This paper was presented at CSEDU 2026 (https://csedu.scitevents.org/?y=2026). Cite as:
Padó, U. and Pampel, B. (2026). TutorBots in Context: A Multi-Lens View on Educational Practice. In Proceedings of the 18th International Conference on Computer Supported Education - Volume 1, ISBN 978-989-758-833-4, ISSN 2184-5026, pages 280-288.
DOI: 10.5220/0014746800004021

TutorBots in Context: A Multi-Lens View on Educational Practice

Ulrike Padó[1][a] and Barbara Pampel[2][b]
[1]*Hochschule für Technik Stuttgart, Schellingstr. 24, 70174 Stuttgart, Germany*
[2]*University of Konstanz, Universitaetsstr. 10, 78464 Konstanz, Germany*
ulrike.pado@hft-stuttgart.de, barbara.pampel@uni-konstanz.de

Keywords: Retrieval-Augmented Generation, Large Language Models, Chatbots in Education.

Abstract: Tutorbots are an increasingly popular tool for enabling students to access individualized chat support for the content of their classes, wherever and whenever. To further facilitate the move from research prototypes to widespread adoption in Higher Education, we identify obstacles and derive recommendations for overcoming them by considering tutorbots through four Lenses (cf. Brookfield, 1995): Our autobiographic Lens, the literature Lens, the Lens of other educators and the Lens of students working with the tutorbot. Our core insights are that many of the educators' and students' initial concerns can be addressed effectively by utilizing Retrieval-Augmented Generation bots. We therefore recommend training educators how to configure and students how to use them, because we find that students profit from instruction on the bots' technical background and on prompting despite self-reported frequent use of chatbots.

1 INTRODUCTION

The advent of chatbots has sparked immense interest in their didactic use. They offer students the opportunity to ask questions at any time of day or night and receive a personalized answer. However, after initial enthusiasm regarding the potential of chatbots in education, limitations became apparent: A common concern is the tendency of chatbots to hallucinate, i.e., to generate answers that sound plausible, but are factually incorrect (Maynez et al., 2020). Also, chatbots can be lured to make undesired outputs through prompt injection (Yu et al., 2024).

Both issues are caused by the Large Language Models (LLMs) that underlie chatbots' text generation capabilities. LLMs learn to predict a likely sequence of next words given their prompts. In this process, they also acquire implicit knowledge of a broad range of subject areas (Petroni et al., 2019) – simply because some sequences of words are more probable than others given the training data. However, since LLMs do not explicitly encode knowledge, they are vulnerable to generating plausible, but incorrect or undesired output.

These concerns can be addressed by the Retrieval-Augmented Generation (RAG) technique (Lewis et al., 2020) which first selects relevant excerpts from a curated knowledge base and passes them on to the LLM along with the user's query and any other prompts (Fan et al., 2024). The LLM then bases its answer on the reliable excerpts and is prompted to refuse to answer in case there is no relevant information in the knowledge base. This reduces hallucinations and makes the chatbot more robust towards prompt injection, as it is likely to refuse prompts that do not refer to facts in the knowledge base.

A popular application of this technique in education is defining a tutorbot, that is, a specialized chatbot that answers students' questions based on the materials of their class. These bots have the advantage of reliably offering relevant information at the right level of detail. This technology is in principle ready for widespread use, but in the context of such broader adoption, the needs and concerns of the intended users become of utmost importance for the acceptance of the technology.

In this paper, our goal is to identify these needs and concerns and to derive actionable recommendations for overcoming them. To this end, we follow a multi-perspective strategy similar to Brookfield's Four Lenses (Brookfield, 1995) as an analytical framework to provide a structured way to examine tutorbots beyond a single viewpoint. From our own perspective, we derive our motivation in Section 2, and then move to sketching the technological possibilities according to the current literature (see Section 3). We

[a] https://orcid.org/0009-0000-0664-7487
[b] https://orcid.org/0000-0001-6492-0381

also consider the concerns of fellow educators about tutorbot use (Section 4), and finally look at both the concerns and experience of actual student users (Section 5 and 6). We integrate our results and derive actionable insights from all four Lenses in Section 7.

2 LENS 1: AUTOBIOGRAPHIC PERSPECTIVE

Exploring the possibilities of using AI chatbots in Higher Education, our earlier work showed how even educators with limited technical expertise can design and configure their own tutorbots (Pampel et al., 2025). The potential applications of these bots are manifold, as they can take on very different didactic roles through careful alignment of the system prompt with the educator's intentions (Lauber et al., 2026).

Building on these considerations, our initial experiences with Retrieval-Augmented Generation (RAG) were highly promising. For straightforward Q&A tasks on text-heavy knowledge bases, early prototypes behaved almost entirely as intended: answers were accurate and the systems reliably refused queries outside the defined scope, limiting the opportunities for prompt injection. This suggested that lectures and seminars could be supported using relatively simple setups, like they are offered by numerous providers, using a medium size standard LLM. However, as soon as the subject matter became more complex, particularly in mathematical or algorithmic domains, response accuracy declined sharply and could not be substantially improved by merely adapting the knowledge base or refining the system prompt.

Hence, we had to take into account more parameters of the chatbot's configuration. We had to carefully choose the LLM and the (creativity) temperature for the generation as well as the embedding model, especially when both knowledge bases and prompts combined content from multiple languages. Depending on the structure of the knowledge base, where information could be either spread out over multiple passages or most likely contained in one section, we had to adjust the parameters for the retrieval step, such as the size of the chunks the knowledge base is split into, their overlap and the number of chunks retrieved for response generation. We further discovered that limiting the buffer memory for the chat history could prevent the chatbot from being lured away from the intended context. Our tutorbot then responded as intended in 95% of the 328 recorded interactions, while only 2.5% of unintended reactions contained incorrect answers (Pampel et al., 2026).

Initial pilot tests in our own courses have also shown how much the acceptance and thus the intensity of use of a bot depends on whether it responds to queries as expected. Many students already use other chatbots regularly and only accept course-specific offers if they seem advantageous to them – for example, if they respond more reliably to questions about the subject area or relate to the explanatory approaches and technical language used in the courses.

Through collaborations with colleagues responsible for other teaching and learning settings, we gained additional insights: These exchanges showed that aligning the chatbot's behavior with the structure of the knowledge base and the intended didactic objectives through parameter settings needs these educators' in-depth familiarity with the teaching context.

To summarize the findings of this lens, our insights and experiences indicate a need for educators to be both capable and able to carefully configure their own RAG-bots, but that we need a more systematic investigation in order to be able to make well-founded recommendations for the dissemination of such concepts and systems.

3 LENS 2: LITERATURE PERSPECTIVE

In response to concerns about chatbot reliability in education, Retrieval Augmented Generation was a source of renewed optimism. Hence, an increasing number of course-specific RAG systems were developed and evaluated – see Swacha and Gracel (2025) and Li et al. (2025) for an overview.

Application Contexts. The majority of studies on the use and effects of RAG chatbots focus on students rather than on the teachers who deploy these systems, and are almost exclusively in the field of Higher Education (Swacha and Gracel, 2025).

A vast majority of the projects reported in the literature rely on commercial models, most commonly GPT-3.5 or different versions of GPT-4 (Swacha and Gracel, 2025; Li et al., 2025), which entails both ongoing costs and a dependence on external providers.

Tutorbots are used by students primarily for review, clarification of terms, and self-study support (Lang and Gürpinar, 2025; Sánchez-Vera, 2025). The amount of usage is correlated with students' prior subject knowledge (Lang and Gürpinar, 2025) and there are indications that students may lack appropriate prompting skills to make the most of their interactions with the chatbot (Sánchez-Vera, 2025).

Learning Outcomes. Regarding the effects, empirical findings remain mixed. Hakim et al. (2024) found that for mandatory engineering lab health-and-safety

training, RAG-chatbots significantly improved their students' learning performance compared to both a generic chatbot and traditional instruction. Similarly, Miladi et al. (2025) report that the availability of a RAG-chatbot improved knowledge retention among participants of a MOOC. Gains from chatbot use level off at high interaction levels instead of rising linearly, possibly indicating an effect of overreliance (Sánchez-Vera, 2025). Other studies did not find significant differences in learning gains (Thüs et al., 2024; Lademann et al., 2025), but positive effects on motivation, subjective understanding, cognitive load, higher self-efficacy and technology acceptance (Lademann et al., 2025; Hakim et al., 2024).

Accuracy
While RAG can drastically increase factual consistency between chatbot responses and course materials and lead to high content reliability (Sánchez-Vera, 2025), results may differ among systems and educational settings, which continues to raise questions about the answer quality in an educational context. Neumann et al. (2024) developed a chatbot that can be integrated into the learning management system Moodle, and while they found a high accuracy rate (88%) in their tests, they emphasize that educators should encourage students to critically assess and verify a chatbot's responses and underscore that a chatbot is a supplementary tool rather than a definitive source of knowledge. For the CS50-Duck system, Liu et al. (2024b) report similar performance numbers for questions on course content, with the caveat that questions about administrative issues are answered less reliably (77%). This is problematic in its own right, as students may be misled about requirements, for example.

Similar to our own experience (see 2), these findings show the need for careful benchmarking of each tutorbot. This should optimally be done by the educators employing it, given their unique ability to assess the pedagogical suitability, curricular alignment, and reliability of a chatbot's responses within their specific educational context. Despite this, the existing literature contains only a very small number of teacher-centered RAG systems. Most reported approaches focus on learner-oriented systems with fixed configurations, while explicit support for teachers, such as in the construction of the knowledge base, pedagogical role modeling, and system control, remains rare. HiTA (Liu et al., 2024a) is a RAG-based platform in the sense of a virtual teaching assistant which explicitly integrates teachers' expertise into the didactic process via curated materials and rules in system prompts. But also here, as well as in most other interfaces for the educators, additional parameters relevant for a RAG-chatbot's configuration, like the chunk size and overlap, the number of retrieved passages and even the temperature for the LLM and the choice of embedding model, remain inaccessible, despite their influence on context alignment and the likelihood of hallucinations (Huang et al., 2025).

To summarize the findings of this lens, the literature supports our autobiographic experiences that RAG-chatbots can be a valuable addition to educational settings, but that their successful integration needs careful configuration.

4 LENS 3: COLLEAGUES' PERSPECTIVE

We now investigate the perspective of colleagues, i.e., educators motivated to use tutorbots in their teaching, to better understand what may facilitate or hinder the actual adoption of such systems in practice. We collected real-life concerns regarding chatbots in teaching settings from educators at German universities. We wanted to hear primarily from educators who were already motivated to use tutorbots, in order to gauge the impediments to real-world tutorbot use that still exist in a "best case" scenario.

We offered two online-workshops for educators and academic staff on how to design specialized chatbots for their own classes in different fields of study. We therefore assume that the participants were representatives of the target user group. We elicited concerns and worries about the use of LLMs from this group, using an open question format in order to distort the participants' opinions as little as possible.

After a brief introduction into generative AI, we asked the participants about their concerns regarding the use of LLM-based chatbots in education in general. (Note that because of the workshops' topic the participants might have had chatbots in mind that are specialized to some extent.)

4.1 Method

The workshops were held online in a video-conference tool that offers a Q&A-feature, where the participants can (anonymously, if they want) post concerns themselves and also vote for someone else's statement, in case they agree on it as well. Responses were collected in German and are translated where they appear below.

Overall, we collected 77 textual responses to our open question, which could also receive supporting votes from other participants (the most supported response had 12 votes and was about about chatbots'

"lack of factual accuracy / hallucinations"). For our analysis, we counted the number of votes for each concern and – adding one for the author – weighted the concerns accordingly.

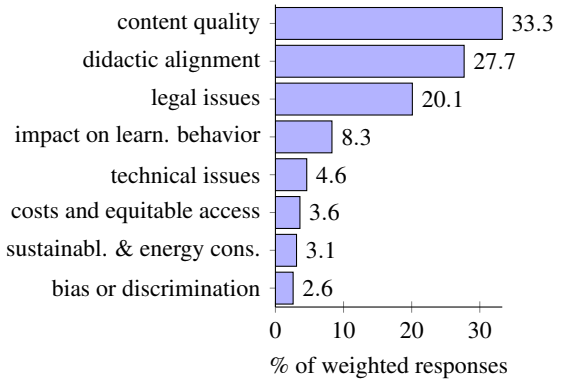


Figure 1: Percentage of weighted responses by educators for each category of potential issues. $N = 194$ is the sum of all weighted responses, where each statement is multiplied by the number of votes plus one for the author.

4.2 Results

The share of weighted responses is shown in Figure 1. By far the most concerns (33.0%) regard the content quality of the bot's answers. Many mention hallucinations, for example that "the chatbot presents seemingly coherent results that are factually incorrect, making it difficult for learners to distinguish them from valid explanations", or outdated information, such as concerns about "how quickly updated materials, for instance current exam templates or exercise sets, can be made available to the chatbot". But also the chatbot's behavior, which in an educational context reflects the didactic alignment, is mentioned several (27.7%) times, for example that "the level of knowledge of the students may be over- or underchallenged", or that "the system prompt should be designed in such a way that it cannot be misused and that learners have to work out the answers themselves".

Among the legal issues mentioned (20.1%) is compliance with the European Union's General Data Protection Regulation (GDPR), but also the European Union's AI Act, such as "am I, by operating the bot, considered an AI operator or a provider?", but also more general issues of legal responsibility, for example "legal certainty in cases where students perceive hallucinations as correct and use this incorrect information". Some concerns (8.3%) were more on a meta-level, regarding general changes in the students learning behavior, like relying too much on support systems like these chatbots, but also including the fear of loss of personal contact between educators and learners, as reflected in the statement that "interaction with motivated human tutors generates higher motivation among students than a bot".

To our surprise, only few participants fear technical issues (4.6%), costs and access (3.6%), sustainability and energy consumption (3.1%) or biased and discriminating behavior (2.6%) of the bots. This might have been influenced by the workshop's announced focus, where participants may have assumed equal accessibility of their own bots for all of their students, for example.

To summarize the findings of this lens, the educators' perspective centers on their desire to provide their students with the best possible tools (high content quality and good didactic alignment) and at the same time safeguard students' data privacy and their self-determination vis-á-vi AI systems (legal provisions).

5 LENS 4: STUDENTS' PRECONCEPTIONS

We next collect input from groups of students, who bring their own experiences, expectations and reservations into the teaching setting. We are interested in two aspects of their perspective: One, their preconceptions and concerns as chatbot users, and two, their actual usage patterns of a tutorbot provided for their classes (see Section 6).

5.1 Method

We distributed an anonymous on-line questionnaire to Computer Science students in their first or second year of study for a Bachelor's degree at a German University. Most of the students were enrolled in one of two iterations of an introductory AI class as part of their compulsory studies. We also contacted students in their first semester in an Introduction to Programming class. Participation in the study was voluntary.

In one AI class, running in the winter of 2024, 34 students participated, in the second AI class (summer of 2025), 47 students answered our questionnaire. The Programming class also ran in the winter of 2024, and 12 students answered. We asked the students in all three classes about their previous experience with chatbots and their current usage before covering any of these topics in class. Since the questions answered by all three groups were the same, we report the data for the total of 93 respondents together. In the summer AI class, we also handed out an additional questionnaire targeting students' preconceptions about AI

systems before the start of formal instruction. 49 students participated.

5.2 Results

Use of Commercial ChatBots. In all three classes, students clearly are experienced users of chatbots outside of class: 62% of students use chatbots at least once a day, 30% use them weekly and only 8% of students use chatbots monthly or not at all.

Students use the chatbots predominantly to research information (69.6%) and for coding support (41.9%), but also to help them revise for classes (25.8%). This underscores the potential usefulness of specialized, relevant and reliable chatbots in teaching. Text generation (and correction) is the least mentioned usage at 19.4% (multiple mentions were possible); at this stage in the degree, few written homework assignments are due.

Students are quite happy with the performance of the chatbots they use - 85% of students say that they are "satisfied for some uses" or "mostly satisfied" and none are fully dissatisfied. However, only 13% of students are fully satisfied by the currently available general-purpose chatbots.

Three frequent reasons for dissatisfaction appear in the data (ignoring reasons mentioned fewer than five times, multiple concerns could be named): Students most frequently report encountering hallucinations (39.7%) or feeling that their prompts are misinterpreted (21.5%). Some students also worry about the cost of a paid subscription (8.6%).

Understanding of AI. In the summer AI class, we also probed students' understanding of LLM-based AI systems before the start of formal instruction in addition to asking about students' LLM usage habits. We selected two salient questions from a questionnaire on AI concepts by Mayer (2024).

Students felt fairly confident of their general understanding of AI (see Fig. 2) and overwhelmingly see (LLM-based) AI systems as "a new kind (literally: generation) of search engine". This last statement is fully consistent with students' reported usage of chatbots to research information.

To summarize the findings of this lens so far, our results indicate that students are frequent users of commercial chatbots, which they use primarily to research information. Therefore, offering them didactically aligned tutorbots opens a natural way of learning for the students. However, the added value of the specialized tutorbot should be clearly described, as students already have their habitually used chatbots to fall back on. Students report that they are generally quite satisfied with commercial chatbots, but the main

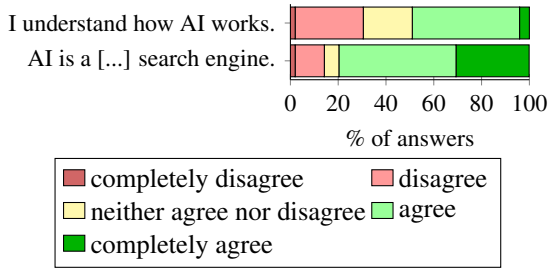


Figure 2: Student feedback on technical understanding of AI before instruction. $N = 49$.

concern they do have closely mirrors the educators' greatest concern, namely the reliability and quality of chatbot answers.

We also see a separate concern about receiving unexpected results that do not fit the intention of the prompt. This was not a salient point for the educators, but is consistent with students' clear intuition that LLM-based AI is "a new generation of search engine" and their matching self-reported usage of chatbots to retrieve information.

6 LENS 4: STUDENTS' USAGE PATTERNS

Our findings in Sec. 5 raise the question which interaction styles the students use with chatbots – do they formulate conversation-like natural language prompts as if speaking to a person, or do they use the keyword strategy usually employed with traditional search engines? We investigate this by analyzing which types of interactions we see in their encounters with the chatbot and their prompting strategies for information searching as well as the impact of instruction.

6.1 Method

We present results from the same classes as in Section 5. In all three classes, students were provided with a tutorbot, a chatbot specialized for the content of their class. They had access to the bot from its introduction in class until after the exams. All three bots were prompted to be a source of explanations of the class materials (but did not have meta information about the class, like test dates and test content). We specifically instructed the tutorbots not to solve exercises or create longer sections of code.

The tutorbots used the RAG paradigm with a knowledge base made up of the lecture slides for the courses (and, in the case of the AI classes, also the transcripts of lecture videos, which helped improve the retrieval results). We initially used

the commercial `OpenAI GPT-3.5 turbo` LLM, and later switched to the `llama-3.1-instruct` and `mistral-large-instruct` LLMs.

We logged students' interactions with the tutorbot (students used the tutorbot voluntarily and were able to opt out of analysis by marking their interactions accordingly, but no one did). In the winter AI class, we logged 520 interactions made up of a student prompt and bot answer; in the summer class, 374, and in the Programming class, 268.

We then annotated the student prompts in each prompt-answer pair by intention (e.g., exploring the tutorbot's abilities, requesting information, social interaction). From these labels, we identified four salient prompt topics: *explore* covers prompts that satisfy students' curiousity about the tutorbot's abilities, both as overt queries and as prompt injection attempts (to explore the bot's robustness). Interactions marked as *information* ask the bot to provide definitions and explanations, in line with its chosen didactic role. Given the students' self-reported expertise with chatbots, we also check whether the students demonstrate sophisticated prompting strategies by modifying the tutorbot's initial output through follow-up *meta prompts* (for example, requesting a shorter answer, an example or a re-phrasing in simpler words). Finally, we report the extent to which the students show social *interaction* with the tutorbot (for example greeting it, thanking it, or expressing frustration directed at it). This will give additional insight into the students' interaction strategies with the tutorbot. Together, these four topics cover 79% of all interactions with the tutorbot in the three classes.

6.2 Results

We begin by presenting analyses of the distribution of the four selected interaction types. We then zoom in on which prompting strategies students use naively, and finally look at the impact of instruction both in usage patterns and students' self-reported behavior changes.

Interaction Type Distribution

Fig. 3 shows the distribution of the four interaction types for each of the three courses. The general picture is similar across all classes: Students used approximately 40% of all prompts to ask for information, as appropriate for the didactic role of our tutorbot. They made few social interaction attempts and hardly used meta prompts.

Three interesting differences emerge between the classes: The Programming students were least interested in exploring or hacking the tutorbot and focused most on asking for information. In both AI classes, where the tutorbot was the topic of instruction, students explored or attempted to hack much more frequently, especially in the winter class. The summer AI class had most attempts at social interaction with the bot. This difference between the AI classes may be explained by the timing of instruction; in the summer, students had learned about human dialogue patterns and the technical background of the tutorbot before first using it, which may have shifted the focus of their exploration towards human dialogue patterns.

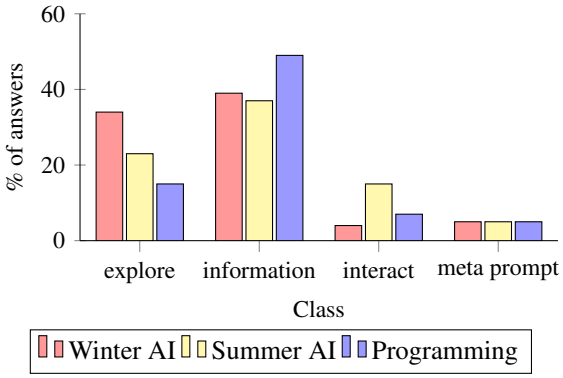


Figure 3: Distribution of student prompt types in three classes, in percent of prompts for that class.

Prior Prompting Strategies. We now look at students' prompting strategies before receiving instruction. We zoom in only on the information prompts, since these indicate interactions that are the center of our didactic interest.

In the winter AI class, we logged 133 information prompts in the weeks before instruction. In the summer class, the class structure was different and students received instruction before their first contact with the bot, so we have no relevant data from this class. However, we can use all 110 information prompts from the Programming class, where no explicit instruction was given.

We then analyzed the language of the prompts in these interactions and find that the majority of prompts is very general and vague: Students often use short sentences like "What is ...?" or even just search terms (as for a search-engine query). Fig. 4a shows a breakdown into five different prompting strategies within the information interaction type: Wh-questions (including instructions to "explain X"), pure keyword queries, and three more sophisticated questioning strategies: First, queries that give additional detail (e.g., "What is the role of neural networks in AI?", which is clearly more focused than asking "What are neural networks?"). Even more advanced are queries that ask about the differences between different concepts and, finally, queries that include a

meta prompt regarding the tutorbot's output, for example "How does Machine Learning work? Give a detailed answer."

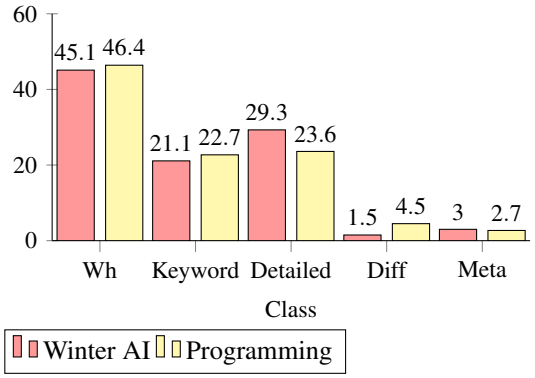
The strategies are distributed similarly across the two classes; terse wh-question and keyword prompts make up more than 60% of prompts in both classes. Among the more sophisticated prompting strategies, the detailed queries that give some additional context to the query are most common at 29% (AI) and 24% (Programming) respectively. The more advanced strategies that ask about the difference between two concepts or give meta-instructions about the level of detail or examples are much more infrequent and are both below 5%, in the data from both classes.

This high proportion of very general queries indicates that students have little understanding of the importance of providing detailed prompts; our finding above on the rare use of follow-up meta prompts (like "Ok. Give more detail.") shows that they also don't often use the opportunity to follow up on general prompts with later refining requests.

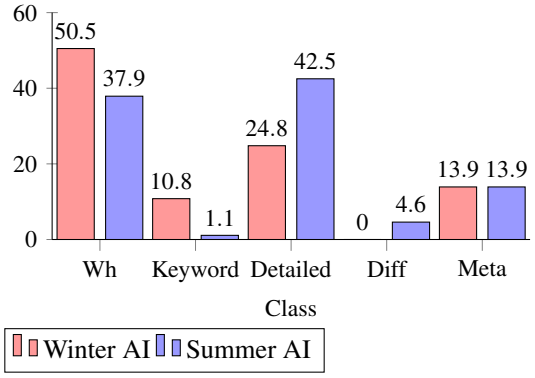
Impact of Instruction. To show the impact of instruction on students' prompting strategies, we now analyze the prompts logged after instruction (excluding the instruction sessions themselves, since the students were explicitly asked to experiment with prompting, which reduces the interpretability of the data). We have 93 information prompts in the winter AI class after instruction, and 88 prompts from the summer AI class. There was no instruction in the Programming class.

Fig. 4b shows the distribution of prompt strategies in the interactions falling under the information prompt types. Quite strikingly, the majority of prompts still falls either in the Wh-Question or the Detailed category, as before, but the number of pure keyword prompts has fallen by half in the winter AI course and this prompt strategy barely exists in the summer AI data. Also, we see a drastic rise in prompts that include meta instructions such as "include an example" or "explain like I'm five". Instruction has therefore helped students to shift away from keyword search strategies to more detailed prompts as the first level of sophistication, and even to prompts that explicitly manipulate the level of detail and complexity of the tutorbot output, at a much higher level of sophistication.

This shift is also mirrored in questionnaire data. In the Summer AI class, we asked students about the impact of instruction on their prompting behavior. At a relatively low completion rate of $N = 15$, 60% of students report that their prompting behavior has changed. All of these students state that they now add



(a) Before instruction on LLMs in two classes.



(b) After instruction on LLMs in two classes.

Figure 4: Distribution of student prompt strategies, in percent of the *information* prompts for that class. Top: Before instruction, bottom: after instruction.

more detail to their prompts and define more clearly what information they are looking for.

To summarize the Findings of this Lens. We saw both curiosity about the tutorbot (expressed by hacking attempts and explorative prompts), as well as a high number of bona fide information requests. This matches both our expectation given that the bot was a topic of instruction in two of the three courses and our didactic intention for the tutorbot as a source of information, as well as the students' self-reported conceptualization of the tutorbots as search engines. The low number of social interaction prompts also match this.

On the level of the prompting language, we found that the level of students' prompting sophistication was overall quite low before instruction about the technical background and functioning of the tutorbot. This pattern is again in line with their view of chatbots as search engines and is a likely explanation for the students' stated dissatisfaction with commercial chatbots at the level of prompt interpretation discussed in Section 5. Probably, the keyword prompting technique frequently returns chatbot answers that do not match the users' intentions.

We therefore identify an information need by the

students in this dimension. When addressing this information need, the impact of instruction is directly visible in the rising sophistication of prompting, which leads to more context for the tutorbot and therefore potentially reduces the amount of misinterpreted prompts.

7 DISCUSSION AND CONCLUSIONS

In order to smoothen the path to wider adoption of RAG-based tutorbots in Higher Education, we have investigated such bots from four perspectives, inspired by Brookfield's Four Lenses (Brookfield, 1995): We have described our motivation from our autobiographic view of tutorbots in education, sketched the available technology from the literature, asked educators about their concerns and expectations, and added an analysis of students' perspectives and prompting patterns across several courses.

Integrating these perspectives, we first and foremost identify important information needs on the side of both students and educators. Similar to our own experience, we see that both groups are concerned that tutorbots might hallucinate; educators also worry that they might not be well-aligned to their didactic purpose. However, from the autobiographical and literature lenses we draw evidence that this concern can be addressed by RAG-tutorbots:

The *content quality* can be improved by carefully configuring the tutorbots' parameters, like the size of the text chunks in the knowledge base, the number of chunks retrieved or the (creativity) temperature for the response generation.

Tutorbot alignment to a didactic role is in turn controlled by the system prompt, which assigns a persona and role to the tutorbot that fits the didactic goals of the class. A locally hosted bot with an open-source LLM cuts out any additional system prompts injected by hosting platforms and services.

The third major concern voiced by educators was questions about *legal issues*, which address both the requirements of the European data protection laws (GDPR) and of the EU AI Act (comparable legislation of course exists in other countries).

The data protection regulations govern the processing of personally identifiable data (e.g., names, matriculation numbers or addresses). Usually, it is not necessary to collect such data in order to establish a chatbot for instruction, and students can be advised not to enter any such personal information into the chat conversation. The AI Act's provisions require any users of university-provided AI systems – students and educators alike – to be trained to use them safely. (Additional measures become necessary in high-risk settings, for example when AI is used for assessment.)

Therefore, our **first recommendation** is to train educators on the technical background of LLMs, and instruct them how to define their own tutorbots. This qualifies them to verify and maintain high-quality tools for their students, and also to communicate this fact to their students, increasing their trust. We also argue that such training qualifies educators for the use of AI systems in the sense of the EU AI act, enables them to better safeguard students' private data and helps them ensure they use low-risk systems in the sense of the AI Act. Finally, this of course also serves to alleviate educators' concerns about the legal ramifications of offering chatbots in class.

Intriguingly, we have identified a second kind of information need in the student group: Although students self-reported as frequent users of chatbots, they are often not satisfied with the answers they receive. We find that they understand LLM-based AI models to be a new kind of search engine, and that they prompt accordingly, using very simple prompts and often only keywords. This lack of context in their prompting causes general and vague, or possibly unexpected chatbot answers. With RAG tutorbots, this is compounded because retrieval of relevant passages suffers. Note that we notice this in a "best case" scenario among Computer Science students who self-reported as experienced chatbot users.

Our **second recommendation** is therefore to carefully instruct students about the technical background of LLM-based AI systems and encourage them to experiment with prompting techniques on any kind of chatbot. Doing so changed students' prompts towards including more detail and directions for the chatbot. Such instruction also helps satisfy the requirements of the EU AI act for training AI users.

In sum, integrating the four different perspectives on tutorbots in the Higher Education classroom has enabled us to identify the biggest concerns about the use of tutorbots, and make recommendations on how to overcome them. Fortunately, none of the issues raised are insurmountable today; instead, they can already be addressed for the most part by careful technical implementation and training of both educators and students.

Future work therefore includes the development of training strategies for educators and students and their evaluation in real life, especially for cohorts that are not as technologically savvy as the Computer Science students considered here.

Funding

This work was supported by a bwDigiFellowship to both authors granted by Stifterverband and the Ministry of Science, Research and Arts of Baden-Württemberg, Germany. The authors gratefully acknowledge the LLM services granted by the KISSKI project of Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen.

REFERENCES

Brookfield, S. D. (1995). *Becoming a critically reflective teacher*. Jossey-Bass.

Fan, W., Ding, Y., Ning, L., Wang, S., Li, H., Yin, D., Chua, T.-S., and Li, Q. (2024). A Survey on RAG Meeting LLMs: Towards Retrieval-Augmented Large Language Models. In *Proceedings of the 30th ACM SIGKDD Conference*, KDD '24, page 6491–6501, New York, NY, USA. ACM.

Hakim, V. G. A., Paiman, N. A., and Rahman, M. H. S. (2024). Genie-on-demand: A custom AI chatbot for enhancing learning performance, self-efficacy, and technology acceptance in occupational health and safety for engineering education. *Computer Applications in Engineering Education*, 32(6):e22800.

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., and Liu, T. (2025). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Trans. Inf. Syst.*, 43(2).

Lademann, J., Henze, J., and Becker-Genschow, S. (2025). Augmenting learning environments using AI custom chatbots: Effects on learning performance, cognitive load, and affective variables. *Physical Review Physics Education Research*, 21(1):010147.

Lang, G. and Gürpinar, T. (2025). AI-powered learning support: A study of retrieval-augmented generation (RAG) chatbot effectiveness in an online course. *Information Systems Education Journal*, 23(2):4–13.

Lauber, A.-M., Martin, S., and Pampel, B. (2026). How Can I Help You? Didactic Alignment for RAG Chatbots in Education. In *The 17th ICETC (2025)*. to appear.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., tau Yih, W., Rocktäschel, T., Riedel, S., and Kiela, D. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Proceedings of NeurIPS*, Vancouver, Canada.

Li, Z., Wang, Z., Wang, W., Hung, K., Xie, H., and Wang, F. L. (2025). Retrieval-augmented generation for educational application: A systematic survey. *Computers and Education: Artificial Intelligence*, page 100417.

Liu, C., Hoang, L., Stolman, A., and Wu, B. (2024a). HiTa: A RAG-based educational platform that centers educators in the instructional loop. In *International Conference on Artificial Intelligence in Education*, pages 405–412. Springer.

Liu, R., Zenke, C., Liu, C., Holmes, A., Thornton, P., and Malan, D. J. (2024b). Teaching CS50 with AI: leveraging generative artificial intelligence in computer science education. In *Proceedings of the 55th ACM Technical Symposium on Computer Science Education V. 1*, pages 750–756.

Mayer, J. (2024). Untersuchung von Schüler:innenvorstellungen zu künstlicher Intelligenz. Master's thesis, University of Konstanz.

Maynez, J., Narayan, S., Bohnet, B., and McDonald, R. (2020). On faithfulness and factuality in abstractive summarization. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, *Proceedings of the 58th ACL*, pages 1906–1919.

Miladi, F., Psyché, V., Diattara, A., El Mawas, N., and Lemire, D. (2025). Evaluating a GPT-4 and retrieval-augmented generation-based conversational agent to enhance learning experience in a MOOC. In *Proceedings of the International Conference on Computer Supported Education*.

Neumann, A. T., Yin, Y., Sowe, S., Decker, S., and Jarke, M. (2024). An LLM-driven chatbot in Higher Education for Databases and Information Systems. *IEEE Transactions on Education*.

Pampel, B., Martin, S., and Padó, U. (2025). Regaining Control: Enabling Educators to Build Specialized AI Chat Bots with Retrieval Augmented Generation. In *Proceedings of the 17th International Conference on Computer Supported Education - Volume 2: CSEDU*, pages 371–378. INSTICC, SciTePress.

Pampel, B., Martin, S., and Padó, U. (2026). From capable to trustworthy: Empowering educators to build reliable ai chatbots with retrieval-augmented generation. In *CSEDU 2025, Revised Selected Papers*. Springer. to appear.

Petroni, F., Rocktäschel, T., Riedel, S., Lewis, P., Bakhtin, A., Wu, Y., and Miller, A. (2019). Language Models as Knowledge Bases? In Inui, K., Jiang, J., Ng, V., and Wan, X., editors, *Proceedings of EMNLP-IJCNLP*, pages 2463–2473.

Swacha, J. and Gracel, M. (2025). Retrieval-augmented generation (RAG) chatbots for education: A survey of applications. *Applied Sciences*, 15(8):4234.

Sánchez-Vera, F. (2025). Subject-specialized chatbot in Higher Education as a tutor for autonomous exam preparation: Analysis of the impact on academic performance and students' perception of its usefulness. *Education Sciences*, 15(1).

Thüs, D., Malone, S., and Brünken, R. (2024). Exploring Generative AI in Higher Education: A RAG System to Enhance Student Engagement with Scientific Literature. *Frontiers in Psychology*, 15.

Yu, J., Wu, Y., Shu, D., Jin, M., and Xing, X. (2024). Assessing prompt injection risks in 200+ custom GPTs. In *Secure and Trustworthy Large Language Models Workshop at ICLR*.