities. One notable area of difficulty was in recognizing the context of previous inquiries and responses. Users expressed frustration over the system’s failure to effectively comprehend and recall prior exchanges, thus undermining the continuity and coherence of the interaction and possibly leading to potential misunderstandings (“Okay, now the system cannot work context-based. So it tells me now that it’s unclear to which previous question the answer refers. This means that I would now be frustrated as a user because it did not maintain the context here” - P2, 29 years, male). Also, the limitation of the system’s knowledge to provide lecture material was partly a matter of concern. Users expressed a preference for a more comprehensive knowledge base, one offering a wider range of examples and extending beyond the limited scope of set exercises. This is linked to the expectation of increased knowledge gain: “Yes, well, I actually think the larger the knowledge base is, the better. So, of course, there is a risk if something else suddenly appears, but in principle, this increases the gain in knowledge or the possible gain in knowledge. That’s why I actually think it’s better” (P3, 58 years, male).

Limiting the system’s capability to existing lecture content was seen as a constraint on its potential to enhance the learning experience through exposure to diverse and added examples by some participants. Another significant issue raised was the lengthy processing time. Users were dissatisfied with the slow pace of response, causing discomfort and delay. There seemed to be a concern as to whether the model’s current speed of deriving responses was compromising the time-efficiency of using it as a learning tool: “But of course, if I do this task by task, then depending on how long the waiting time is, the question might arise again, can’t I work faster by myself by perhaps copying things from exams that I had already set?” (P4, 37 years, male). The average response time of the system was 21 s. This long response time has made it challenging for users, particularly for students using it within the context of an online oral examination. As P3 (58 years, male) comments, “The time required for an answer is so long that a student who has the tool has to bridge a lot during an oral online exam before he gets an answer.” Users have the feeling that the system hangs or takes unusually long: “Ah, here comes the answer. The feeling prevails that something has hung up or that something is taking unusually long” (P2, 29 years, male). According to P8 (44 years, male), this issue, if not addressed, can greatly diminish the ease of use for contemporary users, who are accustomed to quicker response times. They expressed their desire for improvements if the system is to be effectively implemented.

In addition to the concerns about response times, users expressed unease and confusion regarding the interaction pattern with the system, especially in regard to the use of natural conversation. Participants demonstrated uncertainty on how to input

or phrase their queries to receive useful feedback from the system. As P1 (58 years, female) asked, “What do I have to tell it?” and further queried, “So it doesn’t matter how you enter it there?”. This uncertainty could understandably add to the feeling of awkwardness and difficulty in using the system. The constraints of the system became evident in instances where users inputted less structured sentences or used abbreviations not found in the system’s pre-loaded material, as seen in the case of participant P3 (58 years, male). As such, to some participants, it seems clear that the system may require a certain degree of learning and adaptation from the user side for effective use, as suggested by P8’s (44 years, male) remark: “Now you probably have to learn what things are possible.” On the other hand, despite initial confusion about the natural language input modality, in the post-interview P1 (58 years, female) found: “I can actually get what I want relatively easily through simple wording. So, it’s very user-friendly”.

Lastly, expectations about the system usage were predominantly fulfilled and exceeded. Users expressed surprise and fascination at the abilities of the system, especially when it managed to return detailed and correct answers to their queries. E.g., P4 (37 years, male) noted: “Great, I also find that very cool because here, I kept the query very general, where one could have expected anything possible from the model, and then the model has specialized itself on data types that are exclusively for time indications.” P9 (46 years, female), on the other hand, was both surprised and worried about the capabilities of the system, indicating that the expectations were exceeded: “Well, as I mentioned, I am, yes, positively surprised, on the one hand positively surprised, on the other hand somewhat shocked, that this program can do so much.” Users mentioned that the system generated well-formulated sentences and was quite specific in its responses, and one user was particularly impressed with the accuracy when given a vague question (e.g., “Yes, but otherwise, I must say, I am very, very positively surprised by the wording” – P4 (37 years, male) and “So, I’m surprised. I would have expected fragments instead. Not fully formulated sentences that are well-phrased and essentially ready for publication” – P7, 61 years, male). However, P7 (61 years, male) was expecting the system to find typos inside the lecturing material: “I would have thought that it might find typos now, as the algorithm cannot identify content-related issues.” Additionally, P8 (44 years, male) wished for more creative responses to generate new examples fitting the existing lecturing materials.

In conclusion, while the system is highly praised for its impressive output quality, comprehensive knowledge retrieval, and high-quality language use, concerns around the response time, the need for user adaptation, and the desire for a wider range of knowledge base highlight areas for future improvements. This feedback provides crucial insights into the user experience of such systems. A comprehensive list of identified codes pertaining to user experience is available in Table 4, which includes the total count of such codes across all transcripts as well as corresponding sample quotes.

Table 4 Factors of user experience

Factors of User Experience	Code Quantity	Example Quotes
Usability – High ease of use	3	*"I can get what I want from it through relatively simple wording. So, it's very easy to use"* (P1, 58 years, female)
Usability – Low ease of use	19	*"People are so spoiled these days. And that makes it very tedious. So, that's what I would expect if this were a product that was actually in use, [the responses] would somehow have to get faster"* (P8, 44 years, male)
Usability – High user satisfaction	64	*"I was very satisfied to have this. I call [the content of my requests] Mode 1, where we only talk about the reproduction of content"* (P2, 29 years, male)
Usability – Low user satisfaction	10	*"But of course, it's frustrating how long [the response generation] is taking now"* (P8, 44 years, male)
Valued functionalities	15	*"Ah, this is pretty cool, there's a note here where you can essentially look up the Create command again in the script"* (P6, 34 years, male)
Critics about functionalities	12	*"Oh yes, there are no answers to errors in the book. Okay, he can't do that… He can't do that. Well, so he just draws on factual knowledge and combines it"* (*P7*, 61 years, male)
Expectations fulfilled or exceeded	15	*"But otherwise, he did very well, I have to say. So I am, as I mentioned, positively surprised on the one hand, and on the other hand a bit shocked that this program is capable of so much. But looking at it from the perspective of wanting to help students, it obviously makes complete sense"* (P9 46 years, female)
Expectations not fulfilled	9	*"Ah, okay, but now he didn't stop at the first definition, he only took the definition of theory of science, which I didn't actually ask for directly"* (P5, 27 years, male)

Usefulness

Overall, educators positively responded to the potential implementation of AI-based systems in lecturing environments. This section explores the factors contributing to the perceived usefulness of these systems in educational contexts and the reasons behind them. P2 (29 years, male) stated, "Yes, I would now very much like to

make this system available to my students and ask them how they like it," suggesting its general usefulness inside the lecturing context. Similarly, P5 (27 years, male) explained: "Well, I do believe that, for the purpose I had in mind, it fulfills its function quite well. So, it can be used quite effectively for those things". They perceived it to be well suited for lectures and for engaging interaction. One reason for the perceived usefulness of the system can be found in its time-saving functionalities. Several participants expressed that the system could help to reduce the time and effort needed to prepare teaching materials and create tasks based on their own lecturing materials. For instance, P4 (37 years, male) found it helpful to quickly create questions for an exam: "That makes things much easier for me. Of course, I could come up with ten tasks on my own, but that takes time. This significantly reduces that time effort. I don't know by how much, but I think it's high enough to be noticeably beneficial". P5 (27 years, male), P7 (61 years, male), and P8 (44 years, male) indicated that the usage might be time-saving, too, when engaging new topics and including them in lecturing materials quickly. P7 (61 years, male): "So, if a lecturer doesn't feel like dealing with the details and takes a standard textbook, then tells the algorithm to compile teaching materials on a topic, then the algorithm would do it. He would only need to look very roughly into the book. This would be more than sufficient for a lecture". Also, the output quality of the provided information led to increased perceived usefulness, as, e.g., P4 (37 years, male) stated, "These are all phrases that one could simply use in the exam." Participants underlined the high quality of the answers provided and were impressed by the ability of the model to create grammatically sound and contextually appropriate responses. Some found that their answers were presented in a stronger, more comprehensive manner by the system, indicating useful applications in writing tasks: "Well, I wrote it with the same meaning, but not in those exact words. Not exactly, right. And he rephrased it better. So, in principle, my facts are in there, but he made better sentences out of it" (P7, 61 years, male).

From a pedagogical standpoint, participants believed this tool could provide valuable assistance. For instance, the structured way of presenting content with the system would improve knowledge transfer. P1 (58 years, female) rated it to be useful for students because it does not just provide solutions but also helpful explanations: "Because the solution to the tasks is already structured. It doesn't just provide a solution, but firstly you need this, secondly you need that, thirdly you need this". Furthermore, the interactive nature of learning with such a system was perceived as a more effective style of learning compared to traditional sequential reading by P3 (58 years, male): "I'm no expert in didactic neurology, but I suspect that engaging in this way gives a much stronger sense of creativity. Not the boring repetition and reading of books".

P5 (27 years, male) perceived the system to be less useful in his lecturing context when it provided sources inaccurately: "Well, the teaching material that I use is intended for the course in which students, for example, should be taught scientific work. The point is to sensitize them precisely to such things. They should use sources and make sure that the sources match the places where they were used. This is a fundamental part of scientific work that you really cite sources where they were used and not somewhere else". The next Section provides more insights into trust and output explainability and its relationship to transparent sources of information.

In conclusion, the participants found the system to be a valuable tool for enhancing their lecturing activities, particularly due to its ability to save time, create high-quality content, and promote engaging, interactive learning. However, some caution was expressed regarding the precise citation and use of sources, highlighting the need for careful adaptation and monitoring in an academic setting. A comprehensive list of identified codes pertaining to usefulness is available in Table 5. The table includes the total count of such codes across all transcripts as well as corresponding example quotes.

Output Explainability and Trust

This section delves into the output explainability and trust factors in interactions with the system, how users reacted to the provided sources of information, and the influ-

Table 5 Factors of usefulness

Factors of Usefulness	Code Quantity	Example Quotes
Lower usefulness due to missing references	1	"*Well, the teaching material that I use is intended for the course in which students (...) should be taught scientific work. The point is to sensitize them precisely to such things. They should use references and make sure that the references match the places where they were used. This is a fundamental part of scientific work, that you really use references at the right places*" (P5, 27 years, male)
High usefulness due to high quality of answers	4	*"The quality of the posed question is very good. As I just said, the system cannot know where I have placed the emphasis, in the lecture, because it is not present during the lecture"* (P4, 37 years, male)
High usefulness due to didactical reasons	4	*"But I do believe that such an interaction is more interesting for [the students] than this...well, just pure script"* (P9 46 years, female)
High usefulness due to time-saving capabilities	12	*"That makes things much easier for me. Of course, I could come up with ten tasks on my own, but that takes time. This significantly reduces that time effort. I don't know by how much, but I think it's high enough to be noticeably beneficial"* (P4, 37 years, male)
General statements about the usefulness of the system	8	*"Here, there were no [hallucinations], I would say. That it somehow created completely different things than the content of the lecturing material. Which of course can happen with such a completely open model. So I think it's very suitable for teaching"* (P6, 34 years, male)

ence these aspects had on the perceived trustworthiness of the system. Moreover, it explores participants' concerns about the ambiguity of the information's source and their varying levels of trust towards the system.

In certain scenarios (randomly and not as an intended functionality), users were provided with references to their lecture materials when interacting with the system. These included references to specific slide numbers or pages where related content was provided. Users were appreciative of this feature, as it allowed them to cross-verify information. This was denoted by our 'Output Explainability' code. P5 (27 years, male), P6 (34 years, male), and P7 (61 years, male) were among those who spoke positively about the system's provision of source references. P6 (34 years, male), in particular, noted that it was helpful to know where they could find a specific SQL command within the lecture scripts.

The importance of output explainability was further reinforced as participants tried to find the origin of certain pieces of information. This was evident when P4 (37 years, male) questioned the source of an example that was not included in the original lecture materials and when P5 (27 years, male) expressed caution about an answer provided by the system that was outside of their teaching materials. In both instances, it was apparent that the participants were keen to know where the system was pulling its information from. This sentiment was echoed by P7 (61 years, male) and P8 (44 years, male). P2 (29 years, male) even suggested an approach whereby users could query about the origins of the information directly.

The system's trustworthiness was also influenced by the comprehensibility of the information sources it employed. With the exception of P1 (58 years, female), P3 (58 years, male), and P7 (61 years, male), all participants voiced some reservation regarding the sources of information used by the system and indicated reliance on educational materials outside those provided in the lecture at some point during the experiment or interview. P5 (27 years, male) and P8 (44 years, male) demonstrated particularly lower levels of trust in the system. P5 (27 years, male) registered dissatisfaction when the system referred to information not directly connected to the lecture material: "If I knew right now, the system not only accesses my script but also partially supplements it with other sources, it would not be a problem for me. But if I was told the chatbot is only fed with knowledge from the script text, and then other things come up, I would certainly think, okay, this is indeed a bit strange". Additionally, he raised concerns about data privacy: "And where do my contents go then. This is not necessarily always a one-way street". Likewise, P8 (44 years, male) expressed difficulties in ascertaining trust in the system: "Yes, it's always a bit hard to say. I mean, the annoying thing about these models is that they always seem totally plausible, and often they're correct, but they're also often wrong but still look plausible. So, in this context, it's always a bit hard to say".

Conversely, P1 (58 years, female), P2 (29 years, male), P3 (58 years, male), P6 (34 years, male), and P7 (61 years, male) demonstrated a higher level of trust in the system, despite their cautiousness during the experiment. P1's trust is attributed to recognizing many familiar examples within the system's responses and lecture material. P2 (29 years, male) expressed confidence in the system's readiness for direct student use. P3 (58 years, male) supported the system's reference to a wider range of educational materials for enhancing its knowledge base despite potential risks: "The

larger the knowledge base is, the better. Of course, there is a risk if something else suddenly appears, but basically, the potential for gaining knowledge increases with it. So, I would actually prefer that". P7 (61 years, male) rationalized that even if the system utilized information from unspecified sources, it was reminiscent of numerous lecturing books that lacked appropriate citing.

In conclusion, while the system's ability to reference material and provide source information was seen as a positive attribute, concerns about data privacy and the accuracy of the information provided influence user trust. A comprehensive list of identified factors of trust and output explainability is available in Table 6, which includes the total count of such codes across all transcripts as well as corresponding example quotes.

Potential Use Cases

A plethora of potential use cases was recognized with the implementation of such systems. Broadly speaking, these use cases can be categorized under three key areas: preparation of lecture materials, the system as an in-lecture assistant, and the system as a study aid for students. This section diverges from the section about usefulness by elaborating on potential usage scenarios of AI-based systems in educational contexts,

Table 6 Factors of Trust and output explainability

Factors of Trust and Output Explainability	Code Quantity	Example Quotes
Appreciation about references	3	*"So, [the generated exercise] would really be a generic exam question, 100 points. He even gave the page number, that's unbelievable"* (*P7*, 61 years, male)
Assumption about external sources of information	6	*"I'm just wondering, where from? So, I asked for revenue, but actually, there's no task involving revenue in my slides. You can just infer this by simply specifying a sales price per unit. But I've never had that in the script"* (P4, 37 years, male)
Cautiousness	25	*"I believe that [the system] is good for the first step to maybe get an impression for a topic. But I would not trust it 100% at this point so that I would convey things one-to-one to the students without checking them first"* (P5, 27 years, male)
High level of trust	12	*"So, overall, I did have the feeling that I could trust the system quite well, to be honest. Based on the responses I received to my inquiries, it was apparent that they stemmed from my script"* (P4, 37 years, male)

as identified by the participants' post-interaction. In contrast to usefulness, this section provides a general overview of potential use cases for AI-based systems like the one used in the study. It is important to mention, though, that this section does not reflect any experiences and user assessments post-interaction. Instead, it elaborates on users' ideas and expectations for additional use cases for AI-based systems in higher education.

When it comes to preparing lecture materials, the most frequent use case cited by participants, minus P5 (27 years, male), was the generation of exercises along with their solutions. P1 (58 years, female) saw a potential application in exam exercises' creation, while P4 (37 years, male) suggested that while not every question could be adopted verbatim from the system, the drafted exercises could be manually elaborated upon: "(…) if [the system] came up with these questions based solely on the script, I think that's pretty good because then one could just extend the questions based on that and would have an exam for 90 points."

Participants 3 (58 years, male), 5 (27 years, male), 6 (34 years, male), and 8 (44 years, male) found the system's ability to create content summaries particularly useful, especially when tackling new lecture topics. P3 (58 years, male) emphasized the advantage of not being constrained to specific scripts and the value of exploring new topics from various sources and perspectives. Participants 1 (58 years, female), 7 (61 years, male), 8 (44 years, male), and 9 (46 years, female) found the system useful for enhancing existing lecture materials either by introducing more recent examples (P7, 61 years, male; P8, 44 years, male) or by identifying ambiguities when a specific prompt does not yield a sufficient answer (P1, 58 years, female; P9, 46 years, female): "It could show me again what things might be important to summarize and explain in more detail. Especially for people who are not present in the lecture. I could well imagine that.".

Furthermore, the participants envisaged the system as a useful tool in actual lectures. P2 (29 years, male), P3 (58 years, male), and P8 (44 years, male) saw the potential for the system to act as a discussion partner for students, while P4 (37 years, male) proposed the idea of the system continuously incorporating live lecture transcriptions and therefore to be able to respond to queries about new topics in the lectures live.

Lastly, every participant viewed the system as a potentially prized study assistant for students. Predominantly, participants saw it as an examination preparation partner. However, P6 (34 years, male) suggested that it could also be used by students seeking to recap the content of previous lectures, particularly those who were not present: "Possibly summaries once again, for students who want to get an overview of what actually happened, who might not have been able to attend the lecture, to receive some kind of summary of the last lecture. I might find that quite interesting". A comprehensive list of identified potential use cases is available in Table 7. Table 7 includes the total count of such codes across all transcripts, as well as corresponding example quotes.

Table 7 Potential use cases

Potential Use Cases	Code Quantity	Example Quotes
Preparation of lecturing material – Creation of exercises	17	In response to an inquiry about possible applications for this system.: *"In the creation of exam tasks, in the creation of exercise tasks, then also in the evaluation of the results, in the creation of sample solutions"* (P3 58 years, male)
Preparation of lecturing material – Content summaries	5	*"Certainly, you could provide a text and request a summary of the key points. This works quite well"* (P8, 44 years, male)
Preparation of lecturing material – Development of new subject areas	7	*"If I were to do something like Basics of Business Informatics, then, of course, I could give it an introductory reading that I think is cool and tell it: build a lecture script out of this for me"* (P8, 44 years, male)
Preparation of lecturing material – Improvement of existing lecturing material	7	*"It could show me again what things might be important to summarize and explain in more detail. Especially for people who are not present in the lecture. I could well imagine that"* (P1, 58 years, female)
Lecture Assistant	6	*"For example (...) I could use this system as a lecture assistant in the form of a robot. And then maybe incorporate it into the lecture as a sparring partner"* (P2, 29 years, male)
Study Assistant	20	*So, the students do like to work interactively. So, these questions could now be transferred into an online quiz"* (*P7*, 61 years, male)

Recommendations for Improvements

The participants shared valuable insights for the improvement of such systems, predominantly revolving around the expansion of the knowledge base and enhancing the graphical capabilities of the system. Participants 3 (58 years, male), 4 (37 years, male), 6 (34 years, male), and 9 (46 years, female) recommended the incorporation of a feature allowing them to navigate whether or not the system should have access to external knowledge bases, such as the internet. This is reflective of lecturers' desire for control over system outputs: "I would appreciate it if I had control over it as a lecturer" (P9, 46 years, female). Additionally, participants 2 (29 years, male), 4 (37 years, male), and 5 (27 years, male) suggested more innovative ways to distinguish between content derived from lecture material and external sources. P2 (29 years, male), for example, recommended the use of a prompt to receive data like slide numbers where specific information was located, while participants 4 (37 years, male)

and 5 (27 years, male) suggested some kind of internal marking to help differentiate contents.

In reference to graphical feature functionalities, there were suggestions for the analysis of visual content, such as images and videos embedded within the lecture material (P2, 29 years, male; P5, 27 years, male). These enhancements also included generating illustrations to enhance the comprehension of the system's output (P1, 58 years, female; P3, 58 years, male; P8, 44 years, male). For example, P2 (29 years, male) proposed both image and video analysis and transcription of video audio to foster content interactivity.

Participants 2 (29 years, male), 3 (58 years, male), and 5 (27 years, male) proposed the introduction of voice input and output functionalities to improve the ease of interacting with such systems. P3 (58 years, male) and five noted that these features would increase system usability, while P2 (29 years, male) desired more human-like interactions with the system. In a similar vein, P4 (37 years, male) envisioned a possible addition of a visual avatar to give the system more personality and realism in a virtual world.

Table 8 Recommendations for improvement

Recommendations for Improvement	Code Quantity	Example Quotes
Extended knowledge base	9	*"Yes, so I would actually find [access to external knowledge bases] better. The larger the knowledge base is, the better. Of course, there is a risk if something from outside the lecturing material is generated, but basically, the increase in knowledge gain or potential knowledge gain can be achieved"* (P3 58 years, male)
References	3	*"Yes, so for me, the ideal situation would be if [information from outside the lecturing material] gets marked clearly. Then I wouldn't have any problem at all. So, I ask a question, and then [the system] first gives me the information from the script. And then I would need a clear identification from the chatbot, which [external content] may be generated behind that"* (P5, 27 years, male)
Graphical features	7	*"As I said, the graphical processing, I would find that great"* (P1, 58 years, female)
Voice interaction	4	*"Yes, then, what would obviously be even simpler, if it worked well, would be to possibly do such things simply by voice command"* (P5, 27 years, male)
Avatar	2	*"So, for example, to make the whole thing even more realistic in the virtual world, I could imagine providing my avatar for the students. I could give it certain personality characteristics of mine"* (P4, 37 years, male)
Scope of answer	2	*"For example, I could imagine a slider at the top in the middle. Microsoft has implemented something similar with Bing. Exact answers, meaning in the sense of just reproducing, or extended mode: additional content"* (P2, 29 years, male)
Archive	2	*"Definitely I could imagine a kind of archive where one could see which tasks might have already been generated in the past"* (P6, 34 years, male)
Export functions	1	*"For the output, I would like to have an export function. That means, I don't really want to have to mark it and have to copy it into my own Word file"* (P4, 37 years, male)

P2 (29 years, male) suggested the provision of a slider functionality to control the extent of detail in output, serving to further emphasize user control over the system. Lastly, an archiving feature was proposed by P6 (34 years, male) and P9 (46 years, female) to keep track of queries and responses from prior sessions, and the addition of an export button to facilitate the generation of documents in formats such as.pdf was conceptualized by P4 (37 years, male). A comprehensive list of recommendations for improvement is available in Table 8, which includes the total count of such codes across all transcripts, as well as corresponding example quotes.

Challenges within Educational Environments

Several challenges around the usage of such systems within the realm of education surfaced from the comments of the study participants. The main concern was with regard to the incorporation of external material in the responses given by the system. P2 (29 years, male) highlighted the potential risk that students might inadvertently learn erroneous content delivered by the system rather than absorbing the correct content derived from lecture materials. This sentiment was echoed by P5 (27 years, male) and P9 (46 years, female), who further pointed out that students, unlike lecturers, may not have the discerning capacity to differentiate between accurate and incorrect insights. Although P7 (61 years, male) did acknowledge this risk, he believed that it was not a substantial concern since many lecture materials are not always precise when referencing external sources. P9 (46 years, female) offered a different perspective, positing that the rise of AI-generated content also poses a risk of students submitting this AI-produced content as their own coursework.

Regarding the direct use of such systems, P6 (34 years, male) voiced concerns that the process of integrating new material into the system may prove too complex for most lecturers; hence, the importance of including user-friendly features that allow for the seamless addition of new topics into the system. P1 (58 years, female) added that despite the inherent value of lecture material such as scripts and textbooks, it does not encompass the entirety of information pertinent to a subject, as it cannot replace the comprehensive knowledge acquired from classroom attendance. This sentiment was summarized by P1(58 years, female): “Of course, one thing is what’s written in the script; the other thing is what’s added on the audio track.” A comprehensive list of all identified challenges for implementation in education environments can be found in Table 9, which includes the total count of such codes across all transcripts, as well as corresponding example quotes.

Discussion

To address **RQ1** and **RQ2**, our analysis identified several crucial factors. This section will delve into and discuss the key findings for using LLM-based assistants with RAG to interact with lecturing materials in higher educational contexts. Mainly, the discussion will revolve around the potential of LLM-based systems for university lecturers and the specific implications when designing them to enhance educational processes.

Table 9 Challenges within educational environments

Challenges within educational environments	Code Quantity	Example Quotes
Use of external information	5	*"And I see a certain risk, of course. As [the system] has even led me onto thin ice with his first response. These are the kinds of things that, in my opinion, should be clearly explained to the students. One can offer [such a system] as a supportive tool, but one must always check and read it carefully. If in doubt, compare it again with the script"* (P9 46 years, female)
Recognition of LLM generated content	2	*"But if you have to write, for example, homework or something like that, of course, I see the danger that everyone sees, namely, that you might no longer do it yourselves"* (P9 46 years, female)
Effort to integrate lecturing material	1	*"Also, it depends on, how easy or complicated it is for the lecturer to feed his own documents into this model. That might be another point that could potentially discourage me. If I had to go through a complicated procedure every time, I might not use it"* (P6, 34 years, male)
Missing instructions of lecturers	1	*"Of course, one thing is what's written in the script, the other thing is what's added on the audio track"* (P1, 58 years, female)

Usability and User Experience

Our research results primarily delved into factors of usability and user experience. Most notably, many users expressed their satisfaction with the system, praising its accuracy, the quality of its outputs, and its advanced capability to retrieve information. The system's effectiveness as a capable tutor particularly caught their attention. When comparing the perceived output accuracy, quality, and information retrieval abilities, a noticeable gap in research publications exists for the straight application of LLM-based systems in educational settings. Research work by Ruan et al. (2019) outlined the creation of an educational quiz chatbot, which, although not based on an LLM, but on sentence similarity to generate responses, showcased high accuracy. Still, at times, the system experienced difficulties in understanding complete phrases (2019). Analyzing our results and participants' overall satisfaction with the prompt handling, we believe LLMs' NLP capabilities can boost the overall effectiveness of such systems. This enhancement would occur through improved comprehension abilities, setting them apart from non-LLM-based systems.

The system, while largely successful, was not without its shortcomings. User dissatisfaction primarily stemmed from system malfunctions, some missing features, constraints of the system's knowledge bank, and, most notably, response time. The similarity search in our vector database took less than a second; however, the overall process of user inquiry took longer on average due to the time it took the GPT-4 API to process the results of the similarity search. This was identified as the most frustrating factor for users and, consequently, a hindrance to the system's effectiveness. At the time of our experiment, GPT-4 was not publicly available hence we relied on a license for research purposes. It should be noted that the long processing time could be due to the prototypical stage of the API. Another contributing factor to the long processing time was the vast number of tokens to be processed from the similarity search results. One can infer that the response time of LLMs could vary based on the hosting hardware, the number of tokens processed in each request, and the number of simultaneous requests. Our findings share similarities with Gnewuch et al's (2022) research that stresses the importance of a well-balanced chatbot response time. Their study revealed user dissatisfaction in instances of both delayed and rapid response rates. Experienced users found the artificially delayed response time vexing as they were aware of the chatbot's ability to respond instantly. While their study scrutinized the effects on social presence and the likelihood of using the chatbot (2022), our findings suggest that prolonged response time also dampens the user experience.

In relation to the aspect of malfunction, we found minimal instances of participant dissatisfaction regarding their prompt responses. For instance, we found that users can experience frustration due to a misspelled term, which can result in an inaccurate response from such systems. This issue parallels the findings of Ruan et al., whose system also struggled with providing proper responses to prompts that included typographical errors. In our view, the problem was not necessarily linked to the limitations of the LLM but more closely tied to the system's similarity search functionality. We observed that, on other occasions, the system was able to correctly respond to prompts despite them containing typographical errors. In one case, the misspelled term was directly linked to a key subject term from a lecture book. We speculate that this may have caused the similarity search algorithm to falter in locating corresponding pages within the database. At this juncture, we would advocate for additional examination and analysis of such complications when using approaches akin to ours.

Benefits and Drawbacks of LLM-based Systems in Higher Education

Overall, the system met or even exceeded users' expectations. Its value was largely attributed to its capacity for saving time, thus reducing the effort expended in developing educational materials and crafting exercises from lecture materials. Primary features identified as useful were predominantly in relation to interactions with existing lecture materials. Additionally, the capacity to improve or supplement personal teaching materials by utilizing external resources, such as scientific books or the internet, was highly valued.

The beneficial applications of such systems can be grouped into three primary categories: the creation of new lecture materials, support within lectures, and serving as support for students. A predominant use case for the participants was the cre-

ation of exercises initiated both by students and lecturers. This validates the research conducted by Sarsa et al. (2022), who evaluated the use of LLMs for generating programming exercises and code explanations. They concluded that Codex, a model trained to support coding tasks, has significant potential to assist both computer science teachers and students (2022). Consequently, we advocate for further exploration by both practitioners and researchers to determine the most efficient use of LLMs for exercise creation. Participants also suggested that this exercise generation could serve as a tool for exam creation. However, there were a few criticisms about AI-generated exam submissions. This parallels the concerns probed by Dobslaw and Bergh (2023). They studied the use of ChatGPT for answering exam questions and found it was capable of successfully completing exam exercises. For this reason, they recommend altering the format of exams to preclude LLM usage (2023).

Nonetheless, instructors see the implementation of LLM-based systems as a beneficial interactive learning tool for their students. They value the autonomy it gives students to interact with lecture scripts and reading materials. This potential utility for students is reinforced by Leinonen et al.'s (2023) findings, which revealed that students rate the code explanations provided by GPT-3 higher than those given by their peers, suggesting the utility of such systems for students.

To augment the system further, participants articulated the need for larger and more varied knowledge bases. Participants wished for functionalities like the ability to include materials from the internet as well as to combine multiple books in one database. Further research should investigate such functionalities to include multiple sources for the semantic text search. On the other hand, participants were partially worried about the loss of control over the information sources provided by the system. The design implications of these findings will be discussed in the next section.

In conclusion, from the viewpoint of educators, the primary benefits of integrating systems like ours into their teaching process can be found in the creative development of educational content. This can range from designing exercises to other teaching materials, which may be derived from their own resources or from extraneous educational sources. As for students, these systems provide an interactive interface with the learning materials, making it possible for them to receive explanations directly. This is viewed as highly advantageous in enhancing their learning experience.

Despite being viewed as advantageous, the implementation in educational environments comes with certain obstacles. Some lecturers had reservations regarding the precision of citations and sourcing offered by the system. While the system's ability to cross-reference information from lecture materials was highly valued for its transparency and clear output, doubts emerged about the origins of specific data not initially sourced from the lecture content. Future research should explore possibilities that such systems can introduce features that clearly denote information sources.

An additional obstacle is the incorporation of external resources. There is a risk that the LLM-based system might inadvertently disseminate faulty content, misleading students with inaccurate details. We have explored potential solutions to this problem in the *Design Implications* section below. The focus of our study lies on how LLM-based systems with RAG can support the practices of lecturers. However, most participants considered the potential application of such systems for direct interaction with students. To facilitate a secure interaction between students and such systems,

further research and investigation are undoubtedly needed. Additionally, participants mentioned the predicament of students potentially presenting AI-generated assignments as their own, similar to the results from Sarsa et al. (2022) who found that LLMs might enable students to create solutions in exams. The subsequent section will outline the design implications derived from the findings of the study.

Implications for the Design of LLM-based Systems in Higher Education Environments

From our study, we can derive implications for the design of LLM-based assistants in higher education. We will structure and describe these implications into separate paragraphs.

Use Visual Elements to Enhance Interaction with LLM-based Systems

Incorporating visual elements, using avatars, introducing speech recognition and audio output, expanding the knowledge base, and providing clear marking of information sources are some of the revisions that could significantly enhance the effectiveness of such systems. The idea of using virtual avatars as teaching assistants further supports the research conducted by Vallis et al. (2023), which states that avatars are well-received by students in an educational framework. Therefore, using avatars for the visualization of teaching assistants might be relevant for designers of LLM-based systems. Although most participants expressed the need for visual elements such as pictures and graphs, there were others who suggested the necessity for text-based analysis of images when assimilating lecture content into the knowledge database. As an implication for the design of LLM-based chat systems, models such as the high-profile, text-to-image model, stable-diffusion-v1-4 could be introduced for creating illustrations and charts within the chat (Rombach et al., 2022). To convert lecture images into text, systems could incorporate image-to-text models like vit-gpt2-image-captioning before generating embeddings of the lecture content for the vector database (NLP Connect, 2023). Horawalavithana et al. (2023) proposed a framework consisting of both a visual and a textual encoder in order to enable LLMs to process multimodal instructions. Similar approaches might also be applicable to LLM-based systems in higher educational contexts.

Provide Speech-to-text and Text-to-speech Features

Users in our study mentioned that they would appreciate the possibility of generating their prompts via voice inputs. It would ease the process and improve their experience with LLM-based systems. Thus, designers should consider the integration of Speech-to-Text technologies, for instance, open-source technologies like whisper, to allow users a more convenient interaction with the LLM (Radford et al., 2022). In the same context, participants could also imagine the system to generate voice outputs of the LLM's results, for instance with open source applications like SpeechT5 (Ao et al., 2022). In general, text-to-speech and speech-to-text features would improve the

users' experience with the LLM-based systems. Additionally, such features would further enhance the general accessibility of such systems.

Let Users Switch between Open- and Closed-world Settings

A frequently mentioned request among participants referred to the possibility of alternating between an 'open-world' mode, granting access to external resources found online, and a 'closed-world' mode, restricted to referencing information solely from the lecture materials contained within the system database. Such features could increase users' trust in the system and thus promote acceptance. Designers of LLM-based systems could pursue approaches similar to Hussain and Athula (2018), in which they extended the knowledge base with MediaWiki, a Wikipedia API, which equips the chatbot with extra knowledge and, at the same time, limits it to the determined sources.

Allow Individual Handling of Sources

Participants expressed their desire to manage multiple textbooks or manuscripts in the system. Thus, LLM-based assistants for higher education should enable users to upload multiple types and sources of materials into the system. In the same context, many participants asked for options to display the sources of the system's response or at least provide a short notice in case information were not derived solely from the uploaded materials. Such features could potentially amplify the system's transparency and users' trust and might be beneficial for the design of LLM-based systems to be used in higher education settings.

Enhance Processing of Queries that Refer to Previous Interactions

Improvements are necessary in the way such systems process queries related to previous interactions. Our study revealed that users are satisfied when such a system effectively understands queries linked to previous ones but are frustrated when it fails to do so. In the system presented in this study, every query is sent to the vector database, which returns the top k similar pages/slides. The prompt then refers to these results from the similarity search (Top k similar pages/slides + Prompt + "Use only the information from the teaching materials. If there is no content information about the request there, do not answer it."), which made the LLM occasionally struggle to comprehend references to prior requests.

Handling of Misspelled Prompts

LLM-based systems using RAG should incorporate features that effectively manage queries with misspelled terms. Such features may prevent the system from accurately identifying learning materials to answer a given query. For instance, in the case of the participant whose prompt could not be answered by the system due to a misspelled term in the prompt, future systems could offer prompt suggestions if the similarity

between the query input and the existing text embeddings in the vector database falls below a certain threshold. This might guide users to formulate a more precise query.

Optimize Response time

Finally, measures to reduce response time can improve user experience. Response time could be decreased by reducing the amount of processed data, for instance, by providing less context from the semantic text search. Another solution to reduce response time could involve the self-hosting of open-source models, which would allow the manipulation of the response time by adjusting the system hardware and thereby provide the users with options to select the most suitable hardware for their use cases. As described in the Usability and User Experience section above, we believe that the long processing time in our study derived from the early stage of the GPT-4 API. Therefore, the processing time might also decrease in the near future.

Limitations

It is imperative to note that the results of this study come with certain limitations, mainly linked to the demographic setup of the participants and the variances in their experiences. Most participants were male (7 out of 9) and aged from 27 to 61 years, which could have biased the interpretation of the findings. Further, the heterogeneous level of familiarity with LLMs among participants could have affected their interaction with the system. Furthermore, study participants were recruited only from the university business department. Even though the topics of the teaching materials varied, the results may not necessarily be applicable to faculties outside of business disciplines. This highlights the need for extended research in different education disciplines. Lecture materials also varied in type. While some participants appreciated the use of books, others opted for script-based materials. This variability in material preference suggests that system requirements and user experiences could differ based on the type of lecture material. Another limitation showed up in some experiments, where the format of the lecturing material led to system errors due to differences in the length of the context provided. For example, searching the top results from a database containing book pages produced longer prompts sent to the LLM than scripts. This length exceeded GPT-4 model's token limit, which may have negatively impacted the experience of respective participants. We employed split coding in our qualitative analysis. The reliability of this method might be debatable, but it is recognized as an effective technique for small research teams conducting qualitative data analysis. Although the attribution of inter-reliability scores could have potentially strengthened the reliability of our findings, it is important to note that there is no definitive consensus on whether coupling consensus coding and inter-reliability scores results in more reliable outcomes. Therefore, we chose to utilize split coding in this instance. Lastly, at the time of our experiment, GPT-4 was not publicly available. Hence, we relied on a license for research purposes. This becomes a limitation to our study as it potentially affects the speed and efficiency of our system and, therefore, also the user's experience. It is essential to keep these limitations in mind when interpreting our study findings. Lastly, we chose not to incorporate an analysis

of the conversation logs within our study. Instead, we employed the 'think-aloud' method, which encourages participants to vocalize both their queries and responses during system interaction to gather comprehensive insights. However, we acknowledge that an examination of conversation logs could offer additional valuable data on the utilization of such systems in academic environments and should be considered for further studies in this area.

Conclusion

Our research reveals that systems based on LLMs using RAG approaches could be of significant utility in higher education contexts to provide an interactive conversational interface for users interacting with lecture materials. Educators recognize the benefits that such systems offer in improving lecture preparation, assisting in delivering lectures effectively and serving as a study tool for students. Nevertheless, the use of such technologies is not without potential risks and downsides, which must be considered while designing LLM-based systems for higher education. The key concerns revolve around the possibility of misinformation and insufficient referencing. As LLM systems become increasingly relevant in educational settings, our findings aim to shed light on initial design considerations for designing and integrating such systems in higher education contexts. However, more research is needed to uncover a deeper understanding of the students' needs for such systems, their perception, how to maximize its efficacy to enhance learning environments, and the impacts on education quality when using LLM-based systems. In this sphere, our study provides a robust foundation upon which both practitioners and researchers can build further.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s40593-024-00424-y.

Author Contributions All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by Darius Hennekeuser, Daryoush Daniel Vaziri and David Golchinfar. The first draft of the manuscript was written by Darius Hennekeuser and all authors commented on previous versions of the manuscript. All authors approved the final manuscript.

Funding The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Declarations

Competing Interests The authors have not relevant financial or non-financial interests to disclose.

References

Ao, J., Wang, R., Zhou, L., Wang, C., Ren, S., Wu, Y., Liu, S., Ko, T., Li, Q., Zhang, Y., Wei, Z., Qian, Y., Li, J., & Wei, F. (2022). *SpeechT5: Unified-Modal Encoder-Decoder Pre-Training for Spoken Language Processing* (arXiv:2110.07205). arXiv. https://doi.org/10.48550/arXiv.2110.07205

Auernhammer, J. (2020). *Human-centered AI: The role of Human-centered Design Research in the development of AI.*

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., & Amodei, D. (2020). *Language Models are Few-Shot Learners* (arXiv:2005.14165). arXiv. https://doi.org/10.48550/arXiv.2005.14165

Candido Jr, A., Maziero, E. G., Specia, L., Gasperin, C., Pardo, T., & Aluisio, S. (2009). Supporting the adaptation of texts for poor literacy readers: A text simplification editor for brazilian portuguese. *Proceedings of the Fourth Workshop on Innovative Use of NLP for Building Educational Applications*, 34–42.

Dobslaw, F., & Bergh (2023). Experiences with remote examination formats in light of GPT-4. *Proceedings of the 5th European Conference on Software Engineering Education*, 220–225. https://doi.org/10.1145/3593663.3593695

Elkins, S., Kochmar, E., Cheung, J. C. K., & Serban, I. (2023). *How Useful are Educational Questions Generated by Large Language Models?* (arXiv:2304.06638). arXiv. https://doi.org/10.48550/arXiv.2304.06638

Espejel, J. L., Ettifouri, E. H., Alassan, M. S. Y., Chouham, E. M., & Dahhane, W. (2023). *GPT-3.5 vs GPT-4: Evaluating ChatGPT's Reasoning Performance in Zero-shot Learning* (arXiv:2305.12477). arXiv. https://doi.org/10.48550/arXiv.2305.12477

Gnewuch, U., Morana, S., Adam, M. T. P., & Maedche, A. (2022). Opposing effects of Response Time in Human–Chatbot Interaction. *Business & Information Systems Engineering*, *64*(6), 773–791. https://doi.org/10.1007/s12599-022-00755-x

Heilman, M., & Smith, N. A. (2010). Good Question! Statistical Ranking for Question Generation. *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 609–617.

Holmes, W., Porayska-Pomsta, K., Holstein, K., Sutherland, E., Baker, T., Shum, S. B., Santos, O. C., Rodrigo, M. T., Cukurova, M., Bittencourt, I. I., & Koedinger, K. R. (2022). Ethics of AI in education: Towards a community-wide Framework. *International Journal of Artificial Intelligence in Education*, *32*(3), 504–526. https://doi.org/10.1007/s40593-021-00239-1

Horawalavithana, S., Munikoti, S., Stewart, I., & Kvinge, H. (2023). SCITUNE: Aligning large Language models with scientific multimodal instructions. *arXiv*. https://doi.org/10.48550/arXiv.2307.01139. arXiv:2307.01139.

Huang, J. T., Sharma, A., Sun, S., Xia, L., Zhang, D., Pronin, P., Padmanabhan, J., Ottaviano, G., & Yang, L. (2020). Embedding-based Retrieval in Facebook Search. *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, *2553–2561*. https://doi.org/10.1145/3394486.3403305

Hussain, S., & Athula, G. (2018). Extending a Conventional Chatbot Knowledge Base to External Knowledge Source and Introducing User Based Sessions for Diabetes Education. *2018 32nd International Conference on Advanced Information Networking and Applications Workshops (WAINA)*, 698–703. https://doi.org/10.1109/WAINA.2018.00170

K Allen, L., L Snow, E., & S McNamara, D. (2015). Are you reading my mind? Modeling students' reading comprehension skills with Natural Language Processing techniques. *Proceedings of the Fifth International Conference on Learning Analytics and Knowledge*, *246–254*. https://doi.org/10.1145/2723576.2723617

Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., & Hüllermeier, E. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, *103*, 102274. https://doi.org/10.1016/j.lindif.2023.102274

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2023). *Large Language Models are Zero-Shot Reasoners* (arXiv:2205.11916). arXiv. https://doi.org/10.48550/arXiv.2205.11916

Kung, T. H., Cheatham, M., Medenilla, A., Sillos, C., De Leon, L., Elepaño, C., Madriaga, M., Aggabao, R., Diaz-Candido, G., Maningo, J., & Tseng, V. (2023). Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLOS Digital Health*, *2*(2), 1–12. https://doi.org/10.1371/journal.pdig.0000198

Leinonen, J., Denny, P., MacNeil, S., Sarsa, S., Bernstein, S., Kim, J., Tran, A., & Hellas, A. (2023). Comparing Code Explanations Created by Students and Large Language Models. *Proceedings of the 2023 Conference on Innovation and Technology in Computer Science Education V. 1*, 124–130. https://doi.org/10.1145/3587102.3588785

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2021). *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks* (arXiv:2005.11401). arXiv. https://doi.org/10.48550/arXiv.2005.11401

Litman, D. (2016). Natural Language Processing for enhancing teaching and learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, *30*(1). https://doi.org/10.1609/aaai.v30i1.9879

Luckin, R., Underwood, J., Du Boulay, B., Holmberg, J., Kerawalla, L., O'Connor, J., Smith, H., & Tunley, H. (2006). Designing Educational systems Fit for Use: A Case Study in the application of human Centred Design for AIED. *I J Artificial Intelligence in Education*, *16*, 353–380.

Lv, Z. (2023). Generative artificial intelligence in the metaverse era. *Cognitive Robotics*, *3*, 208–217. https://doi.org/10.1016/j.cogr.2023.06.001

Mastery, A. H. J. (2023). *Build and Deploy Your Own ChatGPT AI Application That Will Help You Code* [JavaScript]. https://github.com/adrianhajdin/project_openai_codex

Milano, S., McGrane, J. A., & Leonelli, S. (2023). Large language models challenge the future of higher education. *Nature Machine Intelligence*, *5*(4), 333–334. https://doi.org/10.1038/s42256-023-00644-2

Miltsakaki, E., & Troutt, A. (2008). Real time web text classification and analysis of reading difficulty. *Proceedings of the third workshop on innovative use of NLP for building educational applications*, 89–97.

NLP Connect (2023). *Vit-gpt2-image-captioning*. https://doi.org/10.57967/HF/0222

OpenAI (2023d, August 26). *OpenAI Platform*. https://platform.openai.com/docs/guides/embeddings/use-cases

OpenAI (2023b, August 26). *OpenAI Models*. https://platform.openai.com/docs/models

OpenAI (2023a). *GPT-4 Technical Report* (arXiv:2303.08774). arXiv. https://doi.org/10.48550/arXiv.2303.08774

OpenAI (2023c, August 26). *OpenAI Platform*. https://platform.openai.com/docs/guides/embeddings/what-are-embeddings

Petersen, S. E., & Ostendorf, M. (2009). A machine learning approach to reading level assessment. *Computer Speech & Language*, *23*(1), 89–106. https://doi.org/10.1016/j.csl.2008.04.003

Pitler, E., & Nenkova, A. (2008). Revisiting readability: A unified framework for predicting text quality. *Proceedings of the 2008 conference on empirical methods in natural language processing*, 186–195.

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). *Improving language understanding by generative pre-training*.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, *1*(8), 9.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022). *Robust Speech Recognition via Large-Scale Weak Supervision* (arXiv:2212.04356). arXiv. https://doi.org/10.48550/arXiv.2212.04356

Renz, A., & Krishnaraja, S. (2020, Dezember). *Toward Responsible, Human-Centered AI in EdTech*.

Renz, A., & Vladova, G. (2021). Reinvigorating the discourse on human-centered Artificial Intelligence in Educational technologies. *Technology Innovation Management Review*, *11*(5). https://doi.org/10.22215/timreview/1438

Richards, K., & Hemphill, M. (2017). A practical guide to Collaborative Qualitative Data Analysis. *Journal of Teaching in Physical Education*, *37*, 1–20. https://doi.org/10.1123/jtpe.2017-0084

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). *High-Resolution Image Synthesis with Latent Diffusion Models* (arXiv:2112.10752). arXiv. https://doi.org/10.48550/arXiv.2112.10752

Ruan, S., Jiang, L., Xu, J., Tham, B. J. K., Qiu, Z., Zhu, Y., Murnane, E. L., Brunskill, E., & Landay, J. A. (2019). QuizBot: A Dialogue-Based Adaptive Learning System for Factual Knowledge. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–13. https://doi.org/10.1145/3290605.3300587

Sarsa, S., Denny, P., Hellas, A., & Leinonen, J. (2022). Automatic Generation of Programming Exercises and Code Explanations using Large Language Models. *Proceedings of the 2022 ACM Conference on International Computing Education Research V.1*, 27–43. https://doi.org/10.1145/3501385.3543957

Smith, G. G., Haworth, R., & Žitnik, S. (2020). Computer Science meets Education: Natural Language Processing for Automatic Grading of Open-Ended questions in eBooks. *Journal of Educational Computing Research*, *58*(7), 1227–1255. https://doi.org/10.1177/0735633120927486

Smolansky, A., Cram, A., Raduescu, C., Zeivots, S., Huber, E., & Kizilcec, R. F. (2023). Educator and Student Perspectives on the Impact of Generative AI on Assessments in Higher Education. *Proceedings of the Tenth ACM Conference on Learning @ Scale*, 378–382. https://doi.org/10.1145/3573051.3596191

Strauss, A., & Corbin, J. (1998). *Basics of qualitative research techniques*.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., & Scialom, T. (2023). *Llama 2: Open Foundation and Fine-Tuned Chat Models* (arXiv:2307.09288). arXiv. https://doi.org/10.48550/arXiv.2307.09288

Vallis, C., Wilson, S., Gozman, D., & Buchanan, J. (2023). Student perceptions of AI-Generated avatars in Teaching Business Ethics: We might not be impressed. *Postdigital Science and Education*. https://doi.org/10.1007/s42438-023-00407-7

Wang, J., Yi, X., Guo, R., Jin, H., Xu, P., Li, S., Wang, X., Guo, X., Li, C., Xu, X., Yu, K., Yuan, Y., Zou, Y., Long, J., Cai, Y., Li, Z., Zhang, Z., Mo, Y., Gu, J., & Xie, C. (2021). Milvus: A Purpose-Built Vector Data Management System. *Proceedings of the 2021 International Conference on Management of Data*, 2614–2627. https://doi.org/10.1145/3448016.3457550