big data and cognitive computing

MDPI

Article

Knowledge-Based and Generative-AI-Driven Pedagogical Conversational Agents: A Comparative Study of Grice's Cooperative Principles and Trust

Matthias Wölfel [1,2,*], Mehrnoush Barani Shirzad [2], Andreas Reich [2], Katharina Anderer [1,3]

[1] Faculty of Computer Science and Business Information Systems, Karlsruhe University of Applied Sciences, Moltkestr. 30, 76131 Karlsruhe, Germany; katharina.anderer@kit.edu
[2] Faculty of Business, Economics and Social Sciences, University of Hohenheim, Schloss Hohenheim 1, 70599 Stuttgart, Germany
[3] Faculty of Computer Science, Institut for Anthropomatics and Robotics (IAR), Karlsruher Institut for Technology (KIT), Kaiserstr. 12, 76131 Karlsruhe, Germany
* Correspondence: matthias.woelfel@h-ka.de

Abstract: The emergence of *generative language models* (GLMs), such as OpenAI's ChatGPT, is changing the way we communicate with computers and has a major impact on the educational landscape. While GLMs have great potential to support education, their use is not unproblematic, as they suffer from hallucinations and misinformation. In this paper, we investigate how a very limited amount of domain-specific data, from lecture slides and transcripts, can be used to build knowledge-based and generative educational chatbots. We found that knowledge-based chatbots allow full control over the system's response but lack the verbosity and flexibility of GLMs. The answers provided by GLMs are more trustworthy and offer greater flexibility, but their correctness cannot be guaranteed. Adapting GLMs to domain-specific data trades flexibility for correctness.

Keywords: conversational agent; chatbot; education; large language model; generative language model; retrieval augmented generation; generative AI; digital tutor; digital assistant

Citation: Wölfel, M.; Shirzad, M.B.; Reich, A.; Anderer, K. Knowledge-Based and Generative-AI-Driven Pedagogical Conversational Agents: A Comparative Study of Grice's Cooperative Principles and Trust. *Big Data Cogn. Comput.* **2024**, *8*, 2. https://doi.org/10.3390/bdcc8010002

Academic Editor: Carson K. Leung

Received: 27 November 2023
Revised: 22 December 2023
Accepted: 22 December 2023
Published: 26 December 2023

1. Introduction

The emergence of *generative language models* (GLMs), at the latest since the launch of ChatGPT by OpenAI in November 2022, is drastically changing the way we communicate with computers. This is due to the capability of GLMs to produce human-like text by generating a sequence of words in response to another text. ChatGPT—short for chat generative pretrained transformer—is a large-language model-based chatbot. According to a study conducted by Intelligent.com that surveyed 3,017 high school and college students in the United States, nearly all of them have replaced some of their tutoring sessions with ChatGPT [1]. This demonstrates that profound technological changes are having a major impact on education already. Large language models are trained on a broad variety of text corpora and thus reflect statistical knowledge of word sequences on various topics. Consequently, they can produce coherent and contextually appropriate responses to various given tasks in natural language. This offers great potential in various fields, including education. An example of how to use ChatGPT in the classroom is even given directly by OpenAI [2]. While GLMs can produce surprisingly accurate results that reflect human reasoning in many cases, they also suffer from hallucinations. *Hallucination*, in this context, is defined as a sequence of human-like text that sounds plausible but contains information that is made up [3]. While hallucinations can be useful in poetry and literature, they render GLMs unsuitable in many domains where truthful, reliable, and verifiable answers are required, such as education. To overcome this drawback, the pretrained models can be augmented with domain-specific knowledge that is available in the form

of slides, textbooks, or audio transcripts. Including domain-specific knowledge might be helpful to provide coherent, natural answers that do not alter the original intent or are even made up. This is crucial in an academic context, where it is vital to ensure that content is entirely reliable. To achieve this, we propose and investigate different approaches that allow direct extraction of critical information such as definitions that need to be restated word by word but at the same time allow for the reformulation of less critical information while remaining close to the ground truth. Our study therefore aims to investigate the advantages and disadvantages of different chatbot systems concerning their suitability for an academic context.

Before the introduction of GLMs, the creation of educational chatbot systems was tedious, as it required specific knowledge and hand labor. For each course, it was required to redevelop and reassemble context, and thus the application of chatbot systems for learning was quite limited. To overcome this drawback, Wölfel proposed a conversational tutoring system dubbed *PEdagogical conversational Tutor*, or PET for short, which can be trained exclusively on prestructured PowerPoint slides to reduce the amount of handcrafted dialog [4]. The general architecture of the PET is shown in Figure 1. It includes *data sources* (on the left), *components* (in the center) that are required for the assistive system, and the *human–machine interfaces* (on the right). The individual components analyzed in this publication are described in the following sections.

Following its introduction, it has been used at the University of Applied Sciences Karlsruhe since 2019, as well as at the University of Hohenheim since 2022, to support students in lectures on topics such as engineering and social sciences. The PET can be interacted with via text messages or voice commands and speech synthesis. The system is capable of providing answers to given content-specific questions and evaluating the student's knowledge fully automatically. We found that students utilize the system differently depending on the onboarding process [5]. The onboarding provided two interfaces, a graphical user interface and a chat widget. This leads to students utilizing it either as a search engine, entering single words or word combinations, or adopting a more conversational approach by inputting complete sentences and phrases.

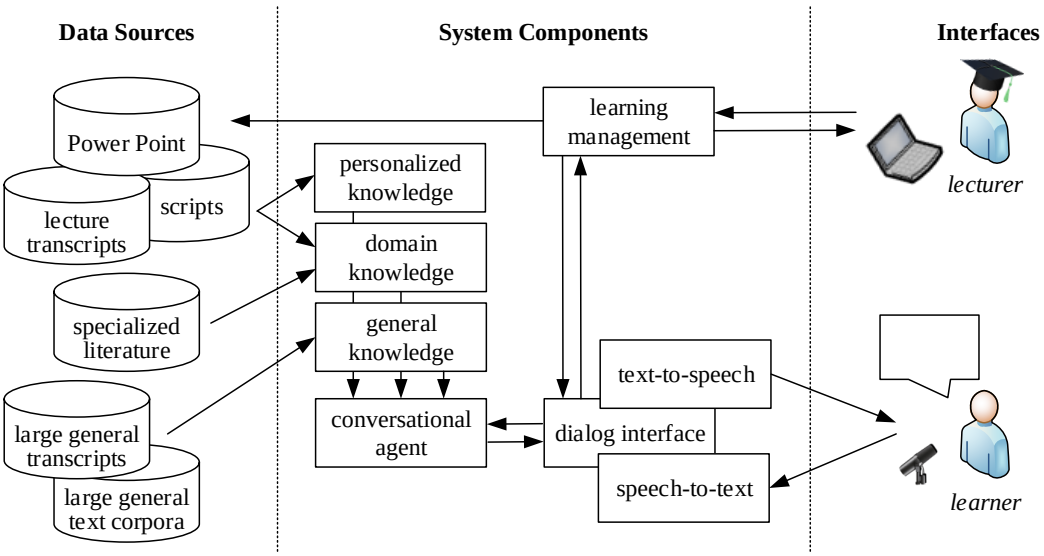

Figure 1. System overview with all components.

The user interface of the PET offers two different layouts: one to handle the information provided by the instructor and the other to handle the dialog with the learners; see Figure 2. By showing a slide that belongs to the text dialog, the system provides additional information not covered in textual form, such as images, diagrams, and charts. In addition, the slides provide context to verify the validity of the answers given by the chatbot and provide a connection to the original lecture the learners might have attended before consulting

the assistant. The users have multiple options to interact with the assistant, either by text or voice, pointing with the mouse or touch, or even through arrow keys to navigate between slides. This paper builds on the PET system, investigating the validation of different chat systems in the context of guided learning.

Figure 2. A conversation with the pedagogical conversational agent PET.

The main contributions are a detailed analysis of knowledge-based versus generative AI chatbots concerning different aspects like truthfulness or trust. Furthermore, this paper explores how large language models can be tailored to enhance the relevance of lecture content.

This paper is structured as follows: The introduction presents the motivation to use chatbots in education and gives a general system overview of such systems. Section 2 starts with a brief review of chatbots and compares knowledge-based vs generative approaches. Data sources are discussed in Section 3, including unlabeled and annotated data. Section 4, Section 5, and Section 6 refer to Section 1 and detail how the conversational agents can be realized. An evaluation of the different systems, investigating Paul Grice's Cooperative Principle, and trust can be found in Section 7. The paper ends with a discussion and overview of the limitations in Section 8 and provides a conclusion and future work in Section 9.

2. Review of Chatbots

Chatbots, also known as conversational agents, natural dialog systems, or virtual assistants, have a rich history that dates back several decades. The idea behind chatbots was to simulate human conversation using machines. The origins of chatbots can be traced back to the 1950s when computer scientists began exploring the possibilities of *artificial intelligence* (AI). In 1950, Alan Turing proposed the Turing Test, a benchmark for determining a machine's ability to exhibit human-like intelligence in conversation. In the 1960s, Joseph Weizenbaum designed ELIZA, a program that kept users engaged through text dialogues by rephrasing their statements as questions. ELIZA, despite being simplistic, revealed the potential of natural language processing and initiated public interest in chatbots. Throughout the 1980s and 1990s, significant progress was made in natural language processing and AI research. However, it was not until the 21st century that chatbot applications became commercially established. In 2011, Apple introduced Siri, a voice-activated virtual assistant, and in 2016, Google launched Google Assistant.

Since approximately 2016, there has been a noteworthy increase in the number of publications on chatbots in general, as well as the number of publications using the terms chatbot and education. This growth can largely be attributed to the advancements in *natural language processing* (NLP) and *machine learning* (ML) methods. Before 2016, chatbots

were primarily rule-based, following predefined rules and decision trees [6,7]. Retrieval-based chatbots became popular after 2016. These chatbots utilize predetermined responses from a fixed dataset or knowledge base and select the most suitable response based on the input [8]. Retrieval-based chatbots use ML algorithms to match user queries with appropriate responses. However, they were unable to generate original responses or achieve the same level of context and semantic similarity understanding as generative chatbots do today.

Starting in 2019, generative chatbots have emerged, with notable breakthroughs from OpenAI, such as GPT-2 [9], GPT-3 [10], and more recently, GPT-4 [11]. GPT models are trained on a large corpus of text to predict the subsequent word in a sentence. This procedure aids in generating new words that are statistically likely. Following pretraining, the model can be fine-tuned for specific tasks, such as translation, summarization, or question answering, to make it more task-specific and improve its performance on those tasks [12].

ChatGPT, as a subsequent advancement of the GPT models, underwent specific fine-tuning to generate conversations resembling those of humans. The training dataset for ChatGPT comprised conversations from a wide variety of sources, as well as human feedback on the quality of the conversations. This enabled the model to gain a better sense of the flow of the conversation, to better maintain context over multiple turns, and to generate coherent and contextually appropriate responses.

Purely generative methods in chatbot development can have drawbacks, including hallucinations, a lack of explainability, biases, and difficulties in verifying model-generated information [13]. In academic settings, these shortcomings are considered unacceptable, and approaches such as specific prompt engineering, fine-tuning, and document embedding have been proposed to mitigate hallucinations and ensure that the model adheres to the given context [14,15]. Recent advancements, including LoRA [16] and prompt-tuning [17], as well as user-friendly frameworks such as the LLM-Adapters developed by [12] or HuggingFace's PEFT library [18], have made fine-tuning more efficient and accessible to researchers with limited computational resources.

Another potential limitation of purely generative approaches lies in their limited ability to control the flow of the dialog and to guide the conversation. To address this, Wang et al. [19] suggest matching teaching strategies to the tutor's response, such as offering hints in their response.

2.1. Chatbots for Education and Learning

Text-related AI systems have a long history of offering supplementary services to students. Although these systems can benefit any student, they may enhance accessibility for students with special needs through text-to-speech and vice versa. While certain services enhance lectures with transcriptions [20,21] or translations [22] of what the lecturer has said, others focus on facilitating conversational interactions between students and the system. These conversational interactions are transforming education by improving communication, engagement, and personalization for students and educators alike. They offer a range of use cases that provide instant feedback and support, available 24/7, and of varying complexity, providing support in the following areas:

- *Personalized learning* can tailor educational content to meet individual learning styles and pacing, accommodating diverse student needs [23],
- *Course queries* providing information about courses, fees, syllabus, and more [24–28];
- *Admission processes* by guiding students through the enrollment [29];
- *Student feedback* by gathering and analyzing answers to specific questions or by analyzing student sentiment through regular interactions and feedback [30,31];
- *Student engagement* by keeping students engaged with personalized messages, progress tracking, and interactive activities [30,32–34];
- *Teaching and tutoring* by acting as a tutor answering topic-specific questions and providing individualized messages and feedback [4,30];

- *Examinations and assessments* by automating course exams and assessments, providing progress reports, and maintaining course quality [4].

In the following, we limit our focus to systems offering teaching, tutoring, examinations, and assessment capabilities.

Noteworthy educational chatbots before the breakthrough of LLMs include AutoTutor [30,35], or the Tactile Language and Culture System by [36], as listed in the review by [33]. Most existing implementations are designed for specific tasks, such as question generation [37,38], increasing social engagement [32,33], language learning [39] or math tasks [40,41], as well as feedback systems like AutoTutor [30,35]. AutoTutor [30,35] is a tutoring system designed to ask students a question, give feedback on the quality of their answers, and engage them in more elaborate answers. Modeling a student's cognitive state is performed by semantic matching algorithms that compare the student's answer with the expected answer [35].

While the number of research papers envisioning powerful chatbots for academic settings is increasing tremendously, e.g., [42–45], implementations involving the use of LLMs are still rare [31,41,46,47]. As an example, [46] used GPT-4 to modify the context of their exercises based on user interests and thereby tried to personalize the learning experience and increase motivation.

Ref. [25] introduced Eduino, a system specifically designed for a lecture at the University of Bergen. Eduino utilized JavaScript and JSON to structure the course and integrated various features, such as quizzes and a help function. While our model also uses JSON to structure and retrieve information from the lecture, their approach intentionally avoided using NLP tools, as they were afraid that this might interfere with the lecturer–student relationship [25]. Their chatbot was therefore limited to simple queries, and the users had to be familiar with the exact structure of these queries. This design choice reflects the trade-off between control and generalizability, where a more controlled system may have limited capabilities but can potentially be more trustworthy.

The educational chatbot systems found in the literature are all text-based only, to the best of our knowledge. There are currently no chatbot systems that guide students through a lecture by directing them to relevant pages for their questions as performed by our system.

A systematic literature review by [48] indicates that the main objectives of implementing chatbots in education are skill improvement, efficiency of education, student motivation, and availability of education.

2.2. Comparing Knowledge-Based vs. Generative Approaches

Chatbots can be implemented in many different ways. We limit our scope to knowledge-based and generative AI chatbots:

Knowledge-based methods rely on structured data (and often require detailed, manual annotation), detailed descriptions of data handling, and sometimes handcrafted features. They involve breaking down language into its smallest parts (such as words or sentences) and analyzing those parts in a step-by-step process. Tasks such as part-of-speech tagging, named entity recognition, and parsing sentences into their grammatical structures are tackled independently. These methods tend to struggle with understanding context and capturing the ambiguities and nuances of human language. Knowledge-based chatbots can be realized using rule-based logic or AI, e.g., for intent classification.

Generative-AI-based methods rely on large amounts of unstructured data and facilitate LLM and machine learning. To understand and generate text, a word sequence is predicted based on the previous word sequence. This covers high language complexity and statistical knowledge. Generative AI approaches can therefore generate fluent, coherent text and understand more nuanced and complex language tasks. Unlike knowledge-based methods, they do not require explicit rules or handcrafted features.

Each approach has its strengths and weaknesses. While large language models can capture subtleties and handle a wider range of language tasks, they can also generate

text that is misleading or inappropriate, or have difficulty with tasks that require a deep understanding of the text. On the other hand, traditional methods can be more interpretable and controllable, but they require more manual work and may not handle complex language tasks as well. Table 1 summarizes the advantages and disadvantages of the two approaches.

2.3. Ethical Concerns of Chatbots in Education

The ethical implications of conversational educational agents need to be discussed, as they have a strong influence on their actual use. Ref. [49] lists several ethical concerns, mostly related to the use of LLMs, including privacy issues, potential bias, lack of transparency, and lack of creativity. Educational agents could be used to cheat on assessments [49]. Furthermore, the use of an AI agent could lead to a lack of full understanding of course content if students rely too heavily on the use of AI [49]. A study by [50] identified other potential issues, specifically for the use of ChatGPT in education. By conducting user interviews, they gathered ethical concerns from the users themselves. These included the potential to breed laziness and reduce critical thinking among students, lack of trustworthiness, and being prone to error [50]. Using LLMs to implement a chatbot can also lead to copywriting issues [43], especially if the source of information remains unknown. Teachers may also rely too heavily on the model [43] when creating tasks or evaluating assessments. The cost of maintaining an LLM chatbot as well as its sustainable use are two other ethical considerations in [43].

Table 1. Comparing knowledge-based vs generative AI Approaches.

Feature	Knowledge-Based Chatbot	Generative AI Chatbot
Functionality	Conversational agent that focuses on end-to-end dialogue management and allows to customize and control the behavior of the chatbot [51]	Conversational agent that facilitates LLMs to generate human-like text responses based on given prompts [52]
Techniques	Includes intent examples, entity annotations, dialogue flows, and employs AI to learn the dialogue policies [6].	Pretrained on a vast amount of text data [10] and uses generative approaches to produce text responses
Advantages	• Domain expertise [14] • Goal-oriented interactions • Intent-oriented dialog [8] • Contextual persistence • Data privacy	• Wide range of knowledge [10] • Contextual understanding and generation [53]
Disadvantages	• Limited understanding of context • Development complexity • Lack of flexibility and adaptability [54] • Dependency on labeled training data • Requires well-defined queries • Releis on expert's annotations	• Hallucinations • Computation-intensive training runs [53] • Lack of domain expertise • Over-reliance on training data • Ethical concerns • Difficult to control output [53]

3. Data Sources

High-quality textual data are the cornerstone of most conversational agents. These data can be broadly divided into two categories: general data and specialized data. *General data* include text from web pages, books, and conversation transcripts, improving language modeling and overall generalization capabilities. In contrast, *specialized data* consist of content-specific data, such as scientific publications, which enable either the development of knowledge-based systems or the fine-tuning of LLMs for specific tasks.

The knowledge-based PET system is based on semiannotated PowerPoint slides. The generative AI PET system includes various knowledge sources, such as preacquired data inherent in the GPT models, plain text, or semiannotated PowerPoint slides, and might also include lecture transcripts.

3.1. Unstructured Data

The success of LLMs can be largely attributed to their pretraining on a rich corpus of text data from a variety of sources. These text corpora span different genres, with sources including text from websites, books, articles, and forums, and in the case of multilingual systems, different languages too [10]. However, pretrained systems that do not have access to the Internet or other real-time data are limited to the information contained in the training data. For example, the training data for GPT-3 were curated from a combination of five different corpora, each of which was given a specific weight in the training process. These five corpora included Common Crawl [55], WebText2, Books1, Books2 [56], and Wikipedia. To date, OpenAI has not disclosed the detailed datasets used to train GPT-4 [11]. GPT-3 contains data until September 2021 and GPT-4 contains data until January 2022.

3.2. Annotated Lecture Slides

PowerPoint slides, which are widely used by lecturers, serve as a reliable source of up-to-date information for most lectures. Therefore, extracting the content of the slides can be a very valuable source of information. Unlike second-hand sources, such as textbooks written by others, lecture slides accurately reflect the outline and structure of the course, its content, and its terminology. Lecture slides also have an advantage over textbooks in that they are tightly organized into a specific structure that can, for instance, be used for the generation of concept maps [57].

Since the structure and information provided in lecture slides may not be sufficient to find good relationships between specific text passages and the query, the slides are augmented with annotations. These annotations provide valuable additional information that allows the system to identify and classify entity classes such as definitions, quotes, and examples.

All PowerPoint slides used are annotated according to the annotation guidelines suggested in [4]. Examples for annotations are ‘Chapter’, ‘Topic’, ‘Subtopic’, ‘Keywords’, ‘Abbreviation’, ‘Definition’, ‘Example’, ‘Objective’, ‘Quote’, ‘Equation’, ‘Image’, ‘Table’, ‘Diagram’, and ‘Slide Number’. On average, 6.7 annotations are used per slide, with a total of eighteen distinct annotation labels available. The system is designed for the easy implementation of additional labels. Since instructors use different labels for the entity classes, the given ontology can be extended. An example of an annotated PowerPoint slide is shown in Figure 3. The inclusion of semiannotations is useful because different types of entities can be handled and explicitly searched for in the PET system.

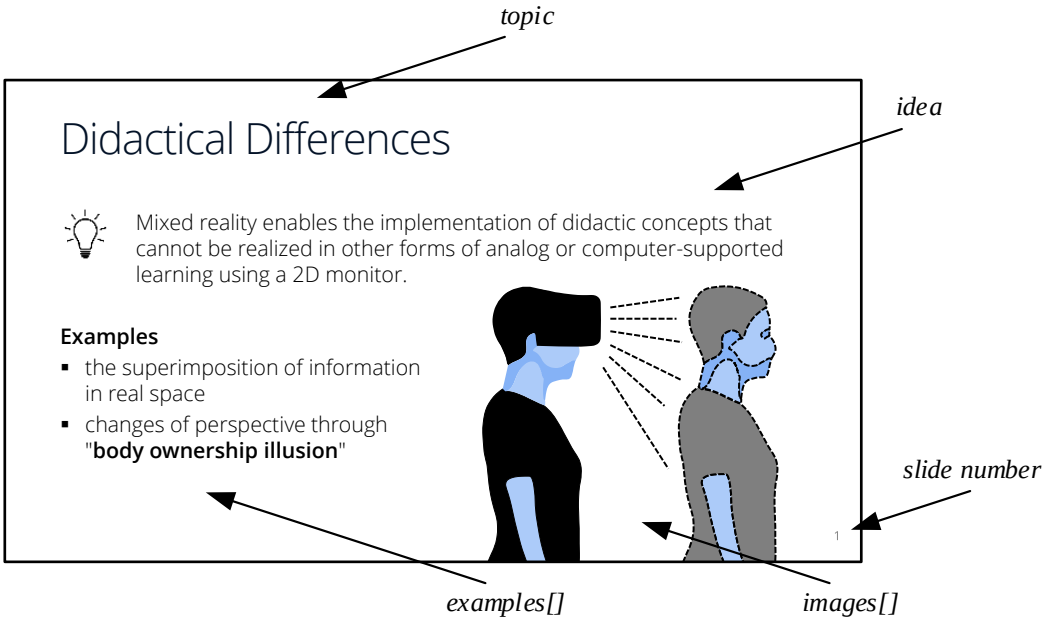

Figure 3. Example of an annotated PowerPoint slide.

To make the preparation of slides as easy as possible for the lecturers, semantically annotated slide templates were developed, which predefined didactically adequate learning paths (as macrolearning paths) for the lecturers through their set of slides. The use of the developed templates avoids additional annotation efforts in the preparation step of the slides to be used by the systems.

3.3. Lecture Audio Transcripts

Automatically transcribing lectures poses specific challenges, as lecture speech differs from other forms of speech due to several factors, including speaking style, the acoustic environment, the number and quality of acoustic channels, and vocabulary and language use [21]. For instance, lectures are less structured and more spontaneous compared with carefully prepared forms of speech, such as broadcast news. Studies have shown that lecture speech shares characteristics with conversational speech, including word contractions, reductions, extraneous filler words, nonlexical filled pauses, partial words, and false starts.

To generate transcripts from lecture videos, we use the automatic speech recognition tool Whisper [58] by OpenAI. It shows good performance in converting spoken language into written text, with a low word error rate (WER) of 6.4% when transcribing German [58]. In the lectures used, the mean average WER is 3.2%.

4. Knowledge-Based Pedagogical Conversational Agents

Knowledge-based chatbots, also referred to as knowledge-driven or knowledge-powered chatbots, rely on a structured repository of factual information, rules, and predefined responses to provide users with responses to their queries. These chatbots are designed to handle specific domains or areas of expertise and are powered by a knowledge base that includes articles and documents. The primary benefit of a knowledge-based chatbot is that it undergoes training using knowledge that is already organized and structured. This knowledge base acts as a central source of information, facilitating the chatbot's access to relevant data and enabling it to respond to user inquiries quickly and accurately. The knowledge-based components are detailed in Figure 4.

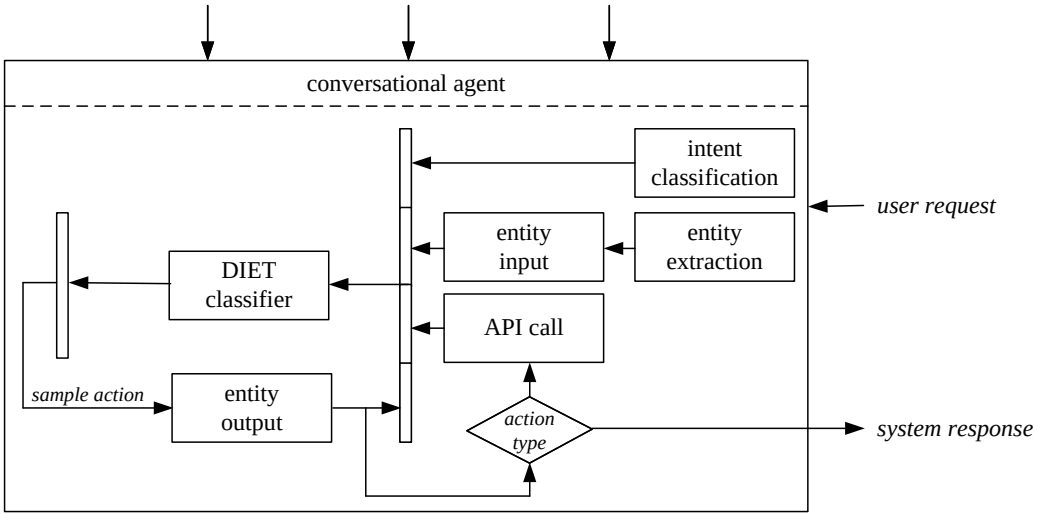

Figure 4. PET with knowledge-based component in Figure 1 based on [59].

It is built around the open-source chatbot development framework RASA [60]. This framework allows to build, test, and deploy conversational interfaces. The PET system relies on RASA's *natural language understanding* (NLU) components to process user input, classify intents, extract entities, and select appropriate responses [51]. The NLU pipeline consists of several components that act sequentially, including the language model, tokenizer, feature analyzer, intent classifier, entity extractor, and response selector. RASA provides example configurations of the NLU pipeline. However, various components have been modified to meet the specific requirements of the application, to improve the quality of the overall system, and to add microservices to provide additional functionality, such as a structured knowledge evaluation of students using semantic similarities. To optimize the NLU pipeline of our virtual assistant, a test set of manually derived test stories was carried out, and a comparative analysis of the F1 scores of six different base models was performed. The use of pretrained BERT embeddings [61] with LaBSE weighting yields an F1 score of approximately 95%, while, e.g., the use of Spacy models results in a comparatively lower performance of approximately 80%. The application of a rigorous five-fold cross-validation methodology enhances the robustness and statistical validity of the obtained results.

The PET system autonomously constructs a dialog flow based on semiannotated slides, as described in Section 3.2. At the same time, additional information is stored in a SOLR database [62], and slide images are extracted for presentation in the React-based front-end. Central to the system's functionality are user-defined actions that are triggered under certain conditions, such as topic searches and student self-tests. The DIET classifier [63] is instrumental in determining the optimal action for each user request. This architecture involves training a recurrent neural network. The training uses generated intents derived from handcrafted rules and slide templates. Accurate categorization improves the system's understanding of user requests and ensures effective interactions. Entities are extracted using regular expressions. If there is a direct match, the system navigates the user directly to the relevant slides. If no exact matches are found, a vector search is used to determine the closest match or to pose a selection of possible topics.

5. Generative-AI-Based Pedagogical Conversational Agents

Generative-AI-based chatbots use GLMs to establish a conversation. In the context of a pedagogical conversational agent, the general model must be prepared for the task through prompt engineering. Prompt engineering is a technique for fine-tuning language models by providing targeted prompts [64]. A prompt may include the desired role, goal, scope, and target audience, and it may include additional documents (see Section 6) to provide evidence for the answers. We used the following prompt in German, followed by the actual question after the colon:

```
Sie sind ein Lehrassistent mit langjähriger Berufserfahrung im Bereich
der intuitiven und perzeptiven Benutzungsschnittstellen. Ihr Ziel ist
es, eine möglichst wahrheitsgetreue Antwort auf die Fragen der Studier-
enden im Bereich der intuitiven und perzeptiven Benutzungsschnittstelle
zu geben und dabei nicht zu sehr auszuschweifen, sondern kurz und
präzise zu antworten. Die Zielgruppe sind Studierende. Die Frage lautet:
```

The prompt can be translated into English as follows:

```
You are a teaching assistant with many years of professional experience
in the field of intuitive and perceptual user interfaces. Your goal is
to answer students' questions about intuitive and perceptual user
interfaces as truthfully as possible, not to ramble, but to answer
briefly and precisely. The audience are students. The question is:
```

The system was prompted to behave like a teaching assistant while remaining relatively generic and open-ended, as more specific prompts would limit the system's capabilities too much and the results would lose their generalizability. Our GLM-based system incorporates ChatGPT [3] versions 3.5 and 4.0 as the kernel for user conversation and information retrieval from documents. The structure of the generative AI component using ChatGPT without additional data is depicted in Figure 5.

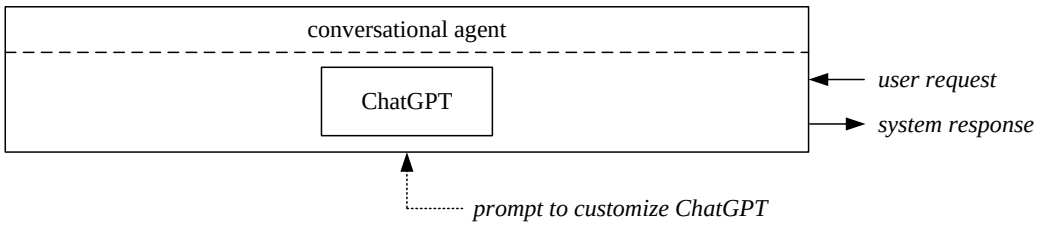

Figure 5. PET with generative AI component (without adaptation) in Figure 1.

6. Adapting Generative AI Pretrained Transformers to Particular Context

Even though ChatGPT is prompted to a particular lesson and target audience, it might miss particular lecture information. Therefore, in addition to employing GPT independently without any supplementary data, we employed GPT in conjunction with presentation slides, as well as presentation slides and the transcripts from the actual presentation, as depicted in Figure 6. By adding these sources, as elaborated in Section 3.1, ChatGPT benefits from access to our custom dataset, extending its pre-existing general knowledge base. The input data, presented as textual information, are ingested into our system using the LangChain framework [65]. LangChain, a versatile tool designed to harness the capabilities of GLMs, is responsible for data loading. Instead of using LangChain's capability to extract information from unstructured data sources, we used the prestructured information to provide well-prepared information. Therefore, we traded in more flexibility in order to use unlabeled data sources to have access to labels such as 'definition' or 'quote'.

In the subsequent phase of our workflow, embeddings are generated, and these embeddings are subsequently utilized by the vectorizer module to construct a comprehensive vector database encompassing the provided source documents. The resulting vector database is systematically stored in a dedicated directory for future reference. For the efficient search and retrieval of documents corresponding to user queries, the specialized vector database system Chroma was deployed. This database leverages embeddings, representing documents as vectors in a multidimensional space. The semantic similarity between user queries and the document vectors is employed to retrieve the most contextually relevant information.

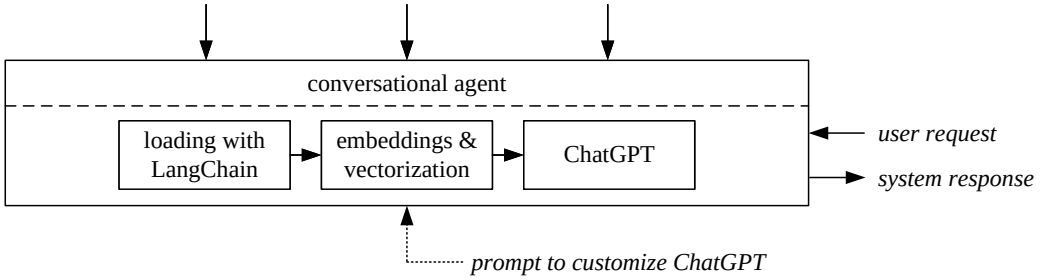

Figure 6. PET with generative AI component and adaptation in Figure 1

7. System Evaluation

In this section, we want to test different implementations of the PET system on their effectiveness and reliability in delivering accurate and informative responses. Evaluating educational chatbots goes beyond measuring their technical functionalities; it delves into aspects of reproducibility, cooperative principles, trust in the answers, and the incorporation of user requests.

To evaluate the quality of the different systems, we rely on the Cooperative Principle introduced by Paul Grice to describe how people use language in context to convey meaning effectively. Grice introduced this principle in his influential 1975 article titled 'Logic and Conversation', ref. [66] where he explored how conversational implicatures, the unspoken meaning conveyed beyond the literal interpretation of words, play a crucial role in successful communication.

Grice's Cooperative Principle and its associated maxims are rooted in the idea that people generally communicate with the intention of being cooperative and mutually understanding. According to Grice, successful communication requires speakers and listeners to follow four maxims, which are general guidelines for how conversation should be conducted. We extended Grice's Cooperative Principle to the Maxim of Trust, as trust in chatbot answers is crucial for both students and educators, as it directly affects their willingness to rely on the provided information and guidance. In the following, we give a short explanation of the different maxims:

- **Maxim of Quality:** This maxim emphasizes truthfulness in communication. Speakers are expected to offer information that is true or for which they have sufficient evidence. They should refrain from providing false or unsupported statements that might mislead the listener.
- **Maxim of Quantity:** This maxim suggests that speakers should provide the right amount of information, neither too much nor too little, to convey their intended meaning. In other words, speakers should be as informative as required but avoid being excessively verbose or overly vague.
- **Maxim of Relation:** The maxim of relation pertains to relevance in conversation. Speakers should contribute information that is pertinent to the ongoing topic or the context of the conversation. Irrelevant or off-topic contributions can hinder effective communication.
- **Maxim of Manner:** The maxim of manner addresses the clarity and effectiveness of expression. Speakers should strive to be clear, orderly, and concise in their speech. Ambiguous or convoluted expressions might confuse the listener and disrupt communication.
- **Maxim of Trust:** This maxim expresses the credibility of the speaker. Whether we believe in an answer or not is influenced by different factors. If the listener does not trust the speaker, this impacts her or his willingness for further interaction.

However, it is important to note that the Cooperative Principle is not always followed rigidly, and sometimes people intentionally flout these maxims to create humor, irony, or rhetorical effects. In such cases, the listener relies on their pragmatic competence, the ability to understand implied meanings and intentions, to interpret the speaker's actual meaning.

In summary, the Cooperative Principle posits that successful communication relies on participants' cooperation and adherence to the maxims of quantity, quality, relation, and manner. By being truthful, informative, relevant, and clear in their communication, speakers contribute to effective conversations where meaning is conveyed efficiently and accurately.

Furthermore, the cooperative principles, derived from Paul Grice's seminal work 'Logic and Conversation', play a significant role in the evaluation of educational chatbots. These principles emphasize that participants in a conversation cooperate and strive to be truthful, informative, relevant, and clear in their communications. Educational chatbots should adhere to these principles, as they directly impact the quality of interactions and the user's perception of the chatbot's competency. A chatbot that respects cooperative principles fosters a sense of trust and engagement, enhancing the overall learning experience.

7.1. Generation of Test Questions

The test data include 281 questions to evaluate and compare the PET system and PET using three variations of ChatGPT. Approximately 200 of these questions were generated with ChatGPT from the transcripts of the lectures using LangChain and prompt engineering. The remaining questions were extracted from the conversation file using the RabbitMQ tool [67], which includes questions asked by student end-users. These questions contain typos and speech recognition errors and are useful to gain insight into the robustness of the systems.

All the collected and generated questions were evaluated by experts to ensure their relevance to the lecture. Each system is presented with the test questions, and the responses generated by the system are collected for evaluation and comparison. To maintain a balanced test set, similar questions were removed.

7.2. Evaluation of Given Answers According to the Maxims

Table 2 evaluates the characteristics and quality of the different systems based on their generated responses. The metrics used in the table provide insights into the system's abilities to generate informative (Quantity), relevant (Relation), clear (Manner), and truthful (Quality) responses, along with their tendencies to ask for requests and whether they leave certain inputs unanswered. The comparisons allow researchers or developers to assess the relative strengths and weaknesses of each system and make informed decisions about their usage in specific applications. The column # *words* represents the average number of words in responses. The column *request* measures the percentage of responses that contain a request for information or action. The column *no answer* indicates the percentage of responses that do not provide an answer to the user's query.

Ratings for quality, quantity, relation, and manner were carried out by two evaluators who were very familiar with the lecture content, while the rating for trust was performed by two evaluators without any lecture-specific knowledge to not be influenced by the correctness of the answer. In total, around 500 ratings were given for each maxim.

To determine the significance of the differences among the various systems, we utilized Kruskal–Wallis tests to account for imbalanced data for all maxims, with responses rated on a scale from 1 to 7 (with 7 representing the highest score). Following this, we conducted a post hoc Dunn's test for pairwise comparisons, with a Bonferroni correction factor to account for multiple comparisons. Scores were calculated while excluding requests and unanswered questions, except for those related to trust. A more thorough examination of the influence of different response types on trust is presented in the relevant section.

Table 2. Results on various aspects of the different systems. The average score value is given for the different evaluations.

	# Words	Request	No Answer	Quality	Quantity	Relation	Manner	Trust
Knowledge-Based	27	27.0%	0.0%	6.54	5.40	5.83	5.20	3.59
Generative AI (GPT 3.5)								
No adaptation	108	0.0%	1.7%	5.42	5.14	5.07	5.49	5.96
Slides	93	3.0%	13.3%	5.87	5.27	5.46	5.45	5.04
Slides and Transcripts	79	0.0%	8.3%	5.89	5.41	5.61	5.57	5.00
Generative AI (GPT 4.0)								
No adaptation	171	0.0%	1.9%	5.43	5.08	5.15	5.38	6.31
Slides	47	7.8%	40.4%	5.86	5.29	5.68	5.68	3.47
Slides and Transcripts	44	0.4%	19.2%	6.06	5.33	5.75	5.60	3.83

The comparison of different systems highlights noteworthy differences in word count, requests, and instances of no response. The knowledge-based system has a word count of 27, while the generative AI systems provide significantly longer responses. GPT 4.0 adaptation results in lower word counts, although still longer than the knowledge-based systems. GPT 3.5 adaptation generally does not significantly impact word counts. GPT systems typically provide answers without requesting clarification from the user. In cases of confusion, knowledge-based systems may offer topic suggestions to clarify the user's query. The GPT 4.0 system seems to refuse more queries than the GPT 3.5 system in the case of adaptation.

7.2.1. Quality

Comparing the quality in Table 2 across different pedagogical conversational agents reveals distinct performance characteristics. The knowledge-based system achieves an impressive quality score of 6.54, indicating its effectiveness in delivering high-quality responses. Among the generative-AI-based systems, the GPT 4.0 system adapted to slides and transcripts scores 6.06, the highest, and it is marginally significantly worse ($p = 0.042$) than the knowledge-based system. The other generative-AI-based systems, which are tailored to the lecture content, consistently score between 5.86 and 5.89, which is significantly lower ($p < 0.01$) than the knowledge-based system. In contrast, generative-AI-based systems without content adaptation achieve significantly lower scores ($p < 0.001$) of 5.42 to 5.43, respectively.

This study demonstrates that customized systems, which are tailored to the specific content, outperform nonadapted systems in terms of quality. This emphasizes the benefits of narrower contexts and the proper utilization of appropriate data.

The nature of a lecture is significantly influenced by the lecture's scope. While the knowledge-based system is trained on sparser data exclusively from the lecture content, generative language models have knowledge of diverse topics, and thus, their performance may be hindered by the differences in definitions across scientific fields. In terms of quality, systems designed for specialized domains tend to perform better.

7.2.2. Quantity

There were no significant differences observed in the quantity metric in Table 2, which ranged from 5.08 to 5.40 among various pedagogical conversational agents. This is interesting because the word count greatly differs among the systems, and the quantity score is quite homogeneous. Each system tries to convey its answer, and some may explain it in longer or more detailed responses. However, these variations in word count do not appear to affect the overall quantity metric.

7.2.3. Relation

The analysis of the relation metric shown in Table 2 indicates that the knowledge-based system performs the best, with an average score of 5.83. The adapted GPT 4.0 models follow closely behind. Specifically, the GPT 4.0 model with access only to slides achieves a score of 5.75, while the adapted GPT 4.0 with both slides and transcripts obtains a mean score of 5.68. Overall, the generative AI models exhibit nominal differences in mean scores compared with the knowledge-based system. However, the generative AI models that lack adaptation scored significantly lower than the knowledge-based system (GPT 3.5 $p < 0.05$ and GPT 4.0 $p < 0.05$). Notably, there was also a significant difference in scores between GPT 3.5 without adaptation and GPT 4.0 adapted to both slides and transcriptions ($p < 0.05$).

These findings suggest that although a knowledge-based system may have a restricted depth of knowledge, it performs well in maintaining relevance to the topic at hand. Conversely, the adapted GPT 4.0 models showcase the ability to adapt and remain topical, with minimal differences in scores compared with the knowledge-based system.

It is worth noting that the nonadapted GPT 3.5 model produced the lowest scores. The enhancement of GPT 4.0 systems may be credited to the substantial amount of training data that have made the model less inclined to produce irrelevant information and improved its coherence.

7.2.4. Manner

According to Table 2, the scores for manner are similar, ranging from 5.20 to 5.68 across the different systems. This indicates that there are no significant differences in the clarity of the answers generated by each system.

Nevertheless, upon comparing the knowledge-based system with the generative-AI-based systems, there is a slight trend showing that the knowledge-based system's responses appeared to be less clear on average. Although this observed discrepancy is not statistically significant ($p > 0.05$), it hints at the possibility that the knowledge-based system may possess comparatively less eloquence, potentially owing to its relatively inflexible and organized nature.

7.2.5. Trust

When comparing the different chatbot systems in terms of trust in Table 2, we see that the GPT 4.0 system was perceived as the most trustworthy, with an average score of 6.31, significantly higher ($p < 0.001$) than all other systems except GPT 3.5. The knowledge-based system and the GPT 4.0 system, adapted to lecture slides, were rated as the least trustworthy, with mean scores of 3.59 and 3.47, respectively, significantly lower ($p < 0.001$) than all other systems, except the GPT 4.0 system adapted to lecture slides and transcripts, which achieved a mean score of only 3.83. These differences in scores result in the communication style of the chatbot. Using clear and concise explanations and avoiding uncommon expressions or technical vocabulary can enhance trustworthiness.

To see if the number of replies that belong to the classes 'request' or 'no answer' influences trust, the scores must be investigated separately. From Table 3, it is apparent that the different types of replies result in different trust scores. On average, answers are rated as more trustworthy than requests. No answer seems to be particularly untrustworthy. This is somewhat surprising, as one could think that honestly answering that one cannot answer the question could also increase trustworthiness. To investigate this further, we conducted a follow-up study on trust, collecting 150 ratings from 15 participants through an online questionnaire. The follow-up study confirmed that request answers and unanswered questions tend to decrease trust in a system. The average score for answers was 5.36, 3.40 for requests, and 3.34 for no answer. It is noteworthy that the scores for the request answers varied widely, indicating that participants were unsure how to evaluate the quality of a system based on a request answer.

As the generative systems did not often provide request answers or no answers, it is difficult to make a proper comparison here. However, there was a tendency for the

requests from the generative systems to be trusted to a higher degree compared with the knowledge-based system. Why the adapted GPT 4.0 systems showed a decrease in performance in giving no answer in contrast to GPT 3.5 could not be justified.

Table 3. Results on trust for the different systems according to the type of answer given. Numbers are not provided where the number of answers within the class were lower than five.

	Trust		
	Answer	**Request**	**No Answer**
Knowledge-Based	3.62	3.58	2.00
Generative AI (GPT 3.5)			
No adaptation	5.98	–	–
Slides	5.60	7.00	3.08
Slides and Transcripts	5.50	–	2.97
Generative AI (GPT 4.0)			
No adaptation	6.34	–	–
Slides	5.53	7.00	1.35
Slides and Transcripts	5.59	–	1.44

To establish and maintain trust, chatbots should demonstrate a deep understanding of the topic, draw information from reliable sources, and provide verifiable references for their responses. But, also, the competence of the user influences how trustfulness is rated. For instance, [68,69] have shown that novices tend to rely more on surface features like text length or image quality, while experts focus more on factual accuracy. In our study, we found a medium positive correlation (Pearson correlation of 0.46) between trust and the number of words in a chatbot's answer, supporting the former findings.

7.3. Reproducibility

Reproducibility is a crucial criterion for evaluating educational chatbots. As the chatbot industry rapidly evolves, validating the reproducibility of results becomes increasingly important. Reproducibility guarantees that the chatbot's responses remain reliable and consistent over time and with different users, establishing a sense of dependability and reliability. Educational institutions and users must have confidence in the chatbot's ability to consistently provide accurate and traceable information.

To compare the reproducibility of the given answers, we posed the same question two times and calculated the text similarity using sentence transformers from Hugging Face [70]. The cosine scores of the embeddings of the responses were utilized to measure their similarity. The results are given in Table 4.

The knowledge-based responses are consistent due to the deterministic nature of the DIET-classifier [63], which does not employ seeds. This is advantageous, as identical inquiries yield identical outcomes. Additionally, the embeddings that represent system information remain unaltered, and new embeddings are also generated deterministically. While asking different questions can lead to varying results, we found that minor modifications in a query, such as correcting typos or addressing grammar errors, typically have minimal impact on search outcomes.

For answers generated by ChatGPT without additional data, the similarity score is 0.88. However, when using ChatGPT adapted to slides with or without transcripts, the score increases to 0.91. These high semantic similarity scores indicate that ChatGPT-generated results have a small variation in their answers in response to the same input prompt. However, as investigated by [71], if prompted to behave differently, e.g., like a layperson instead of an expert, the results have a stronger divergence.

Table 4. Results on reproducibility using cosine scores of the embeddings.

	Reproducability
Knowledge-Based	1.0
Generative AI (GPT 4.0)	
No adaptation	0.88
Slides	0.91
Slides and Transcripts	0.91

7.4. Data Preparation

Data preparation is a crucial phase in any machine learning or AI project, including chatbot development. It involves the process of collecting, cleaning, and transforming data to make it suitable for analysis and model training, adaptation, or fine-tuning. To investigate how the input structure needs to be, we compared, in Table 5, semiannotated slides vs. unlabeled PDFs containing the same content. Although the evaluation of Grice's cooperative principles seems more advantageous for the PDF system, it is apparent that a substantial amount of information could not be extracted from the PDFs. Specifically, 63.6% of the questions were unanswered, as opposed to 25.4%.

Table 5. Results on the generative system using an unlabeled PDF version and a semilabeled PowerPoint version of the of the slides.

	No Answer	Quality	Quantity	Relation	Manner
Unlabeled PDF	63.6%	6.67	6.04	6.04	6.50
Semilabeled PowerPoint	25.4%	6.14	5.63	5.73	6.35

7.5. Observations

In the following, we provide our observations while using the system to provide valuable insights into the strengths and weaknesses of the chatbot systems beyond what can be measured, highlighting areas for improvement and optimization to enhance their overall performance and user experience.

- **Spelling Errors:** Dealing with spelling errors can be challenging for generative AI. ChatGPT attempted to provide answers even with misspelled terms, which may lead to incorrect or irrelevant responses and hallucinations. For instance, the question, which contains typos, 'Was ist Skrupeln?' (meant to be 'scribbeln', Engl. 'what is the meaning of to scribble?') provides a made-up answer in the case of generative AI systems. The knowledge-based system correctly mapped the word to the right meaning and provided a truthful description.
- **Alternative Meanings:** Generative AI systems with no adaptation sometimes choose the wrong definition, leading to inaccurate responses. Knowledge-based systems and systems with adaptation are better at selecting the correct definition.
- **Chatting:** The generative AI chatbot likes to add generic phrases like, 'Of course, I would be happy to explain some criteria', extending the number of words of the given answer without providing additional content.
- **Accuracy of Definitions:** The generative AI chatbot without adaptation delivers highly accurate results with small variations for definitions that are common; however, it struggles if topics are very specific. The knowledge-based system always provides the same definition, ensuring precise responses to user queries.

8. Discussion and Limitations

The results of this study indicate that systems solely based on pretrained large language models cannot provide the same scores in terms of quality, while other aspects are already satisfied, suggesting that other criteria may not be as important as quality when

deciding on an appropriate system. Generative AI systems often provided factually correct definitions that were not necessarily the desired or contextually appropriate ones, emphasizing the importance of embedding context and expert ground truth in these systems. Interestingly, despite the higher quality and relevance, the knowledge-based system was rated lower in terms of trust by the evaluators. This suggests that trust is influenced by various factors and cannot be solely determined by the quality of the information provided. It is important to note that generative-AI-based systems will always try to answer a query, even if it is out of the domain or even if a word does not exist. While this behavior leads to wrong answers, it nevertheless improves trust. Systems that frequently respond with 'I don't know' or request information from the users are generally trusted less.

Current generative AI systems are not inherently proactive. They primarily focus on generating text based on statistical patterns in the training data without considering the learner's specific needs, progress, or need for guidance. These limitations can hinder their effectiveness as a tutor. Knowledge-based systems allow for more control over the dialog flow, and strategies that go beyond a simple question–answer-based interaction are possible.

The limitations of our study might be its small number of evaluators and reproducibility. The lack of reproducibility is caused by the proprietary nature of the generative AI systems and the frequent updates to the models. In [72], it was stated that newer versions of ChatGPT perform worse in certain tasks when compared with older versions. It is difficult to evaluate systems that can produce practically infinite variations, and changes in prompting can lead to different answers and behavior. Furthermore, since our analysis was based on a single lecture, we cannot ensure that these results generalize to other lectures or different topics.

9. Conclusions and Future Work

We compared different pedagogical conversational agents and found that the information provided plays a significant role in how truthful the system can answer questions. This highlights the importance of context in designing effective conversational AI systems and the need to consider domain data. Nevertheless, plain generative AI systems perform reasonably well on general questions, and they can already be used as a sparing partner to support learning.

As for future work, there are several avenues to explore. One potential direction is to investigate how to guide the user with the system. For instance, ref. [46] showed how tasks can be contextualized by prompting such that it is possible to easily adapt the tasks to students' interests. Another key aspect will be on integrating different didactical strategies. While knowledge-based systems can provide verified answers, it is an open point for generative AI systems to verify answers and link to references. While domain-limited generative AI systems already achieve a better balance between good verbosity and truthfulness, hybrid systems that combine knowledge-based with generative AI approaches might be able to overcome some of the current limitations. For instance, a knowledge-based system could prompt a generative AI to use a specific teaching strategy that considers the learner's background, learning goals, and progress [19].

Furthermore, the integration of multimodal models and multilingual models could improve the overall quality of the system, help address issues of inclusion, and overcome language barriers. For instance, chatbots offer opportunities from which learners with visual or auditory impairment might benefit [73]. People with visual impairment could further beenfit from the integration of vision models or even recently introduced vision–language models (e.g., [74,75]) that are capable of translating images to text descriptions.

To sum up, pedagogical conversational agents have the potential to revolutionize education in various ways.

Author Contributions: Conceptualization, M.W.; methodology, M.W.; software, M.W., M.B.S. and A.R.; validation, M.W. and A.R.; formal analysis, K.A.; data curation, A.R.; writing—original draft

preparation, all authors; writing—review and editing, all authors; visualization, M.W.; supervision, M.W. All authors have read and agreed to the published version of the manuscript.

Funding: Mehrnoush Shirzad and Andreas Reich are supported by Stiftung Innovation in der Hochschullehre (project number FBM2020-EA-1670-01800). Katharina Anderer is supported by the Ph.D. scholarship in the Cooperative Doctoral Program Accessibility through AI-based Assistive Technology (kooperatives Promotionskolleg Barrierefreiheit durch KI-Basierte Assistive Technologien (KATE)) provided by the Ministry of Science, Research and Arts Baden-Württemberg (MWK).

Data Availability Statement: The research data is available here: https://github.com/colorfulbit/DataGriceCooperativePrinciples.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Intelligent. New Survey Finds Students Are Replacing Human Tutors with ChatGPT. 2023. Available online: https://www.intelligent.com/new-survey-finds-students-are-replacing-human-tutors-with-chatgpt/ (accessed on 15 December 2023).
2. Teaching with AI. 2023. Available online: https://openai.com/blog/teaching-with-ai (accessed on 4 September 2023).
3. OpenAI. Introducing ChatGPT. 2022. Available online: https://openai.com/blog/chatgpt (accessed on 20 July 2023).
4. Wölfel, M. Towards the automatic generation of pedagogical conversational agents from lecture slides. In Proceedings of the Multimedia Technology and Enhanced Learning: Third EAI International Conference, ICMTEL 2021, Virtual Event, 8–9 April 2021; Proceedings, Part II 3; Springer: Berlin/Heidelberg, Germany, 2021; pp. 216–229.
5. Hoang, T.N.; Reich, A.; Wölfel, M. The First Impression Counts! The Importance of Onboarding for Educational Chatbots. In *20. Fachtagung Bildungstechnologien (DELFI)*; Gesellschaft für Informatik e.V.: Bonn, Germany, 2022.
6. Ramesh, K.; Ravishankaran, S.; Joshi, A.; Chandrasekaran, K. A survey of design techniques for conversational agents. In Proceedings of the International Conference on Information, Communication and Computing Technology, New Delhi, India, 13 May 2017; Springer: Singapore, 2017; pp. 336–350.
7. Adamopoulou, E.; Moussiades, L. An overview of chatbot technology. In Proceedings of the IFIP International Conference on Artificial Intelligence Applications and Innovations, Neos Marmaras, Greece, 5–7 June 2020; Springer: Cham, Switzerland, 2020; pp. 373–383.
8. Gupta, A.; Zhang, P.; Lalwani, G.; Diab, M. Context-aware self-attentive natural language understanding for task-oriented chatbots. In Proceedings of the EMNLP 2019, Hong Kong, China, 3–7 November 2019.
9. Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I. Language models are unsupervised multitask learners. *OpenAI Blog* **2019**, *1*, 9.
10. Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 1877–1901.
11. OpenAI. GPT-4 Technical Report. *arXiv* **2023**, arXiv:2303.08774.
12. Hu, Z.; Lan, Y.; Wang, L.; Xu, W.; Lim, E.P.; Lee, R.K.W.; Bing, L.; Xu, X.; Poria, S. LLM-Adapters: An Adapter Family for Parameter-Efficient Fine-Tuning of Large Language Models. *arXiv* **2023**, arXiv:2304.01933.
13. Alkaissi, H.; McFarlane, S.I. Artificial hallucinations in ChatGPT: Implications in scientific writing. *Cureus* **2023**, *15*, e35179. [CrossRef] [PubMed]
14. Yager, K.G. Domain-specific chatbots for science using embeddings. *Digit. Discov.* **2023**, *2*, 1850–1861. [CrossRef]
15. Feldman, P.; Foulds, J.R.; Pan, S. Trapping LLM Hallucinations Using Tagged Context Prompts. *arXiv* **2023**, arXiv:2306.06085.
16. Hu, E.J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W. Lora: Low-rank adaptation of large language models. *arXiv* **2021**, arXiv:2106.09685.
17. Gao, T.; Fisch, A.; Chen, D. Making pre-trained language models better few-shot learners. *arXiv* **2020**, arXiv:2012.15723.
18. Mangrulkar, S.; Gugger, S.; Debut, L.; Belkada, Y.; Paul, S. PEFT: State-of-the-Art Parameter-Efficient Fine-Tuning Methods. 2022. Available online: https://github.com/huggingface/peft (accessed on15 December 2023).
19. Wang, L.; Sachan, M.; Zeng, X.; Wong, K.F. Strategize Before Teaching: A Conversational Tutoring System with Pedagogy Self-Distillation. *arXiv* **2023**, arXiv:2302.13496.
20. Fügen, C.; Wölfel, M.; McDonough, J.W.; Ikbal, S.; Kraft, F.; Laskowski, K.; Ostendorf, M.; Stüker, S.; Kumatani, K. Advances in lecture recognition: The ISL RT-06s evaluation system. In Proceedings of the Interspeech, Pittsburgh, PA, USA, 17–21 September 2006.
21. Wölfel, M. *Robust Automatic Transcription of Lectures*; KIT Scientific Publishing: Karlsruhe, Germany 2009.
22. Kolss, M.; Wölfel, M.; Kraft, F.; Niehues, J.; Paulik, M.; Waibel, A. Simultaneous german-english lecture translation. In Proceedings of the International Workshop on Spoken Language Translation (IWSLT), Waikiki, HI, USA, 20–21 October 2008.
23. Ashok, M.; Ramasamy, K.; Snehitha, G.; Keerthi, S. A systematic survey of cognitive chatbots in personalized learning framework. In Proceedings of the 2021 Sixth International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET), Chennai, India, 25–27 March 2021; pp. 241–245.

24. Dibitonto, M.; Leszczynska, K.; Tazzi, F.; Medaglia, C.M. Chatbot in a campus environment: Design of LiSA, a virtual assistant to help students in their university life. In Proceedings of the Human-Computer Interaction. Interaction Technologies: 20th International Conference, HCI International 2018, Las Vegas, NV, USA, 15–20 July 2018; Proceedings, Part III 20; Springer: Berlin/Heidelberg, Germany, 2018; pp. 103–116.
25. Khalil, M.; Rambech, M. Eduino: A telegram learning-based platform and chatbot in higher education. In *Proceedings of the International Conference on Human-Computer Interaction*; Springer: Berlin/Heidelberg, Germany, 2022; pp. 188–204.
26. Hien, H.T.; Cuong, P.N.; Nam, L.N.H.; Nhung, H.L.T.K.; Thang, L.D. Intelligent assistants in higher-education environments: The FIT-EBot, a chatbot for administrative and learning support. In Proceedings of the 9th International Symposium on Information and Communication Technology, Danang City, Viet Nam, 6–7 December 2018; pp. 69–76.
27. Gajra, V.; Lakdawala, K.; Bhanushali, R.; Patil, S. Automating student management system using ChatBot and RPA technology. In Proceedings of the 3rd International Conference on Advances in Science & Technology (ICAST), Bahir Dar, Ethiopia, 2–4 October 2020.
28. Chopra, S.; Gianforte, R.; Sholar, J. Meet percy: The CS 221 teaching assistant chatbot. *ACM Trans. Graph.* **2016**, *1*, 1–8.
29. Galko, L.; Porubän, J.; Senko, J. Improving the user experience of electronic university enrollment. In Proceedings of the 2018 16th International Conference on Emerging eLearning Technologies and Applications (ICETA), Stary Smokovec, Slovakia, 15–16 November 2018; pp. 179–184.
30. Graesser, A.C. Conversations with AutoTutor help students learn. *Int. J. Artif. Intell. Educ.* **2016**, *26*, 124–132. [CrossRef]
31. Sonkar, S.; Liu, L.; Mallick, D.B.; Baraniuk, R.G. CLASS Meet SPOCK: An Education Tutoring Chatbot based on Learning Science Principles. *arXiv* **2023**, arXiv:2305.13272.
32. Hew, K.F.; Huang, W.; Du, J.; Jia, C. Using chatbots to support student goal setting and social presence in fully online activities: Learner engagement and perceptions. *J. Comput. High. Educ.* **2023**, *35*, 40–68. [CrossRef] [PubMed]
33. Pérez-Marín, D. A Review of the Practical Applications of Pedagogic Conversational Agents to Be Used in School and University Classrooms. *Digital* **2021**, *1*, 18–33. [CrossRef]
34. Abbas, N.; Whitfield, J.; Atwell, E.; Bowman, H.; Pickard, T.; Walker, A. Online chat and chatbots to enhance mature student engagement in higher education. *Int. J. Lifelong Educ.* **2022**, *41*, 308–326. [CrossRef]
35. D'mello, S.; Graesser, A. AutoTutor and affective AutoTutor: Learning by talking with cognitively and emotionally intelligent computers that talk back. *ACM Trans. Interact. Intell. Syst. (TiiS)* **2013**, *2*, 1–39. [CrossRef]
36. Samtani, P.; Johnson, W.L.; Valente, A. Approaches to Modeling Conversational Agents in the Tactical Language and Culture Training System (TLCTS). In Proceedings of the AAAI Fall Symposium: Adaptive Agents in Cultural Contexts, Arlington, VA, USA, 7–9 November 2008; p. 53.
37. Gabajiwala, E.; Mehta, P.; Singh, R.; Koshy, R. Quiz maker: Automatic quiz generation from text using NLP. In *Futuristic Trends in Networks and Computing Technologies: Select Proceedings of Fourth International Conference on FTNCT 2021*; Springer: Singapore, 2022; pp. 523–533.
38. Dijkstra, R.; Genç, Z.; Kayal, S.; Kamps, J. Reading Comprehension Quiz Generation using Generative Pre-trained Transformers. 2022. Available online: https://e.humanities.uva.nl/publications/2022/dijk_read22.pdf (accessed on 15 December 2023).
39. Massaro, D.W.; Ouni, S.; Cohen, M.M.; Clark, R. A multilingual embodied conversational agent. In Proceedings of the 38th Annual Hawaii International Conference on System Sciences, Big Island, HI, USA, 3–6 January 2005; pp. 296b–296b.
40. Shen, J.T.; Yamashita, M.; Prihar, E.; Heffernan, N.; Wu, X.; Graff, B.; Lee, D. Mathbert: A pre-trained language model for general NLP tasks in mathematics education. *arXiv* **2021**, arXiv:2106.07340.
41. Liang, Z.; Yu, W.; Rajpurohit, T.; Clark, P.; Zhang, X.; Kaylan, A. Let GPT be a Math Tutor: Teaching Math Word Problem Solvers with Customized Exercise Generation. *arXiv* **2023**, arXiv:2305.14386.
42. Harrington, S.A. The Ultimate Study Partner: Using a Custom Chatbot to Optimize Student Studying During Law School. 2023. Available online: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4457287 (accessed on 15 December 2023).
43. Kasneci, E.; Seßler, K.; Küchemann, S.; Bannert, M.; Dementieva, D.; Fischer, F.; Gasser, U.; Groh, G.; Günnemann, S.; Hüllermeier, E.; et al. ChatGPT for good? On opportunities and challenges of large language models for education. *Learn. Individ. Differ.* **2023**, *103*, 102274. [CrossRef]
44. Winkler, R.; Söllner, M. Unleashing the potential of chatbots in education: A state-of-the-art analysis. In *Academy of Management Proceedings*; Academy of Management: Briarcliff Manor, NY, USA, 2018.
45. Adiguzel, T.; Kaya, M.H.; Cansu, F.K. Revolutionizing education with AI: Exploring the transformative potential of ChatGPT. *Contemp. Educ. Technol.* **2023**, *15*, ep429. [CrossRef] [PubMed]
46. Yadav, G.; Tseng, Y.J.; Ni, X. Contextualizing Problems to Student Interests at Scale in Intelligent Tutoring System Using Large Language Models. *arXiv* **2023**, arXiv:2306.00190.
47. Macina, J.; Daheim, N.; Chowdhury, S.P.; Sinha, T.; Kapur, M.; Gurevych, I.; Sachan, M. MathDial: A Dialogue Tutoring Dataset with Rich Pedagogical Properties Grounded in Math Reasoning Problems. *arXiv* **2023**, arXiv:2305.14536.
48. Wollny, S.; Schneider, J.; Di Mitri, D.; Weidlich, J.; Rittberger, M.; Drachsler, H. Are we there yet?-a systematic literature review on chatbots in education. *Front. Artif. Intell.* **2021**, *4*, 654924. [CrossRef] [PubMed]
49. Kooli, C. Chatbots in education and research: A critical examination of ethical implications and solutions. *Sustainability* **2023**, *15*, 5614. [CrossRef]
50. Tlili, A.; Shehata, B.; Adarkwah, M.A.; Bozkurt, A.; Hickey, D.T.; Huang, R.; Agyemang, B. What if the devil is my guardian angel: ChatGPT as a case study of using chatbots in education. *Smart Learn. Environ.* **2023**, *10*, 15. [CrossRef]

51. Bocklisch, T.; Faulkner, J.; Pawlowski, N.; Nichol, A. Rasa: Open source language understanding and dialogue management. *arXiv* **2017**, arXiv:1712.05181.
52. Zhao, W.X.; Zhou, K.; Li, J.; Tang, T.; Wang, X.; Hou, Y.; Min, Y.; Zhang, B.; Zhang, J.; Dong, Z.; et al. A survey of large language models. *arXiv* **2023**, arXiv:2303.18223.
53. Zhou, K.; Qiu, H.; Zhang. Chat-gpt: Potential, prospects, and limitations. *Front. Inf. Technol. Electron. Eng.* **2023**, 1–6. . [CrossRef]
54. Ni, Y.; Pandelea, X.; Cambria. Recent advances in deep learning based dialogue systems:a systematic survey. *Artif. Intell. Rev.* **2023**, *56*, 3055–3155. [CrossRef]
55. Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; Liu, P.J. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.* **2020**, *21*, 5485–5551.
56. Kaplan, M.; Henighan, B.; Chess, C.; Gray, R.; Wu, A. Scaling laws for Artifical hallucinationsal language models. *J. Mach. Learn. Res.* **2020**, *64*, 1–67.
57. Atapattu, T.; Falkner, K.; Falkner, N. A comprehensive text analysis of lecture slides to generate concept maps. *Comput. Educ.* **2017**, *115*, 96–113. [CrossRef]
58. Radford, A.; Kim, J.W.; Xu, T.; Brockman, G.; McLeavey, C.; Sutskever, I. Robust speech recognition via large-scale weak supervision. In Proceedings of the International Conference on Machine Learning, PMLR, Honolulu, HI, USA, 23–29 July 2023; pp. 28492–28518.
59. Tom, B. Rasa AI: Building Clever Chatbots. 2017. Available online: https://www.slideshare.net/TomBocklisch/rasa-ai-building-clever-chatbots (accessed on 19 September 2023).
60. RASA. Open Source Conversational AI | Rasa. 2023. Available online: https://rasa.community/ (accessed on 23 August 2023).
61. Kenton, J.D.M.W.C.; Toutanova, L.K. Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the naacL-HLT, Minneapolis, MN, USA, 3–5 June 2019; Volume 1, p. 2.
62. Apache. Apache Solr. 2023. Available online: https://solr.apache.org/ (accessed on 21 July 2023).
63. Bunk, T.; Varshneya, D.; Vlasov, V.; Nichol, A. Diet: Lightweight language understanding for dialogue systems. *arXiv* **2020**, arXiv:2004.09936.
64. Ekin. Prompt Engineering For ChatGPT: A Quick Guide To Techniques, Tips, And Best Practices. *TechRxiv* **2023**. [CrossRef]
65. Chase. LangChain. 2022. Available online: https://github.com/hwchase17/langchain (accessed on 20 July 2023).
66. Grice, H.P. Logic and conversation. In *Speech Acts*; Brill: Leiden, The Netherlands, 1975; pp. 41–58.
67. VMware. RabbitMQ: Easy to Use, Flexible Messaging and Streaming—RabbitMQ. 2023. Available online: https://www.rabbitmq.com/ (accessed on 25 July 2023).
68. Lucassen, T.; Muilwijk, R.; Noordzij, M.L.; Schraagen, J.M. Topic familiarity and information skills in online credibility evaluation. *J. Am. Soc. Inf. Sci. Technol.* **2013**, *64*, 254–264. [CrossRef]
69. Golbeck, J.; Fleischmann, K.R. Trust in social Q&A: The impact of text and photo cues of expertise. *Am. Soc. Inf. Sci. Technol.* **2010**, *47*, 1–10.
70. Wang, L.; Yang, N.; Huang, X.; Jiao, B.; Yang, L.; Jiang, D.; Majumder, R.; Wei, F. Text Embeddings by Weakly-Supervised Contrastive Pre-training. *arXiv* **2022**, arXiv:2212.03533.
71. Zylowski, T.; Wölfel, M. An NLP Analysis of ChatGPT's Personality Simulation Capabilities and Implications for Human-centric Explainable AI Interfaces. In Proceedings of the 6rd International Conference on Natural Language and Speech Processing, Virtually, 16–17 December 2023.
72. Chen, L.; Zaharia, M.; Zou, J. How is ChatGPT's behavior changing over time? *arXiv* **2023**, arXiv:2307.09009.
73. Gupta, S.; Chen, Y. Supporting inclusive learning using chatbots? A chatbot-led interview study. *J. Inf. Syst. Educ.* **2022**, *33*, 98–108.
74. Zhu, D.; Chen, J.; Shen, X.; Li, X.; Elhoseiny, M. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv* **2023**, arXiv:2304.10592.
75. Dai, W.; Li, J.; Li, D.; Tiong, A.M.H.; Zhao, J.; Wang, W.; Li, B.; Fung, P.; Hoi, S. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. *arXiv* **2023**, arXiv:2305.06500.