Renáta Németh – Annamária Tátrai – Miklós Szabó – Árpád Tamási

Using a RAG-enhanced large language model in a virtual teaching assistant role: Experiences from a pilot project in statistics education

Renáta Németh, professor of sociology, head of department, ELTE Eötvös Loránd University, Faculty of Social Sciences, Department of Statistics, ELTE Research Center for Computational Social Science
Email: nemeth.renata@tatk.elte.hu
Annamária Tátrai, assistant professor, ELTE Eötvös Loránd University, Faculty of Social Sciences, Department of Statistics, ELTE Data for Good Research Group
Email: tatrai.annamaria@tatk.elte.hu
Miklós Szabó, associate professor, head of department, ELTE Eötvös Loránd University, Faculty of Social Sciences, Department of Minority Studies
Email: szabo.miklos@tatk.elte.hu
Árpád Tamási, founder, GoSchool.ai
Email: arpad.tamasi@gmail.com

The role of artificial intelligence (AI) in education is expected to grow, but how it transforms teaching and learning remains unclear. This study explores the use of an AI tutor that is similar to ChatGPT enhanced with retrieval-augmented generation (RAG), in a pilot project at the Faculty of Social Sciences of Eötvös Loránd University in Budapest, Hungary. The tutor provided a searchable knowledge base for students preparing for admission to the MSc in Survey Statistics and Data Analytics. Instructor feedback highlighted the tutor's ability to deliver accurate, textbook-based responses, but noted limitations in addressing real-world complexities. Student feedback, which was gathered through focus groups and surveys, showed high satisfaction and many used the tool for active learning such as comparing concepts and organising material. Students had the flexibility to adapt the tutor to their own learning strategy, and they also noted the importance of the tutor as a time-saving supplement rather than a replacement for comprehensive study. Approximately 15% of student queries demonstrated critical thinking, where students used the AI tutor to confirm their own interpretations. Similarly, around 15% showed active learning, seeking explanations and comparisons or generated study guides, while nearly 30% engaged directly with course material, referencing specific concepts and theories from their readings. Instructor evaluation revealed that 76% of the AI tutor's responses were fully correct, 17% mostly correct and only 6% were misleading. The findings suggest that RAG models hold promise for enhancing learning by offering reliable, interactive and efficient support for students and educators.

Keywords: artificial intelligence, large language model, retrieval-augmented generation, higher education

Generative artificial intelligence (AI), particularly large language models (LLMs), are challenging our perceptions of AI, but not in the dystopian sense that has been familiar in Hollywood. In films, AI often takes the form of a post-human entity; an antagonist plotting against humanity; however, this narrative says more about our collective fears and modern guilt than it does about the reality of this technology. The critical point here is that AI had already become a social **construction** before it became a **reality** in people's everyday lives. Society is not grappling with the technology itself, but with its imagined consequences. This is reflected in how conversations about AI often revolve around exaggerated risks (what could go wrong) rather than informed discussions about its real-world applications and benefits. Fears that AI will overtake human agency distract from the real, more immediate risks posed by the technology such as bias, ethical use and privacy concerns. The danger of society's focus on constructed fears is that it may be harder to identify and mitigate the real risks. Every technology has its dangers, and AI is no exception. The problem is that as long as abstract fears rather than concrete realities dominate public discourse, these dangers could remain unaddressed. In this context, exploring the practical applications of AI is essential. Rather than looking at AI as a distant, existential threat, we should examine its impact in real-world settings. LLMs are already transforming areas such as education, healthcare and customer service and their integration will only increase in the future, whether we are ready for it or not. Understanding and managing their social impact requires a shift from speculative fear to pragmatic engagement. And that is what we have decided to do in this study in the context of higher education.

1. Introduction

1.1 Generative AI as a challenge and opportunity for higher education

The introduction of LLMs, such as the generative pre-trained transformer (GPT), has transformed the AI field in the last few years. These models have performed outstandingly, particularly in the field of analysis and synthesis of natural language and speech *(Kubli–Saboo, 2022)*.

Two characteristically different approaches to LLM tools have emerged in higher education. Some consider them to be dangerous tools that support cheating, while others primarily consider them to be technologies that can support learning and simplify workflows *(Tlili et al., 2024; Hisan–Amri, 2023)*. The benefits and opportunities that such technologies offer include improving the learning process by creating personalised learning materials and automating the generation of educational content, which can reduce educators' workload. AI can also foster innovation by generating new ideas and solutions.

However, excessive use of these tools raises several concerns. Several studies have examined the use of ChatGPT for educational purposes, primarily drawing attention to its controversial perception (e.g. unethical application, cheating, plagiarism: *Hisan–Amri (2023)*; ideological bias: v*an Poucke (2024)* and providing incorrect information: *Lee, 2024*). The oversimplification of learning and scientific creative processes can impede the development of analytical and critical thinking skills, diminishing users' capacity for deep and meaningful learning. Furthermore, generative AI tools may reference other authors' materials and ideas without proper citation, which raises significant ethical issues.

1.2 The challenge of hallucination and RAG model as a solution

LLMs are prone to 'hallucination', i.e. producing factually inaccurate content *(Tonmoy et al., 2024)*, which can become extremely problematic when (as in our case) the models are used for educational purposes. Dozens of solutions have been developed in the past three years to mitigate the hallucination of LLMs *(Tonmoy et al., 2024)*, most notably the RAG technique *(Lewis et al., 2020)*, which can potentially increase LLMs' reliability by using external databases to produce precise responses while explicitly linking references to the generated responses.

Although RAG has demonstrated its potential in various domains *(Fan et al., 2024)*, its application in educational contexts remains underexplored. This research gap motivated us to investigate how LLMs augmented with RAG perform in a higher education context. Our study augments the GPT-4 model with RAG using teaching resources (two textbooks) provided by the course instructors as external databases. In this way, we created an AI tutor that 'knew' all the material of the course and could answer any question the students had (by providing textbook references). Another advantage of this solution is that it equips the LLM with up-to-date, domain-specific knowledge that is controlled by the instructor.

1.3 RAG model in higher education

The risk of unethical use of the RAG model remains; however, our research does not focus on this but on the actual limitations and opportunities for innovation, i.e. possible hallucinations and how the model can work in a higher education context.

The application of RAG in the educational context is under-researched, and examples in the literature have tended to focus on avoiding hallucination, i.e. whether the model provides accurate responses. For example, *Miladi et al. (2024)* tested a GPT-4 model augmented with RAG within a massive open online course (MOOC) focused on AI, determining that it outperformed the standard GPT-4 model and average student performance, providing accurate and context-dependent responses. At the same time, although RAG brought significant improvements in response accuracy, it was not flawless, which is why we also test the accuracy of responses in our research.

1.4 Research design and objectives

Our research introduces an AI tutor in a pilot project at the Faculty of Social Sciences of Eötvös Loránd University in Budapest, Hungary, in a series of consultations to prepare applicants for the entrance examination to the MSc in Survey Statistics and Data Analytics. The tutor provided students with a searchable, retrievable knowledge base referencing the admissions literature to guide students through entrance exam topics as learning objectives. We chose this course because the students here are highly motivated and the stakes for successful learning are high; therefore, we hoped that the students would actively use the AI tutor to achieve their goals. The course's human tutors are among the authors of this article.

Two textbooks provided the exam preparation literature, *Freedman, Pisani and Purves (2005)* and *Rudas (2006)*, which are hereafter referred to as Textbook 1 and Textbook 2. Participation in the research was voluntary, and participating students consented to the analysis of their queries. Nineteen students took part in the research. We ensured that neither the course tutors nor the admissions committee had access to the data prior to the exam to separate the assessment of exam performance from the queries, which allowed students to use the system more freely. Students' access to the AI tutor was terminated the day before the entrance exams started.

Our research questions were broader than whether the model provides accurate responses. In contrast, we sought to understand users' behaviour through in-depth analysis of queries and student/teacher feedback, to determine how well the tutor supported students, how the tutor can be used most effectively and what unknown benefits/limitations the users discovered.

Students' experiences were collected through a focus group and an online survey. The focus group was organised after the second of a series of three consultations, and the online questionnaire was sent out the day before the start of the entrance exams. The information on the phenomenological experience was complemented by an analysis of the corpus of students' queries. To analyse students' queries, we used traditional qualitative (human reading) and automated approaches. We conducted the automated analysis using one of the latest major LLMs, OpenAI's GPT-4o, to evaluate the queries according to 10 criteria. The prompt used for categorisation is presented in the Appendix, and the 10 criteria are as follows. (1) Anthropomorphising the service (yes/no: messages in which students referred to or treated the RAG service as if it had human characteristics or behaviours), (2) questions about course material (yes/no: messages that contained questions directly related to the academic course material), (3) challenging the AI's limits (messages in which students were testing or pushing the boundaries of what the AI can do and/or its limitations), (4) engagement with course material (messages in which students referenced the course literature: directly quoting or paraphrasing the texts and not only asking general questions), (5) critical thinking and analysis (messages in which students demonstrated critical thinking such as comparing and contrasting different perspectives, asking for implications or consequences and questioning assumptions), (6) active learning strategy (messages in which students used the chatbot for active learning strategies such as self-explanation, summarisation or requesting analogies to better understand complex concepts), (7) feedback and satisfaction (messages in which students provide feedback about their experience using the chatbot, expressing satisfaction or dissatisfaction with the responses they receive), (8) interactive pattern (messages in which students used the chatbot more as a tutor and engaged in interactive, back-and-forth exchange), (9) message complexity (rating of the complexity of the messages from 1 to 10, ranging from simple factual questions (1) to complex, multi-part questions that require synthesis of information (10)) and (10) clarity and specificity (from 1 to 10, in which a message was deemed clearer and more specific if the student provided enough context for the chatbot to generate meaningful responses). We did not test the accuracy of the machine classification because this was beyond the scope of our study; therefore, we do not report exact percentage distributions. In contrast, we used the results as an aid to get an idea of which posts were worth qualitatively investigating, e.g. in terms of anthropomorphisation, without reading all the queries.

Our research is relevant to the readers of the Hungarian Statistical Review in several ways. First, machine learning can be considered a branch of statistics, the pilot was conducted in a statistic learning environment and the results were processed using a combination of data collection (focus group, survey, observation) and analysis (qualitative, quantitative, computational) techniques.

In the following chapters we will turn to the results: first the impressions of the course instructors (chapter 2.1), then the results of the focus groups (2.2), the results of the online survey (2.3) and the results of the analysis of the corpus of queries. The latter analysis includes an examination of hallucination (2.4), the distribution of book pages used by the tutor to generate responses (2.5), the temporal pattern of queries (2.6), and a qualitative analysis of the chats (2.7). Finally, we summarise the results and draw conclusions.

2. Results

2.1 Instructor impressions

The instructors first tested the tutor. Notably, complete and accurate responses were obtained for cases in which the information was factually in the textbook (see Query-Answer 1 in the Appendix on the independence of events) as well as cases requiring a more constructive use of the textbooks' knowledge (see Query-Answer 2–4 in the Appendix), in which the practical implications of the difference between correlation and causality, the definition of type I error or the practical/ethical side of the randomised controlled trial were questioned. For example, the AI tutor was deemed to apply the theorem that correlation does not imply causation to the question of whether olive oil consumption is negatively correlated with a person's number of wrinkles and we should consume more olive oil. However, the AI tutor could not find a plausible explanation for the correlation not being causal (e.g. the possible explanation that was always easy for our undergraduate sociology students to find: social status may be a confounder because higher-status individuals consume more olive oil, and their lifestyle and health consciousness may be behind the better skin). In other words, this conveys the impression that this kind of constructivism, which presupposes knowledge of the world, cannot be expected from the AI tutor.

2.2 Focus group results

Participation in the focus group discussion was voluntary and four out of 19 students took part. Despite the relatively small number, the students provided useful insights. One participant did not use the technology and only decided to test the university-provided LLM before the focus group, and one was a casual user. Two 'heavy users' who were familiar with generative AI had used the university-provided system extensively. This understanding naturally influenced the focus group as these participants were the first to respond and their opinions influenced the contributions of the other two.

We broke down the key themes in the focus group discussion around trust, time management and usage strategies.

2.2.1 Trust

Trust was represented in users' general positive expectations. The participants indicated that their level of trust in AI significantly increased during use. One respondent noted that they expected credible information from the AI, primarily since it was based on the obligatory textbooks and they were confident that it would not hallucinate since it did not include other sources. Also, because the university authorised the system, the participants agreed that they 'dared to rely on it' entirely. The majority indicated that they were sure that at one point they would have used some generative AI as aid, but they would have needed to spend more time. This instilled students' optimism regarding the system's reliability.

However, some doubts remained. Participants expressed concerns about how reliable the information provided by the AI was and questioned whether AI systems like ChatGPT and similar technologies were truly dependable for delivering accurate information.

2.2.2 Time management

Participants agreed that using AI saved time. Many noted that they used AI because it helped them to summarise relevant topics, reducing the time needed to study. While they recognised the added value of the LLM in terms of learning, they emphasised that reading was still crucial because AI cannot provide the full context for every subject. However, this sentiment was not specific in that it reflected an activity that they were willing to engage in or agreed to do but did not have the time or energy to do, or perhaps something that they verbally agreed to but thought that the teachers present at the focus group discussion expected them to say.

2.2.3 Usage strategies

The participants reported having used ChatGPT as a supplementary tool, particularly when they felt that getting information directly from books would slow down their learning process. Many mentioned using ChatGPT to write formal letters, answer complex questions and to summarise long texts. This practical reliance on AI illustrates its role as a valuable resource in an academic setting. Overall, AI platforms have become an essential aspect of students' approach to studying and completing academic tasks efficiently.

Concerning our AI tutor, the two heavy users used different learning strategies. One relied on it almost constantly as a time-saving tool to systematically organise the large amount of material when studying. The other used it only to confirm his understanding, e.g. to clarify previously missed definitions and concepts, if he was not working through the textbook in a linear way.

2.3 Online survey results

When we closed access to the AI tutor, we asked students to complete an online survey, including those who had not used the tutor (the questions are presented in the Appendix). Ten users and 14 non-users completed the questionnaire. The typical reason given by students for non-use was lack of time. They were asked whether they used ChatGPT or had used a similar tool during their undergraduate studies, and we also asked them about their trust in these tools. Notably, user and non-user groups did not differ in this respect, which also indicates that students' reasons for non-use were not attitudinal, which could be good news for the introduction of similar AI tutors in higher education.

The users were very satisfied with the tutor. When asked whether they would recommend the tool to other students, they gave an average answer of 9.4 (1: not at all, 10: completely), indicating (in response to an open-ended question) that the tutor was extremely helpful, they received reliable answers and they found it more reliable than ChatGPT.

The students' answers to the open-ended questions provided very useful insights. The students identified specific approaches to using the AI tutor, e.g. using the tutor for short and simple summaries, reviewing material they had already learned and to refer to relevant textbook pages. Students used the AI tutor following different learning strategies. Some actively relied on it to prepare for the exam, summarising chapters and working out examination items. Others mentioned objectives that could be classified as active learning, e.g. requesting comparisons between concepts and models. Finally, others used it more as a supplementary tool that was only used in cases of uncertainty, to supplement their own notes or to generate specific explanations. In response to another open-ended question, students expect the tool to accelerate the learning process, organise material, aid comprehension and remove uncertainty about a difficult topic. Similarly, such uses were noted when recommending the tool to others, and some students stressed that the AI tutor should only be used as a self-check to supplement learning, rather than a substitute for diligence.

When asked about the added value of the tool compared with a textbook, the students indicated that it was more interactive, easier to search, able to answer questions, concise ('you don't have to read all 1000 pages to get a 20 page note'), able to explain or provide a different perspective on a topic and accessible anywhere, anytime.

The students observed no major errors, with only one miscalculation noted and the fact that sometimes things 'stuck' for the tutor, i.e. including its answer to a previous question in the answer to following questions.

All students answered 'yes' to the question of whether they would use the tool in a master's course and whether they would recommend that we offer the tool to our students next year.

2.4 Analysis of the corpus of queries: quality of responses

A total of 252 randomly selected questions were assessed by an instructor with seven years of teaching experience. After excluding 18 non-academic queries and 8 cases in which the system failed to provide an answer, 238 question–answer pairs were analysed. Among them, 5 questions were unrelated to course material, 5 were based on non-existent or contradictory concepts and 7 were simple requests for page references. This left 213 relevant question–answer pairs for evaluation.

In 149 cases (68%), the system provided responses that were accurate, relevant and contained no misinformation, effectively replacing the instructor's role (see Query 7 in the Appendix as an example). Another 23 cases (11%) involved instances in which the system provided correct answers, but during the explanation or elaboration phase, the AI tutor made claims or statements that were not found in the original textbooks or reference materials. Nevertheless, these answers were considered valid, as demonstrated in Query 8. In 10 cases (5%), the answers were mostly correct but lacked some details (see Query 9 for an example). An additional 19 cases (9%) had minor inaccuracies (see Query 10 as an example), while 10 cases (5%) were judged to be clearly misleading or incorrect (as illustrated by Query 11). In 2 cases (1%), the system simply repeated the question without offering meaningful help (as shown in Query 12).

Overall, 79% of the responses were fully correct, 14% mostly correct and 5% misleading. We used a point system to quantify the quality of the responses (+1 for fully correct answers, –1 for completely wrong answers, 0.85 for incomplete but correct answers and 0.75 for mostly correct answers with minor errors). The average score was +0.847, indicating strong performance overall.

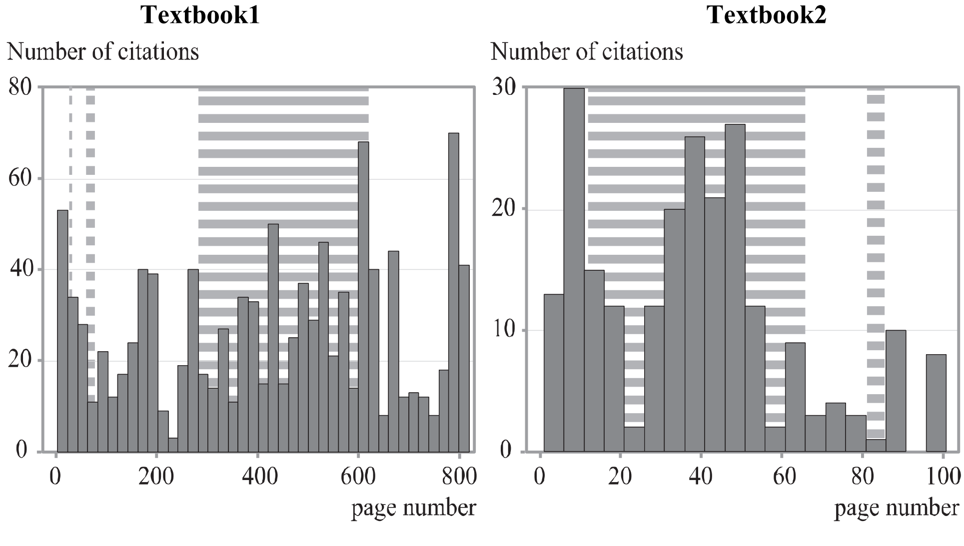
2.5 Analysis of the corpus of queries: references to the textbooks

As the system keeps track of which pages of textbooks were used to generate a response, we can assess the degree to which students processed the textbooks. Figure 1 presents the processing of the two textbooks, with the histogram illustrating how many times each page was used to generate a response. The horizontally striped section shows which chapters were specifically named in the list of items in the entrance exam.

This kind of analysis can be very informative for the instructor of a course using a similar AI tutor as the instructor can see which chapters were more/less frequently referred to in the answers to deduce their difficulty. Figure 1 also demonstrates that the answers to the questions often went beyond the mandatory (striped) chapters, and other pages referenced in the answers were very rarely used (i.e. easier to understand materials or less informative for the items in the entrance exam).

Figure 1

Extent to which textbooks were referenced by the AI tutor



2.6 Analysis of the corpus of queries: temporal patterns

Nineteen students participated in our study, posing 825 questions to the tutor, representing an average 39 questions per person, with the most active student asking 179 questions.

The students had access to the AI tutor for two months and made queries on an average of 15 different days, the least active student used the tutor on only one day and the most active student used the tutor on 29 days.

In terms of the distribution of queries over time, the average number of queries made by a given student on a given day was 4.7, but with a wide variation, with 1 query on the least active student day and 29 queries on the most active student day. If we looked at the average of this indicator per student (i.e. the average number of daily queries per student), we obtained the same average, with a minimum of 1.3 and a maximum of 13.

These findings revealed characteristically different learning strategies, which were consistent with the focus group and online survey results. Some students only submitted a single query to illuminate some point in the course material, and others engaged in interactive, ongoing discussions. The latter students used the chatbot more as a tutor with interactive, back-and-forth questions.

2.7 Analysis of the corpus of queries: qualitative analysis of the categories classified by GPT-4o

We will not go into all 10 aspects below, only analysing the messages that provided the most significant results. **Anthropomorphisation** of the AI tutor was rare and were predominantly cases in which the student addressed the AI tutor as a human being (‘Please explain the following topic in 2 pages!’ – the original Hungarian version addressed the tutor using the term ‘te’ (you), which is used in informal settings, for example among friends). In the non-anthropomorphising questions, students used a general subject or plural first person (‘Why is estimation needed?’ ‘How do we calculate the area under the normal curve?’). Another type of anthropomorphising was accompanying the request with polite phrases (‘Please derive this formula’), to express gratitude for the answer or greet the tutor (‘hello’, ‘how are you?’). Finally, anthropomorphisation was also indicated when students did not ask questions directly, but referred to the AI tutor’s skills at a meta-level, assuming a kind of awareness (‘Why did you suggest this formula to me?’ ‘In other chats you could send it in a regular format before’. ‘So you cannot interpret the values in the tables?’).

Questions that were not directly related to the **course material** concerned the topics of the entrance exam, the entrance exam procedure, about which the tutor had no information. Some students also posed extracurricular questions about the time needed for preparation or the use of the chat platform itself, and some questioned the general factuality of the AI tutor extracurricularly ('Two thirds of six seventh are four seventh? Yes/no?'). Some questions also asked about the tutor's functioning ('Please answer all my questions in this chat concisely and always use the course materials. Answer in English even if the document is in Hungarian').

A relatively high proportion of queries demonstrated **critical thinking** (~15%), these were messages in which students, for example, used the AI tutor to test their own interpretations. Similarly, a relatively high proportion of messages showed signs of **active learning** (~15%), in which the students' aim was to explain, summarise, compare, ask for analogies or generate learning guides for better understanding. For example, for the message 'Please test my knowledge of the test of significance', the student received a summary and several test questions related to the topic. Furthermore, Query 5 in the Appendix demonstrates the student's self-testing with the help of the AI tutor. Query 6 in the Appendix presents an example of a student requesting help with ethical issues related to new statistical knowledge. In alignment with the relatively high rate of critical thinking and active learning, students often considered the **complexity** of the messages to be high (at least 7 on a 1–10 scale in about 15% of the cases). Almost 30% of the messages demonstrated **engagement** with course material, in which students referenced the course textbooks (requesting clarifications on specific concepts, theories or arguments presented in the readings), not only asking general questions.

3. Summary and conclusions

In this study, we summarise our experience with an AI tutor like ChatGPT, where the LLM was augmented using RAG. Our aim was to understand user behaviour by analysing digital tool use in a controlled environment and via student/teacher feedback to determine how well the application supported students' preparation, how they used the AI tutor most effectively and what unknown benefits/limitations were discovered. We used focus groups and an online survey to collect students' experiences, which complemented the analysis of the student–AI tutor interactions. The latter corpus used traditional qualitative and automated (GPT-4o) approaches.

The instructors first tested the tutor, finding that complete and accurate responses were obtained, not only in cases in which the information was factually in the textbooks but also in cases requiring a more constructive use of the textbooks' content. However, we determine that some forms of constructivism that presupposes knowledge of the world cannot be expected from the AI tutor.

According to the focus group and online survey results, AI platforms have become an essential part of our students' approach for studying and efficiently completing academic tasks. Most of the students had experience with ChatGPT and were familiar with the phenomenon of hallucination; therefore, they also approached the tutor with caution, checking it several times, but had a positive experience.

The users expressed high satisfaction with the AI tutor, indicating that it was extremely helpful, they received reliable answers and they found it more reliable than ChatGPT. The students observed no major errors and answered 'yes' to the question of whether they would use the tool in a master's course.

Participants agreed that using the AI tutor saved time. Students used the tutor following different learning strategies. Some actively relied on it to prepare for the exam, summarising chapters and working out examination items. Others noted objectives that could be classified as active learning, e.g. requesting comparisons between concepts and models. Finally, others used it as a supplementary tool that was referenced only in cases of uncertainty to supplement their notes or ask for specific explanations.

These results were also consistent with our analysis of temporal patterns in queries, revealing characteristically different learning strategies. Some students only submitted a single query to illuminate some point in the course material, and others engaged in interactive, ongoing discussions, using the chatbot more as a tutor via interactive, back-and-forth questions.

Qualitative analysis of the students' messages revealed several important findings. A high level of engagement and motivation to learn the material was evident in several aspects of the corpus of queries in which a relatively high number of messages exhibited critical thinking, active learning, engagement and complexity. A few cases of AI tutor's anthropomorphisation were noted and it may be important to prevent this by making students aware of the machine responses to use the tutor critically. The research also identified possible future directions for improvement based on otherwise important questions that were unrelated to the course material.

We coded a randomly selected sample of question–answer pairs by quality. According to the results, 76% of all question–answer combinations were correct, 17% were mostly correct and 6% were completely misleading. Scored on a numeric 0–1 scale (where +1 meant the system answered all questions perfectly and −1 means the AI tutor gave the wrong answer in all cases), the mean of the answers was +0.826. Although using a different design and measurement, these results are broadly aligned with other educational research using the RAG model, e.g. *Miladi et al. (2024)* tests a RAG-augmented GPT-4 model within an AI-focused MOOC and reveals an 85% success rate.

In summary, our findings indicated the potential of using RAG models to enhance the use of LLMs for educational purposes. Our results are promising for students and instructors alike. Using an AI tutor, students can access more personalised and engaging learning experiences and have the flexibility to adapt the tutor to their own learning strategy, which saves time and maintains their engagement and motivation. For instructors, the AI tutor can offer a range of additional support beyond the success of an individual course, helping them to understand how students think, what interests them, what portion of the course material they are processing and what part remains unprocessed by queries. We also found errors in the AI tutor's answers, highlighting the importance of fostering critical thinking skills in our students.

Our analysis also has limitations. This was an observational study. While we acknowledge that the best way to measure the effectiveness of the AI tutor would have been a randomised controlled trial, we did not want to cause undue frustration for students taking the exam by not giving them access to the tutor if they wished. In addition, we only tested the tutor for one type of course and it is possible that the tutor might work differently in different disciplines using different pedagogical methods. Finally, we did not investigate the reliability of the automated classification of queries, which needs to be addressed in a larger study.

For these reasons, we plan to test several types of courses in the near future, design a randomised controlled trial, and include human annotators to investigate the difference between machine and human classification.

Although the scale of our project was limited and the AI tutor application we tested is an initial prototype, we assert that our solution presents a pioneering step towards AI-enabled tutoring systems that could democratise access to personalised education and provide an opportunity to rethink current educational paradigms.

Appendix

Questions used in the online survey
Prompt for categorizing queries
Query-Answer 1 (run by an instructor, translated from Hungarian)
Query-Answer 2 (run by an instructor, translated from Hungarian)
Query-Answer 3 (run by an instructor, translated from Hungarian)
Query-Answer 4 (run by an instructor, translated from Hungarian)
Query 5 (run by a student, translated from Hungarian)
Query-Answer 6 (run by a student, translated from Hungarian)
Query-Answer 7 (example of a correct answer, got by a student, translated from Hungarian)
Query-Answer 8 (example of a correct answer provided by the tutor, though the response was not based on the material covered in the textbook)
Query-Answer 9 (example of an incomplete answer provided by the system)
Query-Answer 10 (example of an answer with minor inaccuracies)
Query-Answer 11 (example of a misleading answer)
Query-Answer 12 (example of a simple repetition of the question without offering meaningful help)

Questions used in the online survey

Is your current or previous work related in any way to artificial intelligence?
Do you use ChatGPT (or similar interactive chat tools)?
- Did you use a similar type of tool during your undergraduate studies?
- What do you use these tools for? You can choose more!
- What other purposes do you use them for? (open-ended)
- Have you ever experienced any of the following when using ChatGPT or similar tools? You can tick more than one answer!
- What other unusual phenomenon have you noticed? (open-ended)

How much confidence do you have in these tools? (On a scale of 1 to 10, where 1 is not at all, 10 is completely)
Have you used the AI tutor we made available?
- Why did you not use the application?
- What other reasons have you had for not using the app? (open-ended)
- In which cases would you have used the app?

How likely would you be to recommend the tool to another student? (scale 0–10, where 0 is not at all, 10 is completely)

What is your primary reason for giving this answer? (open-ended)
What would it take for you to recommend the tool to others? (open-ended)
What did you expect from the tool before you started using it? (open-ended)
What did you fear about the tool before you started using it? (open-ended)
Could you use the tool in an efficient way?
What was your strategy for using the platform? (open-ended)
What user problems did you encounter when using the tool? (open-ended)
What did you like most about the tool? (open-ended)
What could be done to improve the tool? (open-ended)
What are its advantages or disadvantages compared with the 'wild' ChatGPT or other similar free applications? (open-ended)
What is the added value compared with the textbooks? (open-ended)
What would you recommend others use the tool for? (open-ended)
Would you use it in a master's class as a forum in which it is not embarrassing to ask any question?
Would you recommend making the tool compulsory or recommend it for the next year's course?
What incentives do you think could be put in place to encourage university students to use the tool? (open-ended)

Prompt for categorising queries

Categorise and Rate Student Messages
You will be given a random assortment of student messages from a university retrieval-augmented generation (RAG) service chat.
For each message you get the user message, the preceding chatbot response and the course title.
Your task is to analyse these messages, categorise and rate them based on the following criteria:
1. **Anthropomorphising the Service** (anthropomorphising): Messages in which students refer to or treat the RAG service as if it has human characteristics or behaviours.
2. **Questions About Course Material** (course_material): Messages that contain questions directly related to the academic course material.
3. **Challenging the AI's Limits** (challenge): Messages in which students test or push the boundaries of what the AI can do or its limitations.

4. **Engagement with Course Material** (engagement): Messages in which students reference the course literature by directly quoting or paraphrasing textbook content and not just asking general questions. Look for signs of deep engagement, such as requesting clarifications on specific concepts, theories or arguments presented in the readings.
5. **Critical Thinking and Analysis** (critical_thinking): Messages in which students demonstrate critical thinking such as comparing and contrasting different perspectives, requesting implications or consequences and questioning assumptions. These are messages in which students, for example, use the chatbot to test their own interpretations or to challenge the arguments presented in the literature.
6. **Active Learning Strategy** (active_learning): Messages in which students are using the chatbot to support active learning strategies such as self-explanation, summarisation or requesting analogies to better understand complex concepts. In these messages students use the chatbot, for example, to aid in study habits such as generating study guides or practice questions.
7. **Feedback and Satisfaction** (feedback): Messages in which students provide feedback about their experience using the chatbot, expressing satisfaction or dissatisfaction with the responses received.
8. **Interactive Pattern** (interactive): Messages in which students use the chatbot more as a tutor using interactive, back-and-forth questions, referencing the previous answer from the chatbot to continue the exchange.
9. **Message Complexity** (complexity): Rate the complexity of the messages from 1 to 10, ranging from simple factual questions (1) to complex, multi-part questions that require synthesis of information (10). A message is more complex when the student requests summaries, detailed explanations or critical analyses of the course textbooks.
10. **Clarity and Specificity** (clarity): Rate the clarity and specificity of the messages from 1 to 10. Are students using precise language (10) or are they vague and ambiguous (1)? A message is clearer and more specific if the student provides enough context for the chatbot to generate a meaningful response.
For each message, indicate which category or categories it falls into and provide the ratings. A message may belong to more than one category if applicable.
All true/false, except: ‘complexity’: 5, ‘clarity’: 7

Query-Answer 1 (run by an instructor, translated from Hungarian)

How can the independence of events be defined?
The independence of events can be defined as follows:

Two events are independent if the occurrence of one event does not change the probability of occurrence of the other event (Source: *'Statisztika', David Freedman, Robert Pisani, Roger Purves, Daniel Müller, p. 278*).
This means that when two events are considered and the occurrence of one event has no effect on the probability of the other event occurring, then these events are called independent. This definition is essential in the field of probability and statistics as it allows us to investigate and understand the relationships between different events.

Query-Answer 2 (run by an instructor, translated from Hungarian)

People who consume more olive oil have less wrinkles. Would you advise me to consume more olive oil if I don't want to wrinkle?
I have not found any direct information on how increasing olive oil consumption affects wrinkles in the literature; however, the articles highlight that correlation does not necessarily imply causation. For example, in looking at the association between fat intake and cancer, a note that although high fat consumption is associated with a high incidence of certain types of cancer, this does not necessarily mean that fat consumption causes cancer. Similarly, the link between increased consumption of olive oil and less pronounced wrinkling should be treated with caution, as other factors may also be *involved ('Statisztika', David Freedman, Robert Pisani, Roger Purves, p. 180)*.
Therefore, I cannot directly recommend increasing olive oil consumption to prevent wrinkling as no clear causal relationship can be established on the basis of the available information. It may be worth considering other, scientifically based skin care and health maintenance methods.

Query-Answer 3 (run by an instructor, translated from Hungarian)

Can a randomised control group trial be conducted to study the harmful effects of smoking?
It is not practical or ethically feasible to conduct a randomised control group trial to investigate the adverse effects of smoking. In Chapter 2 of the book *'Statisztika' (p. 29)*, *David Freedman, Robert Pisani and Roger Purves* explain that observational studies necessarily deal with the effects of smoking because one cannot expect someone to smoke for years just to do a statistician a favour. In this case, researchers try to establish the effect of smoking by comparing smokers (the 'treated' or 'exposed' group) with non-smokers (the 'control' group). Such a

comparison reveals a strong association between smoking and a number of diseases, providing indirect evidence of causality. However, a randomised control group trial in which participants are divided into two groups and one group is deliberately induced to smoke is unethical and therefore cannot be conducted.
Source: *'Statisztika', David Freedman, Robert Pisani, Roger Purves, p. 29.*

Query-Answer 4 (run by an instructor, translated from Hungarian)

If there is no difference between the average earnings of women and men, and I perform a t-test 100 times to test the difference in means with a type I error of 5%, how many times is the test expected to be significant?
When you perform a t-test to examine the difference in means and no real difference is found between the two populations, a type I error (or alpha error) is the probability that the test will incorrectly report the difference as significant when there is no difference. It is therefore the proportion of false positives.
The type I error of 5% mentioned in the question asked means that an average 5 out of every 100 t-tests are expected to falsely report the difference as significant when there is in fact no difference between the average earnings of women and men. This means that if we perform the test 100 times, we expect the result to be significant 5 times with a 5% first-order error, assuming that there is no difference. This information follows from the basic principles of statistical procedures described in the book *'Statisztika'* by authors *David Freedman, Robert Pisani, Roger Purves (pp. 467–469).* The concept and use of the first-order error is key to the interpretation of statistical tests, especially when the significance level is set at a predetermined value, such as the commonly used 5% level.

Query 5 (run by a student, translated from Hungarian)

With the Z-test, we cannot and should not estimate the sample variance since the test requires that the population variance be known. For the t-test, we can estimate the population variance with the sample variance, but it is recommended to use a corrected variance, since the number of elements is small for the t-test. Right?

Query-Answer 6 (run by a student, translated from Hungarian)

If you are conducting a randomised controlled experiment, does the use of chance raise ethical questions? I mean, we know that we are randomly assigning x and y

to treatment and control groups but there will still be people who don't get vaccinated.
In the design and implementation of randomised controlled trials, the use of randomisation can indeed raise ethical issues, especially in situations in which one group receives some potentially beneficial treatment (e.g. vaccination) while the other group does not (control group). These ethical issues are particularly important in medical research where the effects on the health of participants directly affect their quality of life. [...]

Query-Answer 7 (example of a correct answer from a student, translated from Hungarian)

What is an estimate?
Estimation is a statistical method to approximate an unknown population parameter (such as the mean or proportion of the population) using data from a sample. The value calculated in the estimation process is called an estimate. The term estimation can also refer to the method itself and its result. Mathematically, an estimate is represented by the statistic obtained as a function of observations, which is considered the best approximation of the population parameter. Random error, which arises from the fact that the sample is only a portion of the whole, plays a fundamental role in the estimation process. This random error, or sampling error, is the determining factor in the accuracy of the estimate *(Statistics by David Freedman, Robert Pisani, Roger Purves, p. 389).*

Query-Answer 8 (example of a correct answer provided by the AI tutor, although the response was not based on the material covered in the textbooks, translated from Hungarian)

What is imputation?
Imputation is a statistical technique used to fill in missing data. Missing data can cause problems in statistical analyses as it can bias the results and reduce the representativeness of the sample. Imputation aims to replace missing data with estimates so that the full data matrix can be used in the analysis.
Types and Methods

Simple Imputation:

- **Replacement by average**: missing values are replaced by the average of the variable. This is a simple but often sub-optimal method as it can reduce the variability of the data.

- **Median or mode imputation**: Missing values are replaced by the median or most frequent value (mode) of the variable. This method is similar to mean substitution but is less sensitive to extreme values.

More Advanced Imputation Methods:

- **Regression Imputation**: missing values are estimated using a regression model that predicts missing values based on available data.
- **Hot Deck Imputation**: missing values are replaced by the value of another observation with similar characteristics.
- **Multiple imputation**: this method imputes missing values several times (usually 5–10 times) using different estimation methods and then analyses the resulting data together, considering the uncertainty of the imputation.

Advantages

Data completeness: imputation ensures that the data matrix is complete, allowing the full dataset to be used in the analysis.

Preserving representativeness: filling in missing data helps to preserve the representativeness of the sample, especially if the missing data are randomly distributed.

Imputation increases the **accuracy and reliability** of statistical analyses by reducing bias due to missing data.

Disadvantages and Limitations

Introduction of bias: If imputation is not conducted properly, it can introduce bias into the data. For example, if missing values are missing non-randomly, imputation can bias the results.

Reducing Variance: simple imputation methods such as mean-replacement can reduce data variance, which can distort the analytical results.

Complexity and Resource Requirements: more advanced imputation methods such as multiple imputation are more complex and require more computational resources.

More complex imputation methods require more resources.

In a poll, some respondents may not answer a particular question. If, for example, there are no answers to a question on income, the missing values can be estimated using a regression model based on other respondent characteristics (e.g. age, gender, occupation).

Source:

- *Statistics* by *David Freedman, Robert Pisani, Roger Purves*, which discusses in detail methods of dealing with missing data and their implications.
- *Public opinion polling* by *Tamás Rudas*, which describes imputation techniques used in public opinion polls and their applicability.

***Remark from the paper's authors: *Freedman et al.*'s book Statistics does not include the term 'imputation', while *Tamás Rudas'* book on public opinion polling contains only this short summary (translated from Hungarian):
"If the results of the analysis are more complex indicators than this, then substituting missing data (in the jargon: imputing) has the negative consequence that the responses in the completed data set appear more uniform than they are in reality. The short reason for this is that otherwise identical people, if one answered a question and the other did not, are essentially imputed under the assumption that the non-respondent would have given the same answer as the respondent, which is not necessarily true. There are more complex ways of making up for missing observations than those outlined here that do not have this shortcoming, but it is certainly true that the missing information cannot be made up in any meaningful way. At most, imputation can provide an approximation, the goodness of which requires that the actual (not fully observed) data have some advantageous property" *(Rudas, 2006, pp. 45–46.)*.

Query-Answer 9 (example of an incomplete answer provided by the system, translated from Hungarian)

How do we calculate the expected value?
An example of how to calculate the expected value can be found in the document *'Statistics'* by *Daniel Müller*. In this example, the way to determine the expected value is as follows: first, the 'observed value' is determined, which in this case was 2006 hits. The expected value resulting from the null hypothesis is 1875 hits. Then, the difference between the observed value and the expected value is calculated, which in this case is 131 hits (2006−1875). The difference between the expected value and the observed value is divided by the standard error of the number of hits, which in this case is 37, to obtain the z-statistic, which in this case is 3.5 *(Statistics by David Freedman, Robert Pisani, Roger Purves, p. 534)*.
This example relates to a specific statistical problem, but in general the expected value is calculated by multiplying each possible outcome by its probability of occurrence and then summing them. However, this method may vary depending on the type of data and distribution you are working with.

Query-Answer 10 (example of an answer with minor inaccuracies, translated from Hungarian)

What is the difference between one and two-sample tests?
The difference between one- and two-sample tests lies in their basic purpose and application.
One-sample test
The purpose of a one-sample test is to determine whether the mean of a sample is significantly different from a given value or population mean. For example, if you want to test whether the mean test score for a class is significantly different from 70, you will use a one-sample test.
Two-sample test
A two-sample test is designed to test the difference between the means of two different samples. This test analyses whether the difference between the two samples is significant or simply the result of random fluctuation. For example, if you want to compare the average test scores of men and women, you will use a two-sample test.
Examples and analyses
One-sample test: large samples help to measure variation accurately, but the z-test compares variation to its standard error, so for large samples a small difference can give an impressive z-score. For example, the increase in average scores on mathematics tests between 1978 and 1992 *(Statistics* by *David Freedman, Robert Pisani, Roger Purves*, *p. 611).*
Two-sample test: a two-sample z-test can be used in an analysis, but a box model is needed to interpret the results. For example, comparing male and female enrolment rates in a university *(Statistics* by *David Freedman, Robert Pisani, Roger Purves, p. 614).*
One- and two-sample tests are thus designed to answer different statistical questions and use different methods to examine differences between samples.

Query-Answer 11 (example of a misleading answer, translated from Hungarian)

What is the point of Freedman Chapter 22?
The substance of Chapter 22 of the *'Statistics'* document has not been directly presented on the basis of the information available; however, a section found in the document suggests that the chapter is likely to deal with the accuracy of measurements and random error that can cause measurements to deviate from the exact value. It shows through an example how the variation between repeated

measurements is reflected in random error changes, and how these variations can be quantified by the standard deviation of the data.
Specifically, the textbook provides an example of a measurement in which the average of 100 measurements was 405 micrograms under 10 grams, which is very likely close to the exact weight of NB 10. In this context, the textbook discusses the concept of random error and its calculation, pointing out that the relationship between random error and the standard deviation of measurements can be mathematically determined.
It also refers to a later chapter (Chapter 24), which discusses how to calculate the likely magnitude of the random error for this type of average. Based on the review, it appears that Chapter 22 presents important principles and methods to help understand how to deal with and interpret random errors in statistical measurements.
Source: *'Statistics'* by *David Freedman, Robert Pisani, Roger Purves*.

Query-Answer 12 (example of a simple repetition of the question without offering meaningful help. Translated from Hungarian)

What does it mean that the correlation coefficient measures the closeness of the compression relative to the standard deviations and not in absolute numbers?
The correlation coefficient (r) measures the closeness of a linear relationship between two variables; not in absolute values but relative to the variances of the variables. This means that the correlation coefficient measures the closeness of the compression between points relative to the standard deviations of the data.
This can be explained as follows:

Standardisation: when calculating the correlation coefficient, variables are converted into standard units, i.e. deviations from the mean are divided by the standard deviation. This makes the correlation coefficient unitless.

Relative measure: The coefficient shows how closely points cluster around a line; not in absolute numbers but relative to the standard deviations. Thus, the value of the correlation coefficient remains the same regardless of the unit in which the data are expressed.

An example from the book *Statistics:*

- Two scatter plots may have the same correlation coefficient (e.g. 0.70), but the closeness of the locations of the points may differ. If the points on one of the charts are more closely packed around a line, this may be due to a lower standard deviation. In this case, the correlation coefficient remains the same because it is calculated in relative units rather than absolute values.

As the *Statistics* book says:
'When calculating the correlation coefficient, we convert our variables into standard units: the deviations from the mean are divided by the standard deviation. Thus r measures the closeness of compression in relative terms, relative to the standard deviations, rather than in absolute numbers' *(Statistics* by *David Freedman, Robert Pisani, Roger Purves*, *p. 171).*
It is therefore important to understand that the correlation coefficient does not measure absolute values, but rather measures the closeness of the relationship relative to the variances of the variables.

References

Fan, W. – Ding, Y. – Ning, L. – Wang, S. – Li, H. – Yin, D., ... – Li, Q. (2024): A survey on rag meeting llms: Towards retrieval-augmented large language models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining,* 6491–6501.

Freedman, D. – Pisani, R. – Purves, R. (2005): *Statisztika.* Typotex, Budapest.

Hisan, U. K. – Amri, M. M. (2023): ChatGPT and medical education: a double-edged sword. *Journal of Pedagogy and Education Science,* 2(1), 71–89.

Kublik, S. – Saboo, S. (2022): *GPT-3*. O'Reilly Media, Inc.

Lee, H. (2024): The rise of ChatGPT: Exploring its potential in medical education. *Anatomical Sciences Education*, 17(5), 926–931.

Lewis, P. – Perez, E. – Piktus, A. – Petroni, F. – Karpukhin, V. – Goyal, N. – Kiela, D. (2020): Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems,* 33, 9459–9474.

Miladi, F. – Psyché, V. – Lemire, D. (2024): *Comparative performance of GPT-4, RAG-augmented GPT-4, and students in MOOCs.* TELUQ University, Canada.

Rudas, T. (2006): Közvélemény-kutatás – Értelmezés és kritika. Corvina, Budapest.
https://r-libre.teluq.ca/3266/1/Comparative%20Performance%20of%20GPT-4.pdf

Tlili, A. – Shehata, B. – Adarkwah, M. A. – Bozkurt, A. – Hickey, D. T. – Huang, R. – Agyemang, B. (2023): What if the devil is my guardian angel: ChatGPT as a case study of using chatbots in education. *Smart Learning Environments,* 10(1), 15.

Tonmoy, S. M. – Zaman, S. M. – Jain, V. – Rani, A. – Rawte, V. – Chadha, A. – Das, A. (2024): A comprehensive survey of hallucination mitigation techniques in large language models. *arXiv preprint* arXiv:2401.01313.

van Poucke, M. (2024): ChatGPT, the perfect virtual teaching assistant? Ideological bias in learner-chatbot interactions. *Computers and Composition,* 73, 102871.