On the Effectiveness of Prompt-Moderated LLMs for Math Tutoring at the Tertiary Level

Sebastian Steindl, Fabian Brunner, Nada Sissouno,
Dominik Schwagerl, Florian Schöler-Niewiera, and **Ulrich Schäfer**
Ostbayerische Technische Hochschule Amberg-Weiden, Germany
{s.steindl, fa.brunner, n.sissouno, d.schwagerl,
f.schoeler-niewiera, u.schaefer}@oth-aw.de

Abstract

Large Language Models (LLMs) have been studied intensively in the context of education, yielding heterogeneous results. Nowadays, these models are also deployed in formal education institutes. While specialized models exist, using prompt-moderated LLMs is widespread. In this study, we therefore investigate the effectiveness of prompt-moderated LLMs for math tutoring at a tertiary-level. We conduct a three-phase study with students (N=49) first receiving a review of the topics, then solving exercises, and finally writing an exam. During the exercises, they are presented with different types of assistance. We analyze the effect of LLM usage on the students' performance, their engagement with the LLM, and their conversation strategies. Our results show that the prompt-moderation had a negative influence when compared to an unmoderated LLM. However, when the assistance was removed again, both LLM groups performed better than the control group, contradicting concerns about shallow learning. We publish the annotated conversations as a dataset to foster future research.[1]

1 Introduction

The publication of ChatGPT (OpenAI, 2022) and the following developments around other LLMs have allowed easy, informal access to Language Models even without technical knowledge.

The combination of the large amount of training data, model size, and instruction-following ability has led to LLMs becoming not only a subject of research but also a tool used by the broader population due to its ease of use and perceived usefulness and novelty (Wolf and Maier, 2024). However, many members of the general public equate LLMs to Artificial Intelligence and can lack an understanding of the limitations of these models. Problems such as hallucinations and bias can be obfuscated by the very strong Natural Language Understanding and Natural Language Generation capabilities of current state-of-the-art models. This can lead to a naive, unreflected usage of LLMs for knowledge tasks (Krupp et al., 2024).

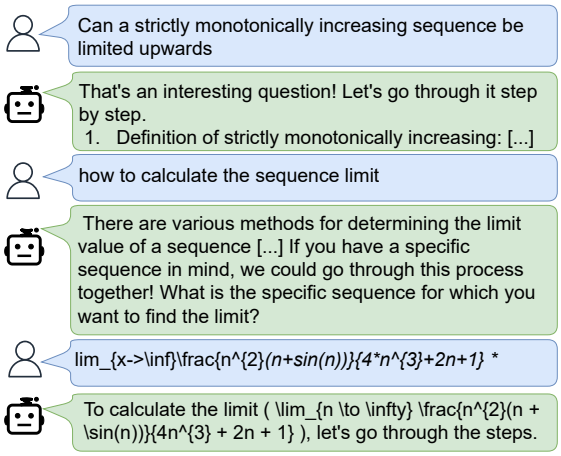

Figure 1: Beginning of an example dialogue from the GMADE dataset (translated from German). Omissions made for readability. The user was interacting with the tutoring version of the LLM.

In this work, we extend previous research into the effectiveness of LLMs in education by providing empirical results for teaching math in formal, tertiary education. To facilitate real-world, large-scale adoption of LLMs in educational institutions, we deem two factors as necessary: ease-of-use and low cost. We therefore use a realistic technical setup to perform our experiment on students' performance with and without LLM assistance. The students use the LLM GPT-4o-mini to solve challenging math assignments of different formats, incorporating the topics sequences, continuity, and differential calculus (DC). We test an out-of-the-box, unmoderated LLM and its prompt-moderated version. Prompt-moderation in this case means that the model was prompted to not give away the solution directly, but to act as a tutor performing

[1] https://github.com/sebastian-steindl/GMADE_Data

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 11310–11323

scaffolding.

We provide the resulting conversations as the German Math Assistant Dataset for LLMs in Education (GMADE), which, to the best of our knowledge, is the first of its kind. An exemplary excerpt of such a conversation can be found in Fig. 1. Our main contributions are, firstly, the investigation into the effect of LLMs, prompt-moderated and unmoderated, in formal tertiary education. We thereby uncover shortcomings of their tutoring competence. Secondly, we perform an in-depth analysis of the students' test performance. Lastly, we publish GMADE to enable future work into understanding and hopefully improving LLMs in education, especially regarding math tutoring.

Our results show the prompt-moderation to be ineffective and thus align with Krupp et al. (2023) while contradicting results by Bastani et al. (2024). This indicates that future work needs to be done on LLM-tutors making use of scaffolding to be efficient in educational settings.

2 Background and Related Work

Learning Theories and LLMs. The concept of tutoring or scaffolding, where an expert helps a less-skilled student to arrive at the solution to a problem, is long-established (Wood et al., 1976; Anghileri, 2006). Moreover, generative AI integrates well with certain learning theories such as self-efficacy and self-determination but might also introduce challenges such as cognitive overload and reduced social contact with peers (Wu, 2023). Another risk is students becoming overly dependent on LLMs (Wang et al., 2024). Thus, the prospect arises that an adept combination of tutoring LLMs and classic teaching can benefit education (Kumar et al., 2023; Ishida et al., 2024).

LLMs in Education. LLMs have become prominent in education, with research focusing on assisting the student, the teacher, and adaptive learning (Wang et al., 2024). Student assistance can come in the form of Question Solving (QS), Error Correction and Confusion Helper (Wang et al., 2024). The problem-solving capabilities of LLMs for knowledge tasks have become evident by multiple benchmarks (White et al., 2024). Even though these capacities stem only from sampling from a learned distribution, they have proven effective for many non-trivial problems.

Combined with their widespread accessibility, it is obvious that students will use such tools extensively for their assignments, especially for QS.

Previous studies showed that unmoderated use of LLMs can impede the students' learning (Bastani et al., 2024; Krupp et al., 2024). As reported by Krupp et al. (2024), students mostly chose the convenient way of copy & pasting a question directly into the chatbot instead of trying to come up with an adequate prompt. This indicates limited reflection and cognitive attention. Bastani et al. (2024) found that during the assisted usage, both the normal and tutor LLM groups showed improved performance over the control group, with the students using a tutor LLM achieving the best scores. However, during an unassisted exam, the cohort using an unmoderated LLM showed worse performance than the group without any assistance, while the tutor cohort was on par with them. They conclude that unmoderated LLM assistance can impede effective learning. Albeit, in the study conducted by Krupp et al. (2023), the scaffolding LLM was the worst approach, and students tried to evade the scaffolding.

In this study, we further investigate the impact of LLM tutoring in a mathematical QS scenario under exam conditions, comparing unmoderated LLM and tutoring LLM use as well as no LLM support.

Pedagogical Alignment of LLMs. To improve the usefulness of LLMs in educational settings, previous work has focused on pedagogical alignment of LLMs. Tack and Piech (2022) propose to measure a model's educational abilities by evaluating if its responses resemble those of a human teacher, show understanding of the student, and are helpful to the student. Similarly, Maurya et al. (2025) define eight evaluation dimensions across which the pedagogical abilities of an LLM can be measured and publish a benchmark towards this. Lastly, Sonkar et al. (2024) constructed a preference dataset designed for pedagogical alignment to fine-tune open-weight LLMs.

The focus of our work is not on improving the pedagogical alignment of LLMs but instead on measuring the effect of prompt-moderation as alignment and LLM-assistance in general on student performance. We measure this effect in a realistic setting by comparing the students' performance with and without LLM-assistance in an exam situation.

Datasets for Mathematical Question Solving. To measure mathematical QS, multiple datasets have been proposed. English benchmarks for

measuring problem-solving capabilities include GSM8K (Cobbe et al., 2021) or MATH (Hendrycks et al., 2021). Moreover, research has turned towards synthetic datasets to tackle the lack of training data. Macina et al. (2023b) synthesize a mathematical QS dataset by using LLMs to simulate students while using human teachers. Chevalier et al. (2024) simulate both students and tutors to source dialogues about a given document. To the best of our knowledge, GMADE is unique in that it provides real, German mathematical QS conversations between students and the LLM assistants.

3 Experimental Design

Our experiment aims to measure the effectiveness of LLM assistance for math tutoring. The setup of our study is analogous to Bastani et al. (2024), who found that prompt-moderation improved the tutoring. We conducted the experiment with voluntary participants from the first semester mathematics for engineers and computer scientists courses. To reduce selection bias, the study was conducted during the last lecture of the semester and also served as a preparation for the exam. This resulted in N=49 participants, which equals roughly 70% of the eligible students.

The experiment was divided into three phases: *review*, *exercise*, and *exam* phase – each lasting 40 minutes with a break of 20 minutes in between. In the *review* phase, the lecturer presented topics that were relevant to the following phases to all participants. Since these subjects had already been discussed during the normal lecture period, this served as a review for the participants and aligned their prior knowledge. In the *exercise* and *exam* phases, the students were presented with a number of mathematical questions. The assignments were printed on paper, and the students were required to answer on paper. This avoids the naive copy & paste behavior that Krupp et al. (2024) observed.

3.1 Cohorts

Starting with the *exercise* phase, the participants were divided into three groups with separate rooms: a reference cohort using coursework material only (referred to as *NON* in the following), one using the unmoderated LLM (*GPT*), and one with the prompt-moderated LLM which should act as a tutor (*TUT*). Following Bastani et al. (2024), the tutor LLM only differs in its system prompt. While approaches exist that specifically fine-tune models for tutoring (Team et al., 2024; Ross et al., 2025), the case of only prompt-moderated LLMs is very relevant. First, this approach is easily accessible and allows to always use the latest models. Moreover, in an informal setting, i.e., when doing their coursework on their own, students are likely to rely on standard "general purpose" LLMs and not specific tutor models. Lastly, we can see LLM assistants that are, if at all, only prompt-moderated, being actively deployed in formal educational institutions (AILeap, 2025; OneTutor, 2025; fobizz, 2025; CSUN, 2025; Kabir, 2025; CNA, 2024).

The participants were randomly assigned to the groups. This mitigates differences in individual proficiency levels and, together with the *review* phase, allows for comparability between the cohorts. The *NON* cohort did not have any LLM assistance. They were allowed to use the material from their coursework during the *exercise* phase only. The present supervisor did not actively teach them or work out the assignments with the group, as this would have falsified any results of their performance in the *exercise*. The *NON* group thus mimics a standard exam-preparation scenario from the pre-LLM-era.

The second group, *GPT*, was given access to the LLM[2] chosen as the assistant. We utilized this model due to its accessibility via an API, widespread usage, and its low cost. These two factors are essential if LLMs were to be used in formal education as teaching support. The last group, *TUT*, had access to the same model as *GPT*, but with a different system prompt. In this prompt, it was specified that the model should not provide the solution to the mathematical questions but rather act as a tutor and guide the user to arrive at the solution himself. In contrast to Bastani et al. (2024), we did not incorporate any problem-specific guidance steps or problem solutions into our system prompt, making it more flexible and reducing the time required to set up the tutoring system for different tasks. The system prompts for both groups are shown in full in Appendix B. Fig. 2 shows a comparison between a conversation with the unmoderated and the moderated LLM.

The LLM assisted groups were given a short introduction on how to use the chatbot and tips on how to input mathematical symbols. However, no mathematical questions were answered during the *exercise*. All groups were supervised during the

[2] gpt-4o-mini-2024-07-18

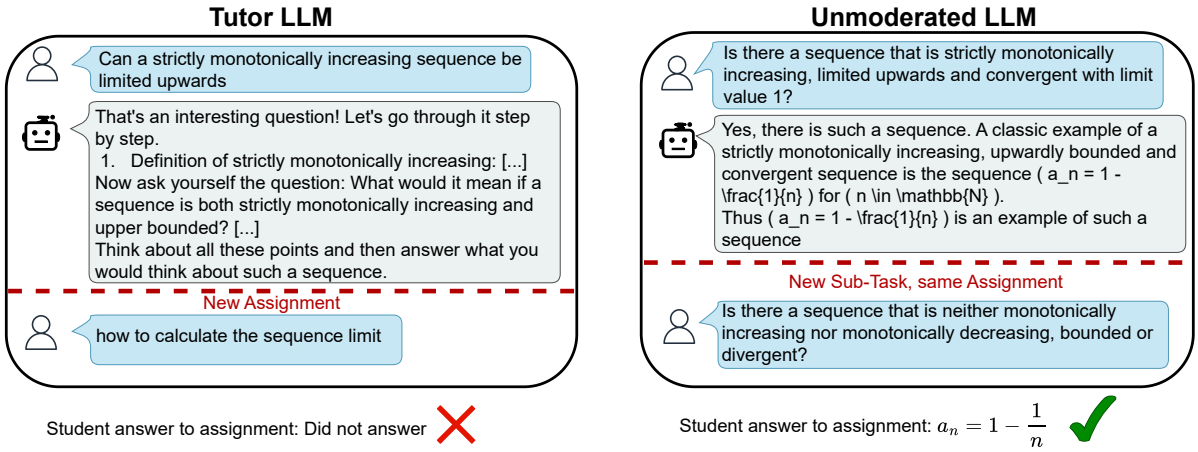

Figure 2: Comparison of a dialogue with the Tutor LLM (*left*) and the unmoderated LLM (*right*).

exercise and the *exam* to ensure adherence to the experimental design. During the *exam*, no group was allowed any assistance.

3.2 Mathematical questions

We focused on mathematical topics from the first semester that are typically challenging for students, since (i) the first semester content is the basis for many engineering and science study programs, (ii) mathematical questions allow for an objective evaluation of performance, and (iii) it has a high failure rate, stressing the motivation for innovative teaching.

The assignments for the *exercise* were designed in such a way that the students were not expected to be able to complete all assignments within the given time. This allows us to measure if LLM assistance increases the amount of worked-on tasks, which, in combination with the total score, can indicate an effect on time efficiency.

The students were given six tasks for both the *exercise* and the *exam*. The topics included sequences and limits, continuity, and DC. The types of assignments ranged from True/False statements, factual questions, calculation tasks, and argumentative questions to comprehension questions. The structure and content of the questions were designed to be analogous between *exercise* and *exam*. All questions are given in Appendix F and G.

We ensured that the utilized LLM can solve all assignments.

4 Method

In the following section, we describe the methods and interventions used in our experiment, including the setup of the two LLM-based tutoring systems and the reference group. We also outline the procedures for data collection and evaluation.

4.1 LLM as a Tutor

Effective LLM-based tutors are desirable from an educational point of view, as it is widely recognized that tutoring can significantly impact learning outcomes (Vail et al., 2016). They would allow more personalized learning without insurmountable human effort. Bastani et al. (2024) found that an LLM prompted to act as a tutor led to more effective student learning compared to the unmoderated LLM. However, their prompt includes the correct solution and detailed instructions on how to guide the students towards it. We deem this as not feasible for large-scale (real-world) scenarios because it requires human supervision in the form of creating the prompts for each task. Ideally, an LLM should be able to act as a tutor just by specifying this role in the prompt. We therefore use minimal prompt moderation that only adds the instruction of behaving like a tutor. The prompts can be found in the Appendix B.

4.2 Grading and Evaluation

Both the *exercise* and the *exam* were graded based on a uniform assessment standard by trained lecturers from the field of mathematics to determine the scores/marks. We define two metrics to measure student performance: the *score* and *completion*. The score simply measures the points a student achieved in a (sub-)task based on comparison with the correct solution of the assignment. The completion metric measures the level to which a student worked on an assignment. For this, we only consider whether a student had attempted to answer

Assignment	*GPT* (N=17)	*TUT* (N=15)
A1 (Sequences)	0.82 ± 0.25	0.57 ± 0.42
A2 (Limits)	0.76 ± 0.31	0.50 ± 0.34
A3 (T/F various)	0.76 ± 0.26	0.40 ± 0.34
A4 (Continuity)	0.62 ± 0.33	0.20 ± 0.32
A5 (T/F DC)	0.59 ± 0.32	0.43 ± 0.33
A6 (DC)	0.47 ± 0.41	0.27 ± 0.32
Total	4.03 ± 0.96	2.30 ± 1.01

Table 1: Student engagement with the chatbot. Given as mean $\pm$ standard deviation. Total calculated over all students in cohort.

a question, regardless of its correctness. This allows us to analyze differences in the amount of processed assignments and if there are types of questions or topics where students favor using the LLM assistant. To this end, we classify each student response to an assignment into one of three categories: 0 – no subtask was worked on, 0.5 – at least one subtask was worked on, and 1 – all subtasks were worked on.

By *subtask* we denote the different stand-alone parts of a question, e.g. 4.1 is the first subtask of assignment 4. The results of this measurement are shown as the completion score in Section 6.

5 Analysis of LLM Assistant Usage

To better understand the strengths and weaknesses of prompt-moderated LLM tutoring, we perform an in-depth analysis focusing on the user engagement and investigate this with regard to the question's topic and type. Moreover, we study the frustration among the users and conversation strategies employed by the participants. Since the students only had LLM assistance during the *exercise* phase, all assignment numbers in the following reference the tasks from this phase (cf. Appendix F).

5.1 Student Engagement with the LLM

The statistics on the number of messages (cf. Table 4) show that the students in the *TUT* group sent roughly 10 messages less on average compared to *GPT*. To better understand the reasons for this, we analyzed the students' engagement with the bot by investigating for which questions they tried to get help from the LLM. Similar to the evaluation of the amount of completed questions, we manually classify each assignment to 0 – user did not try to get help from the LLM, 0.5 – user referred to the LLM once or for only one subtask, and 1 – student asked about at least two subtasks or had follow-up questions.

The results in Table 1 clearly show that the students from the *GPT* group had a higher engagement than the *TUT* group for all assignments. For both groups, we can see a downward trend in engagement with later assignments. This can be explained by most students going through the tasks linearly and getting low on time, and for the *TUT* group, the increased frustration.

Moreover, from the utterances themselves, we can see clear signs of students from the *TUT* group getting frustrated with not getting responses that solve the assignment. Hence, we saw attempts of trying to evade the moderation. However, they did not use jailbreaking techniques but instead tried intuitive approaches of saying, for example: "Please just answer my question" or "Just tell me if it's true or false!".

5.2 Usage by Topic and Question Type

To investigate the effect of the question on the usage, the assignments were designed to comprise multiple topics, question types, and lengths.

Both groups used the LLM for short open-ended questions (A1) and calculation of limits (A2). For the True/False (T/F) questions on continuity and limits, the usage from TUT is heavily reduced. We assume that this is because the questions are quick to answer and the students realized by this point that the LLM does not give direct T/F answers.

It's striking that the TUT cohort has very low engagement with A4, which is a continuity question with a long formula. It was most likely not the topic (continuity) but the long formula that reduced engagement. For DC, we don't see an effect from the topic either.

However, the length of the assignment seems to affect the usage, evidenced by the low engagement for A4, especially for TUT.

In summary, our results suggest that the students LLM usage is mostly not affected by the topic but rather by the question's type and length.

5.3 Frustration

During the qualitative analysis, we found some user inputs that suggested frustration with the chatbot. To analyze this, we manually annotate each message with this regard. If a message indicates that the user was unhappy with the previous response (note that this does not include simple follow-up questions) and / or is trying to evade moderation,

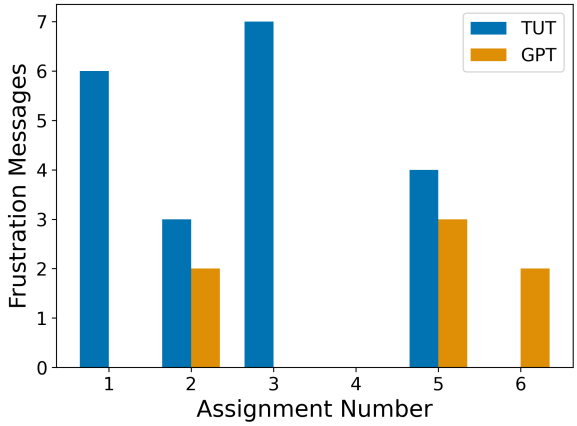

Figure 3: Visualization of the amount of messages signaling frustration per assignment.

Type	*GPT* (N=17)	*TUT* (N=15)
other	4.43 %	0.59 %
chit-chat	6.64 %	2.35 %
direct	66.05 %	68.24 %
abstract	13.28 %	17.06 %
malicious	2.58 %	0 %
follow-up	7.01 %	11.76 %

Table 2: Fraction of each user input type (strategy) for both LLM assisted cohorts in percent.

we label the message as signaling frustration. We only consider messages that are also in relation to the assignments.

While in the *GPT* conversations, only 2.58% of the messages signal frustration, this is true for 11.76% in the *TUT* group. The frustration per assignment is shown in Fig. 3. It is striking that *TUT* saw the highest frustration for A1 and A3. A1 has very short questions, and A3 is additionally a T/F type. This suggests that students expected a clear answer, especially for these seemingly easy questions. In both groups, roughly half of the students showed at least one sign of frustration. However, in the *GPT* cohort, each student has at most one message signaling frustration, while for *TUT* the maximum is six and the minimum is two. This highlights that the scaffolding does not always lead to frustration. Instead, it suggests differences in individual perception. We saw no clear link between frustration level and score.

5.4 Conversation Strategies

The qualitative analysis of the conversations allowed us to identify certain strategies employed by the students. The simple strategy of (mostly) copying a task description does appear in our data. However, in contrast to Krupp et al. (2024), it is not a common approach. This is most likely due to the assignments being printed out, making copy & pasting impossible, and students instead mostly opted to shorten questions. This paraphrasing already incites cognitive work. A frequent type of interaction in both groups is the student asking for mathematical knowledge abstracted from the specific task. These inquiries include, e.g., "How do I test a function for continuity?" and "Please tell me everything about differentiability". On the other hand, attempts to employ the LLM to directly solve the task or parts of it are frequent in both cohorts as well. A striking example is that multiple students ask if the function x^x is continuous. Moreover, there were multiple questions for the value of e^6, which is an intermediate result of one task, exhibiting the attempt to use the LLM as a calculator. Based on these observations, we manually classify all utterances into chit-chat, direct, abstract, follow-up, malicious, and other. The taxonomy is explained in the Appendix C.

The statistics of the different message types are depicted in Table 2. We can see that both groups behaved comparably, with input trying to directly answer an assignment making up roughly two out of three messages. Albeit the statistics show that the *TUT* group had more follow-up messages, this is not wholly comparable, since they include messages trying to escape the moderation.

Lastly, one student tested a malicious attack by trying to convince the bot to execute `sudo rm -fr /*`. This example shows that students may try to circumvent safety guards or try to cause other malicious behavior, stressing the importance of safety mechanisms when employing LLM-based tutors in real-world scenarios.

6 Analysis of Student Performance and Engagement

In the following, we will discuss the effect of (prompt-moderated) LLM tutoring on student performance in our study. First, we consider the student performance and engagement with the chatbot as a whole, then on an assignment-level, and lastly, we give special consideration to transfer tasks.

We measure the student performance with regard to their point score as well as the number of tasks completed. The results in Table 3 show that during the *exercise* phase, the *NON* and *TUT*

	NON (N=17)	*GPT* (N=17)	*TUT* (N=15)
Score *exercise*	39.7 ± 15.1	46.3 ± 12.9	36.8 ± 18.0
Score *exam*	24.3 ± 15.6	38.2 ± 17.2	29.5 ± 21.9
Compl. *exercise*	81.4 ± 12.0	86.3 ± 11.4	75.0 ± 18.6
Compl. *exam*	73.5 ± 16.5	85.3 ± 17.3	69.4 ± 31.4

Table 3: Student performance in percent as mean $\pm$ standard deviation. Score denotes the points they received during grading, and Completion (Compl.) the extent to which they worked on the assignments. Both are shown as percentage of the maximum. A visualization is given in Fig. 4 in Appendix E and the results for each individual assignment can be found in Table 5.

groups perform similarly, while the *GPT* cohort achieved a higher score on average. Moreover, both cohorts that had LLM assistance during the *exercise* achieved a higher mean score on the unassisted *exam* than the group without any LLM assistance. The difference in score between the *exercise* and *exam* was highest for the *NON* group, while the drop-off for the assisted groups was milder and comparable between them. The difference between *exercise* and *exam* and the group comparison *GPT* vs. *NON* is statistically significant. More details are provided in Appendix D.

Lastly, the *TUT* group had the lowest mean completion rate of all cohorts, while the *GPT* group worked on the most tasks on average. This can be explained by the *TUT* group having longer conversations, and due to this process taking longer to get to the solution.

6.1 Assignment-Level Evaluation

The assignment-level evaluation in Table 5 in Appendix A allows for insights with regard to the type of assignment. First, it is evident that LLM-assisted cohorts performed remarkably better on the first assignment during both phases. This task (cf. Appendix F) is fully formulated in natural language, resulting in a low entry barrier for LLM assistance. On the other hand, assignment A4 has the most complex math formula within the *exercise*. Here, *GPT* achieved the best results, while *TUT* was remarkably worse than *NON*. However, *TUT* has a much lower engagement for A4 than *GPT* (0.2 vs. 0.62). The long formula might have been discouraging the students from even trying to enter it, especially if they got frustrated by the scaffolding from previous questions.

During the *exercise*, the *NON* group achieved higher scores than both LLM groups for questions A2 and A6, which require calculations. Additionally, for A5, which, like A5, treats DC, the assisted groups are only slightly better than *NON*. These results suggest that it the assistant was less helpful for tasks that require explicit calculation or that treat DC.

The third assignment during the *exercise* consisted of True/False questions on various topics. This is the only task that all students tried to answer completely. The LLM assisted cohorts achieved a higher score on average and had high engagement with the assistant. A visualization of the completion can be found in Fig. 5 in Appendix E.

In summary, these results show varied effects of the LLM assistance based on the question type and topic. In combination with the engagement analysis, we conclude that the students used the LLM even for tasks where it proved not beneficial. But at the same time there are also assignments where the assistance greatly improved their performance.

Next, we investigate how these results translate to the *exam* phase, where no assistance was given to any group. Here, we observe that the *GPT* group achieved higher average scores than *NON* on every assignment and higher than *TUT* for every assignment but A5. In total, both LLM assisted groups achieved a higher average score on the exam than *NON*. This indicates that LLM assistance, even when not optimal, can still have a beneficial effect on learning *compared to no assistance*, and contradicting concerns regarding shallow learning.

6.2 Performance on Transfer Tasks

It is a natural assumption that LLM assistance might be used as a crutch, provoking shallow learning, as cautioned by Bastani et al. (2024). We argue that this would be most evident for transfer tasks, where one needs to apply previously acquired knowledge to new situations. Therefore, the performance on transfer tasks with and without the assistance is of special interest and will be analyzed below.

The assignments A1, A3, and A5 from the *exercise* and A1, A4, and A6 from the *exam* phase were designed as transfer tasks. From the performance on these (cf. Table 5), we can see that during the *exercise*, the LLM-assisted groups perform remarkably better than NON for A1 and, while still better, less so for A3 and especially A5. These assignments have different question types. A1 is an open-ended task, while A3 consists of T/F questions (50% chance to be correct randomly) and A5 is T/F plus an explanation.

Metric	*GPT* (N=17)	*TUT* (N=15)
# Messages*	31.9 ± 11.6	22.7 ± 9.5
# Messages total	542	340
Message Length user*	59.6 ± 46.5	55.7 ± 45.3
Message Length LLM*	877.6 ± 722.57	1370.5 ± 638
Time delay user* (s)	157 ± 39	222 ± 104

Table 4: GMADE dataset statistics. The message length is given in number of characters. The time delay between two user messages in seconds. *: mean $\pm$ standard deviation.

For the exam phase, the previously assisted participants clearly outperform the *NON* group for A1, and A6, but not for A4. A1 here is similar to the *exercise*, and A6 asks the student to provide a reason why certain statements about DC are wrong. The information the LLM has provided during the *exercise* has likely helped the students to answer this assignment.

Thus, our results indicate that the assumption that students with LLM assistance would perform worse on transfer tasks when unassisted, due to using the LLM as a crutch, is not confirmed.

7 The German Math Assistant Dataset for Education

We publish the annotated conversations between the students and the LLM assistants as GMADE to foster future research into tutoring LLMs. It includes 32 multi-turn German conversations between students and GPT-4o-mini that occurred over the span of 40 minutes, accumulating in roughly 900 messages. The students interact with the LLM to solve a series of mathematical questions.

To the best of our knowledge, GMADE is the first dataset of its kind and could be useful for studies that currently rely on fully synthetic data (e.g., Macina et al., 2023a; Chevalier et al., 2024).

Each user message is annotated as either chit-chat, direct, abstract, malicious, follow-up, or other. Additionally, the annotation includes a binary field marking if the message suggests frustration and a mapping to the assignment number (if applicable). Lastly, we provide the timestamps at which the message was sent, allowing for analysis of processing time.

The statistics of our dataset are presented in Table 4. This reveals that the participants in the *GPT* cohort sent, on average, 10 messages more than their tutored counterparts. Additionally, we can see that while the user messages are of comparable length, the tutor LLM responses are roughly 50% longer. Consequently, the mean time delay between two user messages is about one minute longer in the *TUT* cohort than in the *GPT* group. This can be explained by the longer response needing more time to read and comprehend. Given that the *exercise* was timed, this also limits the number of messages the participants are able to send.

GMADE shows that the unmoderated LLM did not only answer the user's question but also, unsolicitedly, provided additional information such as definitions, examples, and conclusions. This can be attributed to the LLM being trained as a helpful assistant and humans preferring the more detailed responses during Reinforcement Learning from Human Feedback (Christiano et al., 2017).

8 Conclusion

This study investigates the impact of pretrained LLMs as learning assistants in a challenging mathematical QS scenario. The questions are on the level of tertiary education, span multiple topics, and test different competencies. We compare the effect of students being provided with an unmoderated LLM, a prompt-moderated LLM, and no assistance in a three-phase experiment.

Our results indicate that both during the assisted *exercise* phase and the unassisted *exam*, the cohort using the unmoderated LLM achieved the highest score on average. Moreover, on average, they also had the largest proportion of tasks worked on during both phases.

While the tutoring group had a lower mean score in the *exercise* than *NON*, their average score in the *exam* was higher, even though they had worked on fewer assignments.

One possible reason why the *TUT* group performed worse in total than *GPT* is information overload caused by the long responses and frustration stemming from the tutor refusing to answer the question directly, which was similarly observed by Krupp et al. (2023).

The investigation into the student engagement with the assistants shows that the usage is not governed by LLM helpfulness or topic, but by question length and how easy it is to enter it into the chat. However, when considering the actual effect the engagement had on the QS, we see that this varies by question type and topic.

A more detailed analysis of the conversations reveals that the students utilizing the tutor LLM got frustrated with its evasive responses and tried

to extract a direct answer. Additionally, they received longer responses, which increased the time between two user messages and probably led to them working on fewer assignments than the other two groups during the *exercise*.

This indicates that LLM assistance, even when not optimal, can still have a beneficial effect on learning *compared to no assistance*, and refutes concerns regarding shallow learning.

To conclude, the LLM assistance had a positive effect on the students performance, even when unassisted. This suggests that concerns of shallow learning might be exaggerated and stresses the need for further studies into LLM tutors. However, prompt-moderation with the intention of having the LLM act as a tutor, could actually be harmful to the learning effect.

Given our results, we believe future work can focus on how prompt-moderation can become beneficial, how to best enhance traditional lectures with LLM assistance and the role of the subject or domain in this context.

9 Limitations

While our study reveals interesting notions about the efficiency of LLM assistance on math learning, it also has some limitations that we needed to accept to make it feasible. First, the focus of our study was not to investigate how the students perceive the chatbot, since their general technology acceptance and perception of ease-of-use are well-established (Yilmaz et al., 2023; Chan et al., 2024; Strzelecki, 2024; Shahzad et al., 2024), and it has been shown that students are mostly positive towards computer-based learning, especially when used as a supplement to traditional lectures (Dewhurst et al., 2000). Instead, we limited our investigation to the actual effect on students' performance.

Importantly, our results should not be interpreted as a comparison between LLM tutoring and human tutoring but instead between LLM tutoring and no tutoring.

While we stayed close to the study design described by Bastani et al. (2024), one important difference is the system prompt. We do not include any problem-specific guidance steps or solutions, which requires considerably less time to set up and is, in our opinion, a more realistic setting. This, however, probably also impacts the behavior of the tutoring LLM and thus the reception (and maybe performance) by the students.

Due to the elaborate setting, the collection of dialogues is expensive, leading to GMADE having a limited number of dialogues. Moreover, GMADE does not include human tutoring utterances and can thus not be directly used for the training of LLM tutors.

Since the students with unmoderated LLM access were able to work on more problems, as evidenced by the higher completion score, they could have a slightly higher task familiarity, which might lead to improved performance when unassisted. While this effect cannot be avoided, we still deem the comparison as fair, as such would be an advantage of the assistance they received.

Our study imposes a strict time limit on the students preparation, i.e., the *exercise* phase, for the *exam* phase. This is required to have comparable results and a realistic time frame for the study. We argue that students also experience time pressure when preparing for real exams, due to the workload of preparing for multiple exams in a short period of time. Nevertheless, it is possible that the results would be different if students were given as much time as they liked to prepare with their designated assistance for an unassisted exam.

Another limitation arises from the fact that the Tutor LLM did not perfectly follow its instruction to refrain from giving away solutions. When a student, e.g., asked for "a divergent limited sequence", the response included an example that was a correct solution to the first task. In at least one case, the Tutor LLM also gave away the full (and correct) solution to a task.

10 Ethical considerations

All participants in our study were volunteers that agreed to donate their data for research. All participants were informed about how their data was going to be used and actively agreed to it before taking part in the study. Their performance did not influence their semester's grade, which was determined by an additional, standard exam.

The GMADE dataset was carefully reviewed to not contain any personal information before publication. Through the usage of the API, it was ensured that the generated input will not be used to train future LLMs, acting as an additional security mechanism in case a participant entered personal data.

We used a website wrapper around the ChatGPT

API as the user interface for the groups *GPT* and *TUT*. This allows us to easily collect the conversations and ensures GDPR compliance. The usage of the GPT-4o-mini via the API is convincing due to its ease-of-use and low price-point. For approximately 900 messages (both user and assistant), the cost was only 0.21 euros. This could allow educational institutions such as high schools and universities to offer AI tutors to their students.

References

AILeap. 2025. AILeap. https://aileap.ee/en. Accessed: 2025-04-29.

Julia Anghileri. 2006. Scaffolding practices that enhance mathematics learning. *Journal of Mathematics Teacher Education*, 9(1):33–52.

Hamsa Bastani, Osbert Bastani, Alp Sungu, Haosen Ge, Ozge Kabakcı, and Rei Mariman. 2024. Generative AI can Harm Learning. *Available at SSRN*, 4895486.

Shiau Wei Chan, Nur Intan Shahira Norhisham, Fadillah Ismail, and Md Fauzi Ahmad. 2024. Students' Perceptions and Intentions Regarding ChatGPT Usage in Higher Education. In *Proceedings of the 2024 10th International Conference on E-Society, e-Learning and e-Technologies (ICSLT)*, ICSLT '24, pages 49–54, New York, NY, USA. Association for Computing Machinery.

Alexis Chevalier, Jiayi Geng, Alexander Wettig, Howard Chen, Sebastian Mizera, Toni Annala, Max Jameson Aragon, Arturo Rodríguez Fanlo, Simon Frieder, Simon Machado, Akshara Prabhakar, Ellie Thieu, Jiachen T. Wang, Zirui Wang, Xindi Wu, Mengzhou Xia, Wenhan Xia, Jiatong Yu, Jun-Jie Zhu, Zhiyong Jason Ren, Sanjeev Arora, and Danqi Chen. 2024. Language Models as Science Tutors. *Preprint*, arXiv:2402.11111.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30:4299–4307.

CNA. 2024. The project in Denmark introduces artificial intelligence into schools. https://www.cna.al/english/tech/projekti-ne-\protect\penalty\z@danimarke-fut-ne-shkolla-inteligjencen-\protect\penalty\z@artificiale-i387221. Accessed: 2025-04-29.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

CSUN. 2025. ChatGPT Edu (IT) | CSU Northridge. https://www.csun.edu/it/software-services/chatgpt. Accessed: 2025-04-29.

David G Dewhurst, Hamish A Macleod, and Tracey A. M Norris. 2000. Independent student learning aided by computers: An acceptable alternative to lectures? *Computers & Education*, 35(3):223–241.

Ronald A. Fisher. 1919. The correlation between relatives on the supposition of Mendelian inheritance. *Earth and Environmental Science Transactions of the Royal Society of Edinburgh*, 52(2):399–433.

fobizz. 2025. fobizz. https://fobizz.com/en/. Accessed: 2025-09-13.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*.

Toru Ishida, Aditya Firman Ihsan, Rikman Aherliwan Rudawan, et al. 2024. Advancing global south university education with large language models. *arXiv preprint arXiv:2410.07139*.

Omer Kabir. 2025. AI in every classroom? Inside Israel's ambitious education overhaul. https://www.calcalistech.com/ctechnews/article/lde52kasn. Accessed: 2025-04-29.

Lars Krupp, Steffen Steinert, Maximilian Kiefer-Emmanouilidis, Karina E. Avila, Paul Lukowicz, Jochen Kuhn, Stefan Küchemann, and Jakob Karolus. 2023. Challenges and Opportunities of Moderating Usage of Large Language Models in Education. *Preprint*, arXiv:2312.14969.

Lars Krupp, Steffen Steinert, Maximilian Kiefer-Emmanouilidis, Karina E. Avila, Paul Lukowicz, Jochen Kuhn, Stefan Küchemann, and Jakob Karolus. 2024. Unreflected acceptance–investigating the negative consequences of chatgpt-assisted problem solving in physics education. In *HHAI 2024: Hybrid Human AI Systems for the Social Good*, pages 199–212. IOS Press.

William H Kruskal and W Allen Wallis. 1952. Use of ranks in one-criterion variance analysis. *Journal of the American statistical Association*, 47(260):583–621.

Harsh Kumar, David M. Rothschild, Daniel G. Goldstein, and Jake M. Hofman. 2023. Math Education with Large Language Models: Peril or Promise? *Preprint*, Social Science Research Network:4641653.

Howard Levene. 1960. Robust tests for equality of variances. In *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*, pages 278–292. Stanford University Press.

Jakub Macina, Nico Daheim, Sankalan Chowdhury, Tanmay Sinha, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. 2023a. MathDial: A dialogue tutoring dataset with rich pedagogical properties grounded in math reasoning problems. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5602–5621, Singapore. Association for Computational Linguistics.

Jakub Macina, Nico Daheim, Sankalan Pal Chowdhury, Tanmay Sinha, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. 2023b. MathDial: A Dialogue Tutoring Dataset with Rich Pedagogical Properties Grounded in Math Reasoning Problems. *Preprint*, arXiv:2305.14536.

Kaushal Kumar Maurya, Kv Aditya Srivatsa, Kseniia Petukhova, and Ekaterina Kochmar. 2025. Unifying AI tutor evaluation: An evaluation taxonomy for pedagogical ability assessment of LLM-powered AI tutors. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1234–1251, Albuquerque, New Mexico. Association for Computational Linguistics.

OneTutor. 2025. OneTutor. `https://onetutor.ai/`. Accessed: 2025-04-29.

OpenAI. 2022. OpenAI: Introducing ChatGPT.

Emily Ross, Yuval Kansal, Jake Renzella, Alexandra Vassar, and Andrew Taylor. 2025. Supervised fine-tuning llms to behave as pedagogical agents in programming education. *arXiv preprint arXiv:2502.20527*.

Muhammad Farrukh Shahzad, Shuo Xu, and Iqra Javed. 2024. ChatGPT awareness, acceptance, and adoption in higher education: The role of trust as a cornerstone. *International Journal of Educational Technology in Higher Education*, 21(1):46.

Samuel Sanford Shapiro and Martin B. Wilk. 1965. An analysis of variance test for normality (complete samples). *Biometrika*, 52(3-4):591–611.

Shashank Sonkar, Kangqi Ni, Sapana Chaudhary, and Richard Baraniuk. 2024. Pedagogical alignment of large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13641–13650, Miami, Florida, USA. Association for Computational Linguistics.

Artur Strzelecki. 2024. Students' Acceptance of ChatGPT in Higher Education: An Extended Unified Theory of Acceptance and Use of Technology. *Innovative Higher Education*, 49(2):223–245.

AnaÃ¯s Tack and Chris Piech. 2022. The AI teacher test: Measuring the pedagogical ability of blender and GPT-3 in educational dialogues. In *Proceedings of the 15th International Conference on Educational Data Mining*, pages 522–529, Durham, United Kingdom. International Educational Data Mining Society.

LearnLM Team, Abhinit Modi, Aditya Srikanth Veerubhotla, Aliya Rysbek, Andrea Huber, Brett Wiltshire, Brian Veprek, Daniel Gillick, Daniel Kasenberg, Derek Ahmed, et al. 2024. Learnlm: Improving gemini for learning. *arXiv preprint arXiv:2412.16429*.

John W. Tukey. 1949. Comparing individual means in the analysis of variance. *Biometrics*, pages 99–114.

Alexandria K Vail, Joseph B Wiggins, Joseph F Grafsgaard, Kristy Elizabeth Boyer, Eric N Wiebe, and James C Lester. 2016. The affective impact of tutor questions: Predicting frustration and engagement. *International Educational Data Mining Society*.

Shen Wang, Tianlong Xu, Hang Li, Chaoli Zhang, Joleen Liang, Jiliang Tang, Philip S. Yu, and Qingsong Wen. 2024. Large Language Models for Education: A Survey and Outlook. *Preprint*, arXiv:2403.18105.

Colin White, Samuel Dooley, Manley Roberts, Arka Pal, Ben Feuer, Siddhartha Jain, Ravid Shwartz-Ziv, Neel Jain, Khalid Saifullah, Siddartha Naidu, et al. 2024. Livebench: A Challenging, Contamination-free LLM Benchmark. *Preprint, arXiv:2406.19314*.

Vinzenz Wolf and Christian Maier. 2024. ChatGPT usage in everyday life: A motivation-theoretic mixed-methods study. *International Journal of Information Management*, 79:102821.

David Wood, Jerome S. Bruner, and Gail Ross. 1976. The role of tutoring in problem solving. *Journal of child psychology and psychiatry*, 17(2):89–100.

Yi Wu. 2023. Integrating generative AI in education: How ChatGPT brings challenges for future learning and teaching. *Journal of Advanced Research in Education*, 2(4):6–10.

Halit Yilmaz, Samat Maxutov, Azatzhan Baitekov, and Nuri Balta. 2023. Student Attitudes towards Chat GPT: A Technology Acceptance Model Survey. *International Educational Review*, 1(1):57–83.

A Results per assignment

The results per assignment are shown in Tab. 5.

B Full System Prompts

B.1 System Prompt for *GPT*

You are a friendly and concise assistant. Provide clear and direct answers. Always answer in the language you have been addressed in. If your response includes math, use Latex notation.

B.2 System Prompt for *TUT*

You are an educational tutor who helps the user improve his understanding of mathematical concepts. Provide detailed and helpful answers to the user's queries. As a tutor, you are not allowed to answer

	NON (N=17)		*GPT* (N=17)		*TUT* (N=15)	
Assignment	Score	Completion	Score	Completion	Score	Completion
A1	23.5 ± 35.9	73.5 ± 35.9	77.9 ± 35.2	100.0 ± 0	70.0 ± 25.4	83.3 ± 24.4
A2	45.1 ± 27.3	88.2 ± 28.1	40.2 ± 23.6	97.1 ± 12.1	40.0 ± 31.4	83.3 ± 36.2
A3	52.9 ± 24.8	100.0 ± 0	67.6 ± 26.2	100.0 ± 0	70.0 ± 23.5	100.0 ± 0
A4	52.2 ± 24.8	91.2 ± 19.6	61.4 ± 24.0	91.2 ± 19.6	26.3 ± 38.5	60.0 ± 43.1
A5	15.7 ± 23.9	55.9 ± 16.6	17.6 ± 20.8	76.5 ± 31.2	20.6 ± 28.3	66.7 ± 30.9
A6	37.6 ± 23.1	79.4 ± 30.9	34.1 ± 32.4	52.9 ± 45.0	29.3 ± 28.9	56.7 ± 45.8
A1	11.8 ± 21.9	64.7 ± 34.3	41.2 ± 40.4	91.2 ± 26.4	36.7 ± 44.2	73.3 ± 41.7
A2	31.8 ± 30.0	88.2 ± 21.9	45.9 ± 29.8	91.2 ± 19.6	26.7 ± 38.3	76.7 ± 37.2
A3	23.5 ± 43.7	82.4 ± 39.3	47.1 ± 49.1	100.0 ± 0	23.3 ± 41.7	60.0 ± 50.7
A4	42.6 ± 27.0	88.2 ± 21.9	42.6 ± 25.0	85.3 ± 29.4	47.5 ± 28.8	73.3 ± 41.7
A5	11.0 ± 10.3	70.6 ± 30.9	15.1 ± 14.2	64.7 ± 29.4	14.2 ± 18.4	66.7 ± 40.8
A6	16.2 ± 26.4	47.1 ± 37.4	50.0 ± 33.9	79.4 ± 30.9	30.0 ± 29.0	66.7 ± 30.9

Table 5: Student performance per assignment in percent as mean $\pm$ standard deviation. Grey background denotes *exercise* phase.

the user's mathematical question directly. Instead, you should guide him towards finding the solution himself in incremental steps. Always answer in the language you have been addressed in. If your response includes math, use Latex notation.

C Taxonomy of User Messages

The taxonomy that was used to classify the user messages is shown in Tab. 6.

D Statistical Significance

We evaluate if there are statistically significant differences between the groups for both the score and the completion. The Shapiro-Wilk test (Shapiro and Wilk, 1965) showed that the performance scores follow a normal distribution ($p = 0.37$, $p = 0.14$, $p = 0.38$ for *NON*, *GPT*, and *TUT*, respectively). Moreover, the homogeneity of variances could be established with the Levene test (Levene, 1960) ($p = 0.61$ and $p = 0.42$ for *exercise* and *exam*, respectively). Given this, we performed a two-way ANOVA (Fisher, 1919) test. This revealed that there is a significant ($p < 0.01$) difference between the scores during the *exercise* and *exam*, and that type of assistance had an influence on the scores ($p = 0.03$). However, there was no significant ($p = 0.57$) interaction between the level of assistance and the difference in score between the *exercise* and *exam* phase (cf. Table 3). We further evaluate the statistical significance of the difference in score between the groups with the Tukey's HSD test (Tukey, 1949). The only group comparison that had significant differences was *NON* vs. *GPT* ($p = 0.046$, $CI = [0.162, 20.38]$). For *NON* vs. *TUT* ($p = 0.96$, $CI = [-9.30, 11.58]$) and *GPT* vs. *TUT* ($p = 0.099$, $CI = [-19.58, 1.31]$) no statistical significant difference was observed.

The completion rate is not normally distributed. We thus use the Kruskal-Wallis-Test (Kruskal and Wallis, 1952), which revealed that there is no significant ($p = 0.16$. $p = 0.08$ for *exercise* and *exam*, respectively) difference in completion between the cohorts.

E Visualizations

Figures 5 and 4 visualize the performance and completion percentage.

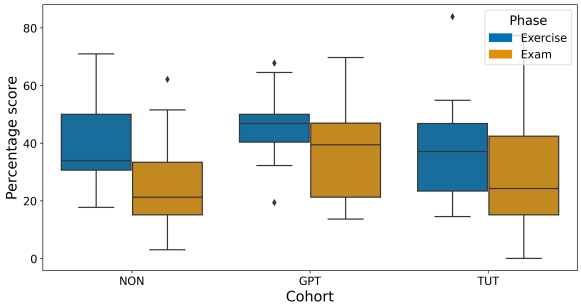

Figure 4: Boxplot of the performance as measured by the percentage score.

Name	Explanation
other	Messages that don't fit any other category.
chit-chat	Messages not related to the assignments, but fitting a chit-chat conversation, where the chatbot is supposed to keep the conversation going. Includes for example salutations.
direct	The user attempted to directly solve an assignment with his message. Includes for examples directly entering an assignment or trying to use the chatbot as a calculator.
abstract	The user attempted to get more indirect, abstract knowledge from the chatbot, expecting the response to help him solve an assignment. Includes for example asking for the rules of differentiation.
malicious	The user entered malicious input, that should be deflected by the Bot. Only one instance of this was found.
follow-up	The user message is a follow-up to his previous input. Could be a question, but also a clarification. Includes for example the user asking "and concretely?" or providing the domain of definition.

Table 6: Taxonomy used to classify the user messages.

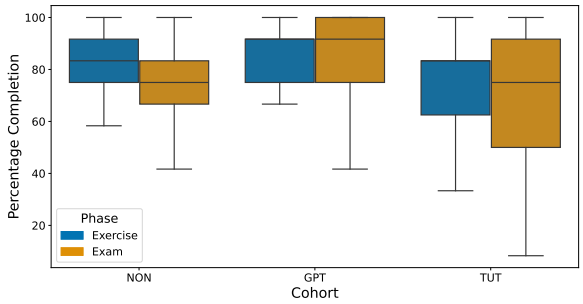

Figure 5: Boxplot of the completion percentage.

F Assignments for the *exercise*

A1: Sequences Provide, if possible, a sequence for each case that satisfies the given properties. If no such sequence exists, briefly justify your answer. Find a sequence that:

1. is strictly monotonically increasing, bounded above, and convergent with the limit 1.
2. is neither monotonically increasing nor monotonically decreasing, bounded, and divergent.

A2: Sequence Limits Calculate the following (possibly improper) sequence limits

1. $\lim\limits_{n\to\infty} \frac{n^2(n+\sin(n))}{4n^3+2n+1}$
2. $\lim\limits_{n\to\infty} \sqrt{2n} - \sqrt{n+1}$

A3: True or False?

1. If a function is differentiable at x_0, then it is also continuous at that point.
2. If $f : \mathbb{R} \to \mathbb{R}$ is continuous at x_0, then $\lim\limits_{x\to x_0} f(x) = f\left(\lim\limits_{x\to x_0} x\right)$ holds.
3. If the right-hand and left-hand limits of a function agree at a point x_0, then f is continuous at x_0.
4. The function $g : \mathbb{R}^+ \to \mathbb{R}$ with $g(x) = x^x$ is continuous.

A4: Continuity of Functions We consider the function $f : [-1, \infty) \to \mathbb{R}$ with

$$f(x) = \begin{cases} e^{2x} - x - 2, & x \in [-1, 3], \\ \frac{x^2-x-6}{x-3}, & x > 3. \end{cases}$$

1. Check whether the function is continuous at $[-1, \infty)$. In particular, examine the point $x_0 = 3$.
2. Does f have a zero in the interval $[-1, 3]$? Justify your answer without explicitly calculating it.
   Use the estimate $2 < e < 3$ and no (!) calculator.

A5: Differential Calculus Decide whether the following statements are true or false. Give a short justification or counterexample for each.

1. The sum $f+g$ of a non-differentiable function f and a differentiable function g can result in a differentiable function.
2. If an inverse function $f : \mathbb{R} \to \mathbb{R}$ is differentiable, then the inverse function $f^{-1} : \mathbb{R} \to \mathbb{R}$ is also differentiable.

3. If $x_0 \in [a, b]$ is a (local) minimum of a function $f : [a, b] \to \mathbb{R}$, then f is differentiable in x_0 and $f'(x_0) = 0$.

A6: Differential Calculus The continuous function $f : [0, 1] \to \mathbb{R}$ is given as follows:

$$f(x) = \begin{cases} \sqrt{x} \ln x & \text{for } x > 0, \\ 0 & \text{for } x = 0. \end{cases}$$

1. Using a suitable difference quotient, show that f at $x = 0$ has no right-hand side derivative.
2. Determine the derivative f' of f for $x \in (0, 1)$.

G Assignments for the *exam*

A1: Sequences If possible, specify a sequence that has the properties mentioned. If no such sequence exists, give a brief explanation. We are looking for a sequence that

1. is strictly monotonically decreasing, bounded and divergent.
2. is recursively defined, bounded and divergent.

A2: Sequence Limits

Calculate the following (possibly improper) sequence limits

1. $\lim\limits_{n\to\infty} e^{-n}((-1)^n + 3)$
2. $\lim\limits_{n\to\infty} \sqrt[n]{5 + 3 \cdot 2^n}$

A3: Continuity

Investigate whether the following function is continuous: $f : [-1, 1] \to \mathbb{R}$ with

$$f(x) = \begin{cases} e^{-\frac{1}{1-x^2}} & \text{for } x \in (-1, 1) \\ 0 & \text{for } x = -1. \end{cases}$$

A4: Continuity Answer the following questions and give reasons for your answer.

1. Can the sum of two discontinuous functions result in a continuous function?
2. Can the sum of two continuous functions result in a discontinuous function?
3. Is the function $f : \mathbb{R} \to \mathbb{R}$ with $f(x) = x^{\text{sign(x)}}$ continuous?

A5: Differential Calculus A function $g : [0, 1] \to \mathbb{R}$ is given as follows:

$$g(x) = \begin{cases} \sqrt{x} \ln(1 - x) & \text{for } x > 0, \\ 0 & \text{for } x = 0. \end{cases}$$

1. Using a suitable difference quotient, justify that g has a right-hand derivative at $x = 0$ and calculate it.
2. Specify the derivative of g.
3. Show that g has no critical points for $x \in (0, 1)$.

A6: Myths about Differential Calculus Give a suitable example to explain why the following statements are all false:

1. The concatenation of a non-differentiable function and a differentiable function cannot result in a differentiable function.
2. If $f : \mathbb{R} \to \mathbb{R}$ is differentiable and strictly monotonically increasing, then $f'(x) > 0$ must hold for all $x \in \mathbb{R}$.
3. If for a function f the limit values $\lim_{x\to x_0-} f'(x)$ and $\lim_{x\to x_0^+} f'(x)$ exist and coincide, then f is differentiable at the point x_0.