Jana von Dielingen[1], Tobias R. Rebholz[2] & Frank Papenmeier[3]

Empirical investigation of a GPT-4o mini-based tutor for the learning of R programming

Abstract

This study investigated the impact of an AI-based tutor using GPT-4o mini on R programming task outcomes and subjective evaluations among psychology students at the University of Tübingen. Students were divided into three groups: AI tutor, video tutorial, or both. Confirmatory analyses showed no significant differences in performance points and subjective evaluations across the three groups. Descriptive results and exploratory analyses suggest that our AI tutor improved subjective evaluations of the learning environment without affecting time on task or performance. We discuss the implications of our results for future research on the use of AI in higher education.

Keywords

artificial intelligence, higher education, GPT-4o mini, tutor, programming

1 Corresponding author; Eberhard Karls University of Tübingen, Germany; jana.von.dielingen@gmail.com; ORCID 0009-0004-2047-9423

2 Eberhard Karls University of Tübingen, Germany; tobias.rebholz@uni-tuebingen.de; ORCID 0000-0001-5436-0253

3 Eberhard Karls University of Tübingen, Germany; frank.papenmeier@uni-tuebingen.de; ORCID 0000-0001-5566-9658

https://doi.org/10.21240/zfhe/SH-KI-2/13

Empirische Untersuchung eines GPT-4o mini-basierten Tutors zum Erlenen von R Programmierung

Zusammenfassung

Diese Studie untersuchte den Einfluss eines KI-basierten Tutors unter Verwendung von GPT-4o mini auf die Lernergebnisse im R-Programmieren und die subjektive Einschätzung unter Psychologie-Studierenden an der Universität Tübingen. Die Studierenden wurden in drei Gruppen eingeteilt: KI-Tutor, Video-Tutorial oder beides. Die konfirmatorischen Analysen zeigten keine signifikanten Unterschiede in der Aufgabenleistung und den subjektiven Einschätzungen. Die deskriptiven Ergebnisse und explorativen Analysen deuten darauf hin, dass unser KI-Tutor zu verbesserten subjektiven Einschätzungen der Lernumgebung führte, ohne die Lernzeit oder Lernleistung zu beeinflussen. Wir diskutieren die Implikationen unserer Ergebnisse für zukünftige Forschung zum Einsatz von KI in der Hochschulbildung.

Schlüsselwörter

Künstliche Intelligenz, Hochschulbildung, GPT-4o mini, Tutor, Programmierung

1 Introduction

The rapid adoption of AI technologies in education is exemplified by the widespread use of ChatGPT with around 37% of university students turning to this tool for assistance with their assignments reflecting the growing reliance on AI (*4 in 10 College Students Are Using ChatGPT on Assignments*, 2024). The most recent innovation was the introduction of conversational, generative AI, which is based on Large Language Models (LLMs), such as the interactive chat interface of OpenAI's GPT-3. Digital teaching methods have been utilized and researched for a substantial period (e.g. Bilyalova et al., 2020; Paul et al., 2018). However, the introduction of LLM-based conversational AI tutors enables more personalized and scalable support, facilitating independent learning by using natural language processing and machine learning techniques to assess student responses and monitor progress through the analysis of individual learning patterns (Lin et al., 2023). Moreover, AI-driven tutoring systems play a crucial role by offering an innovative platform that enhances educational accessibility. This approach allows students to learn at their own pace and from any location, thereby increasing flexibility in the learning process. Nevertheless, there are also initial studies that highlight the risks of using AI-based tutors. For example, Bastani et al. (2024) demonstrated that a GPT-4-based tutor significantly enhanced the math performance of high school students. However, when access to the tutor was later removed, the students' performance declined to a level lower than that of those who had never used the AI tool, highlighting a risk of over-reliance on such tools.

In this evolving landscape, universities have the chance to reconsider their teaching methods and learning strategies to effectively incorporate these tools (Vargas-Murillo et al., 2023) and provide guidance for their use. Our approach is therefore to include and test an AI-based tutor based on the *GPT-4o mini* LLM in the learning of R programming, a statistics software (R Core Team, 2024), for undergraduate psychology students at the University of Tübingen. Specifically, we test whether the AI-based tutor influences the performance of students in this specific context, for which LLMs are particularly helpful and effective (e.g., Tian et al., 2023), compared to a

more conventional teaching approach using video tutorials. While some studies have explored the potential of AI in education (e.g. Cowen & Tabarrok, 2023; Zografos & Moussiades, 2023), empirical evidence on the impact of GPT-based tutors on actual learning outcomes in higher education remains scarce. Frankford et al. (2024) explored a GPT-3.5-based AI tutor in Artemis, focusing on personalized interaction during a Pascal's Triangle exercise. They analyzed user experiences and identified varying user types but found the AI's feedback effective only 66.6% of the time, often being vague, incorrect, or overly solution-focused. In contrast, Baillifard et al. (2024) reported that psychology students using their GPT-3.5-based app significantly outperformed peers in a Neuroscience exam. In addition to assessing actual performance, understanding students' subjective opinions about the use of our AI tutor is crucial. The perception of AI tools can strongly influence their adoption and integration into regular study habits, predicting their continued use (Isaac et al., 2019; Shaengchart, 2023). Therefore, our study also investigated how students evaluate our AI tutor using the following scales based on the findings of Isaac et al. (2019). They demonstrated that these factors influence students' intentions to use or their actual use of AI-based tutors.

1. *System Quality* refers to how strongly users perceive the system as user-friendly and easy to connect with.
2. *Information Quality* pertains to how users assess the information provided in online learning environments in terms of its accuracy, comprehensiveness and timeliness.
3. *Compatibility* refers to how well new innovations are perceived to fit with the existing needs and values of their users.
4. *User Satisfaction* is defined as the extent to which users find systems to be useful.
5. *Task-Technology Fit* is defined as how well systems align with the tasks at hand and meet specific requirements as well as the degree to which technologies support users in completing coursework or jobs.

6. *Performance Impact* refers to how system use enhances work quality by speeding up task completion, increasing job control, improving accuracy, and boosting overall efficiency.
7. *Future Usage* is a measurement of actual intention to use the learning environment again.

To the best of our knowledge, no study has so far examined the differences in students' perceptions of AI-based teaching methods versus video-based teaching methods. However, Kim et al. (2020) demonstrated that the perceived usefulness of AI-based teaching methods is rated highly by students. Consequently, we also hypothesized that the learning environment with the AI tutor would receive higher scores in subjective evaluations, also when AI tutor and video are both available (the combination performs as well as the better individual component). The literature presents mixed findings on Information Quality, with some studies indicating that people are equally skeptical of AI- and human-generated information (e.g. Buchanan & Hickman, 2024), while others suggest that AI-generated information is trusted less (e.g. McClain, 2024). We hypothesized that, compared to the conventional teaching methods using video tutorials, participants would place less trust in an AI tutor, leading to a lower rating for Information Quality in the groups where an AI tutor was involved, including the combined tutor + video group. Figure 1 shows all assessed scales and the corresponding hypotheses regarding our three experimental groups: AI-based tutor (henceforth: "AI tutor"), video-based teaching (henceforth: "video"), and a combination of both (henceforth: "AI tutor + video").

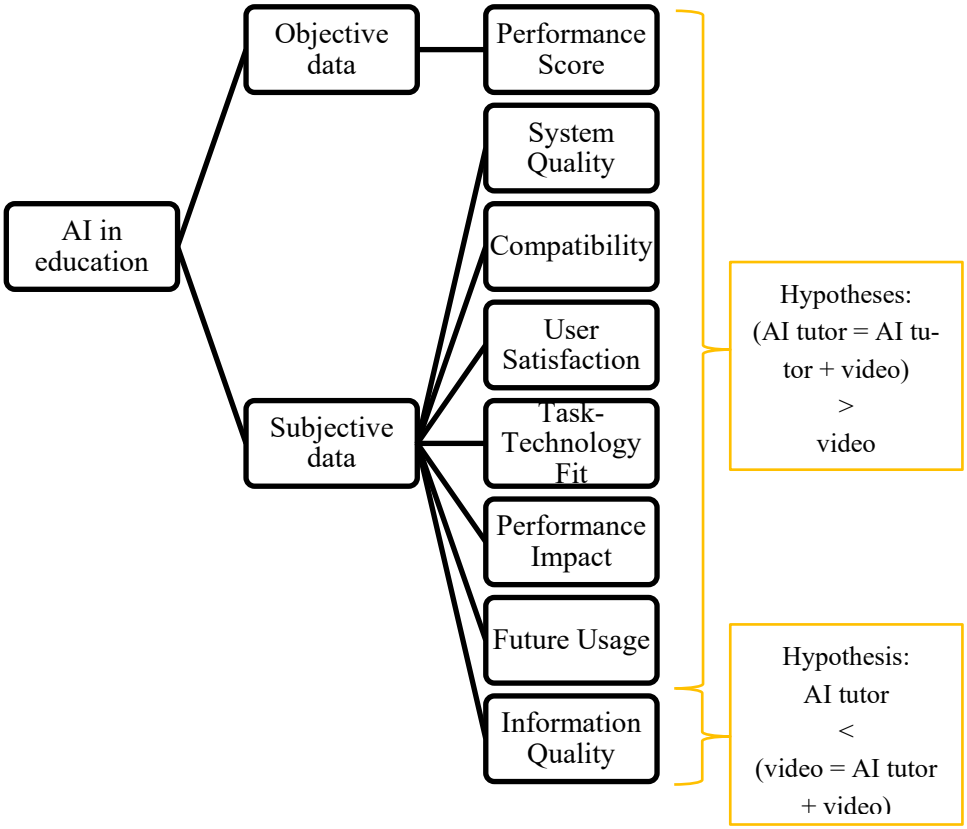

Fig. 1: Assessed scales in our study categorized according to objective and subjective data with corresponding hypotheses regarding our three experimental groups (AI-based tutor, video-based teaching, and a combination of AI tutor + video).

By examining both performance outcomes and subjective evaluations, this study aimed to provide a comprehensive evaluation of the role of a LLM-based AI tutor in higher education. Our results contribute to the ongoing debate on the efficacy and acceptance of AI-based educational tools, particularly in comparison with more traditional digital learning material such as video tutorials.

2 Methods

2.1 Study Design

This study was preregistered on Open Science Framework (OSF, see https://osf.io/6zm8r). Students were assigned to one of three groups: a *GPT-4o mini*-based AI tutor, which provided personalized guidance and answered questions (AI tutor group), a short video explaining the topic with examples (video group), and a hybrid approach combining both (AI tutor + video group). We measured the performance points in the R tasks and the subjective evaluations regarding System Quality, Information Quality, Compatibility, User Satisfaction, Task-Technology Fit, Performance Impact and Future Usage.

2.2 Participants

Participants were recruited from among Bachelor Psychology students at the University of Tübingen. The study was advertised with recruitment e-mails and messages to student groups that had already participated in a course teaching R (see Supplementary Materials E: https://osf.io/kpuqd). Thus, our participant pool was limited from the outset, and our primary objective was to recruit as many participants as possible within this small potential sample.

A total of 33 participants completed the experiment. We did not collect any demographic data such as age or gender to ensure the anonymity of our participants. We documented which university courses on R programming the participants had previously attended.

Given the restricted participant pool, we opted for performing a sensitivity analysis after data collection rather than a power analysis beforehand. The sensitivity analysis indicated that the smallest effect size detectable with 30 participants (i.e., actual-use groups sample size of analyses after exclusions, details see below) and a power of 0.80 ($1 - \beta = .80$) at a significance level of $\alpha = .05$ was $\eta^2 = .26$.

2.3 Materials

2.3.1 The AI Tutor

The AI tutor used in this study was built with a custom version of GPT, tailoring the GPT-4o mini model specifically for our educational needs with the following system prompt:

> "You are a R Tutor that focuses exclusively on R programming. You are designed to encourage self-discovery and learning through errors. When a student asks about a general R topic, you provide a short introduction to the topic along with an example. When a student inputs an incorrect answer, you give a hint about where the mistake is without providing the correct answer at all. This method helps in reinforcing learning and understanding, ensuring students engage deeply with the concepts and think critically about their approach to problem-solving in R programming. You will not respond to questions outside of R programming."

The prompt was designed to guide the AI tutor in delivering educational content, providing feedback, and responding to student inquiries in a pedagogically sound manner instead of just providing the correct solution. We left the LLM's temperature (controls randomness) and the Top P (controls diversity) parameters at their default values of 1.

To implement the AI tutor, we created a web page that interfaced with the GPT via the API and was hosted on a secure server managed by the university. A screenshot of the website is shown in Figure 2.

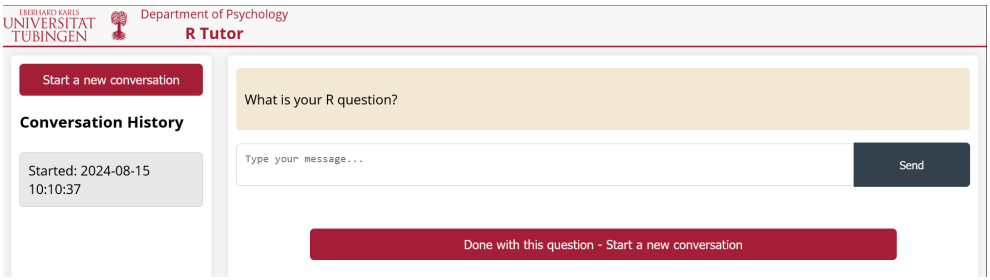

Fig. 2: Screenshot of the website that hosted the AI tutor.

2.3.2 The Video

To reflect the traditional teaching method in the R programming course of the Bachelor Psychology program at the University of Tübingen, we recorded a video that explains the basics of regular expressions in R. In this video, typical regular expression commands and metacharacters were demonstrated in R Studio using a sample data set. The video was almost 14 minutes long and was screen recorded by the adjunct lecturer of the R Programming seminar. A screenshot of the video is shown in Figure 3.

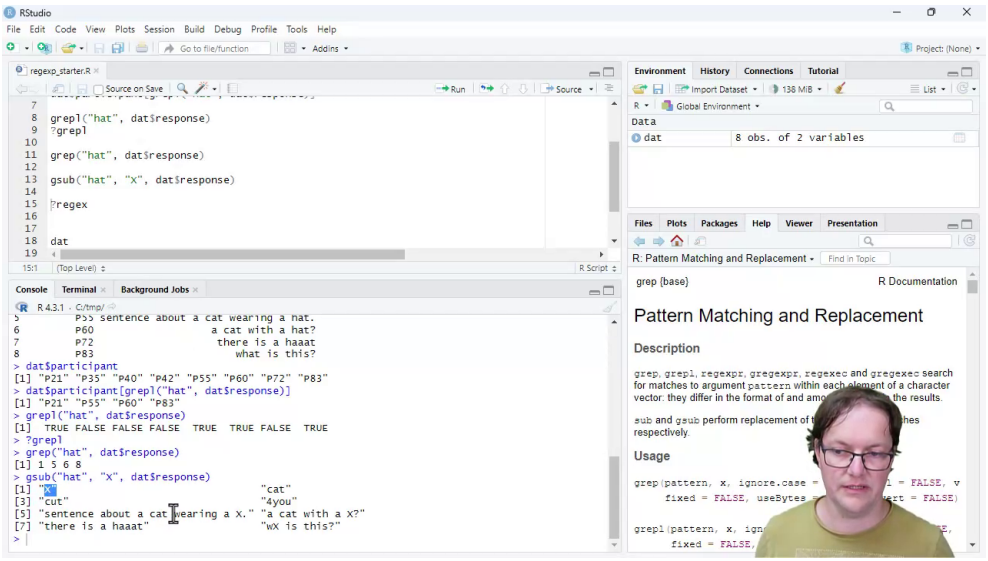

Fig. 3: Screenshot of the video tutorial explaining regular expressions in R.

2.3.3 The R Programming Performance Tasks

To measure learning outcome, we asked participants to solve six tasks with the help of regular expressions. In the recruitment email, we specified that participants should conduct the experiment on a computer with R installed. Trying out the R code on the computer to solve the tasks was voluntary, for which a code to generate the data set was provided. We provide the full task description including all regular expression tasks in Supplementary Materials A (see https://osf.io/6k7sy).

2.3.4 The Subjective Evaluations

The subjective evaluations were collected using seven scales: System Quality (three items), Information Quality (five items), Compatibility (three items), Task-Technology Fit (three items), Performance Impact (nine items), User Satisfaction (one item), and Future Usage (one item). Participants rated the statements on each scale using a

7-point Likert scale, ranging from “strongly disagree” to “strongly agree”. All statements were formulated in a positive direction, meaning that higher scores indicate a more favorable evaluation, ensuring a consistent response scale. The scales and statements were adapted from Isaac et al. (2019). We made adjustments to some of the statements to better fit our experimental setting, such as omitting items that did not directly relate to the learning environment the participants used. We provide the full questionnaire that we used in Supplementary Materials B (see https://osf.io/6uzgy).

2.4 Procedure

Participants started this study on SoSci Survey. After providing informed consent, participants were given a brief textual introduction to regular expressions. They were then directed to a page presenting the task description, R code for generating the data set, and all six R programming tasks. Based on their assigned group, participants had access to either a button that opened the video, a button that launched the AI tutor in a separate window, or both buttons. Participants were free to utilize these buttons as often as they wished throughout a maximum of 45 minutes to complete all tasks. The remaining time was displayed as a timer on the screen and they received a reminder five minutes before the time was up.

Upon completing the tasks, participants proceeded to the subjective evaluations. Each scale was presented on a separate page. At the end of the experiment, participants were asked for their consent for data usage once more and could choose to enter their ID to receive course credit and/or provide their email address to receive an example solution for the R tasks via email. Personal data was saved separately from the other responses. Participants were free to drop out of the experiment at any time without giving any reason.

2.5 Data Analysis

Two independent raters scored the R tasks (condition blinded) using a predefined scoring scheme (see Supplementary Materials C: https://osf.io/h63r8). The performance points for each participant were calculated as the average of the points assigned by the two raters. In addition to performance points, we calculated the average scores for each subjective scale by transforming the Likert scales into numbers (-3 to +3) and averaging the responses to all items within a scale. Following a detailed manipulation check, we reassigned participants to different groups based on their actual behavior during the study, correcting our preregistered analysis. Three participants from the AI tutor + video group were re-assigned to the AI tutor group because they did not open the video once. One participant from the AI tutor + video group was re-assigned to the video group because of opening the AI tutor but not sending at least one prompt to the tutor. This allowed us to increase the statistical power by retaining and analyzing more of the collected data as reported in the following. To analyze the data, we performed one-way ANOVAs on the average scores and pairwise t-tests for each scale. We made our data and analysis script available as open data on OSF: https://osf.io/9387s

3 Results

3.1 Assigned Groups

For our analyses following preregistered criteria, we analyzed the data according to the groups the participants were initially assigned to and we performed the preregistered exclusions. Specifically, we excluded three participants from the AI tutor + video group because they did not open the video once, two participants because they did not use any buttons of the assigned learning environment (1 x video group, 1 x AI tutor + video group) and one participant because of responding "NO ANSWER" to every task, leading to a total sample size of $N = 27$ participants with $n = 10$ in the AI tutor group, $n = 10$ in the video group and $n = 7$ in the AI tutor + video group. We conducted the preregistered one-way ANOVAs with assigned groups. They showed no significant difference in performance points or for any of for the subjective scales among the three groups. The full details regarding data preparation and results of our preregistered analysis are provided in Supplementary Materials D (see https://osf.io/hxuce*)*.

Given the restricted sample size, we performed the same analyses focusing on participants actual use of the learning environment, that is, actual-use groups instead of the groups they were initially and randomly assigned to (see also Methods section for details).

3.2 Actual-use Groups

The reassignment following the actual usage of learning environments led to a total sample size of $N = 30$ participants with $n = 13$ in the AI tutor used group, $n = 11$ in the video used group, and $n = 6$ in the AI tutor + video used group. Among the participants, $n = 16$ had previously completed the university course "Computergestützte Methoden", $n = 12$ had completed "R-Programmierung", and $n = 2$ had not completed either course prior to the experiment.

Cronbach's alpha (α) was calculated to assess the reliability of each scale: System Quality (α = .88), Information Quality (α = .83), Compatibility (α = .93), Performance Impact (α = .92), and Task-Technology Fit (α = .80) showed good reliability. User Satisfaction and Future Usage consisted of one item each, which is why Cronbach's alpha could not be calculated.

An Intraclass Correlation Coefficient (ICC; model: one-way, type: consistency) was calculated to assess the reliability of ratings in terms of total performance points assigned by the two raters across the 30 participants. The analysis yielded an *ICC(1)* = .99, 95% CI [.97, .99], indicating excellent reliability.

One-way ANOVAs and pairwise *t*-tests were conducted for each scale (see Table 1 and Figure 4). We observed no significant effect of the used learning environment on students' performance score. Regarding the subjective scales, we observed a significant influence of the used learning environment on System Quality and Task-Technology Fit. We observed a descriptive trend with the AI tutor used group rating their experience most favorably, which was also supported by statistically significant pairwise comparisons for System Quality, Task-Technology Fit, User Satisfaction, Compatibility, and Performance Impact. For these scales, the AI tutor used group was ranked highest, and the video used group was ranked lowest. The AI tutor + video used group was in between the other two groups for Task-Technology Fit, User Satisfaction, Compatibility, and Performance Impact, and it was like the video used group for System Quality. For the remaining subjective scales – Information Quality and Future Usage – we observed no significant effects.

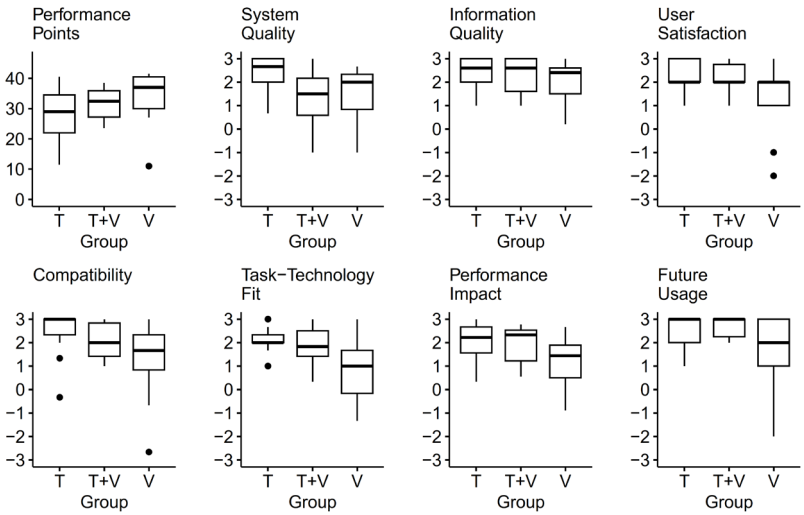

Fig. 4: Boxplots for each scale using the grouping of participants according to their actual usage behavior (N = 30; T: AI tutor used; T+V: AI tutor + video used; V: video used).

Table 1: Descriptive data and results of the ANOVAs for each scale using the grouping of participants according to their actual usage behavior (N = 30).

Scale	Group	Descriptives		ANOVA		
		M	*SE*	*F*(2, 27)	*p*	η^2 [95% CI]
Performance Points	AI tutor used	27.90	2.35	0.71	.503	.05 [.00, .24]
	AI tutor + video used	31.54	2.42			
	video used	32.16	3.42			
System Quality	AI tutor used	2.46[a]	0.19	4.04	.029	.23 [.00, .46]
	AI tutor + video used	1.28[b]	0.59			
	video used	1.36[b]	0.38			
Information Quality	AI tutor used	2.42	0.17	0.67	.523	.05 [.00, .23]
	AI tutor + video used	2.27	0.36			
	video used	2.05	0.26			
User Satisfaction	AI tutor used	2.38[a]	0.18	2.71	.085	.17 [.00, .40]
	AI tutor + video used	2.17[a,b]	0.31			
	video used	1.36[b]	0.47			
Compatibility	AI tutor used	2.46[a]	0.27	3.03	.065	.18 [.00, .41]
	AI tutor + video used	2.06[a,b]	0.35			
	video used	1.21[b]	0.49			
Task-Technology Fit	AI tutor used	2.08[a]	0.14	5.08	.013	.27 [.02, .50]
	AI tutor + video used	1.83[a,b]	0.39			
	video used	0.82[b]	0.42			
Performance Impact	AI tutor used	2.04[a]	0.23	2.36	.113	.15 [.00, .38]
	AI tutor + video used	1.91[a,b]	0.38			
	video used	1.22[b]	0.33			
Future Usage	AI tutor used	2.54	0.18	2.11	.141	.14 [.00, .36]
	AI tutor + video used	2.67	0.21			
	video used	1.73	0.49			

Note. Group means that differ in their superscript showed significant differences in the pairwise comparisons using *t*-tests with pooled SD.

3.3 Exploratory Analyses

We performed an exploratory one-way ANOVA to investigate whether the three groups differed in the time spent for solving the tasks. This analysis revealed no significant difference, $F(2, 27) = 0.43$, $p = .654$, $\eta^2 = .03$, 95% CI [.00, .19], in the time spent for solving the tasks across the three groups: AI tutor used group ($M = 33.40$ minutes, $SE = 2.59$), AI tutor + video used group ($M = 36.98$ minutes, $SE = 1.87$), video used group ($M = 33.29$ minutes, $SE = 2.74$).

Finally, we also conducted an exploratory analysis to examine the number of prompts used within the two groups that interacted with the AI tutor. A *t*-test revealed a significant difference, $t(17) = 2.61$, $p = .018$, $d = 1.29$, 95% CI [0.21, 2.33]), indicating that the AI tutor used group ($M = 10.46$, $SE = 1.28$) interacted significantly more with the tutor using prompts than the AI tutor + video used group ($M = 5.00$, $SE = 1.26$).

4 Discussion

We examined the effect of our AI tutor on students' performance and subjective experience compared to conventional digital teaching methods. Our findings using assigned groups reveal no significant differences in performance points or for any of the subjective evaluations among the three groups. We attribute these findings to the limited sensitivity of our preregistered analysis, as our sample size only allowed for the detection of very large effects.

The analyses using actual-use groups, however, revealed significantly higher scores for the AI tutor used group than video used group for some subjective scales. This suggests that participants found the learning environment consisting of the AI tutor to be easier, more flexible, and more understandable, while also meeting their expectations of technology support. These findings are consistent with those of Kim et al. (2020), who reported that the perceived usefulness of an AI tutor is rated highly by users, and, together with ease of communication, is a key predictor of the intention to actually use the AI tutor.

Our exploratory analysis also revealed that the AI tutor used group interacted significantly more with the tutor using prompts compared to the group that had both access to the AI tutor and video resources. Further, there was no significant difference in the time required to complete the tasks across the three groups. Although some literature posits an increase in efficiency by using AI tutors (e.g. Amdan et al., 2024), there is a lack of research specifically addressing time efficiency. Our results indicate that there may not be a time-saving advantage. Instead, the time needed to achieve similar performance points appears to be constant across the three groups. This raises important questions for further research, particularly regarding whether the anticipated efficiency gains from AI-assisted learning are indeed realizable in practice.

A limitation of this study is the quasi-experimental nature of the results using self-determined groups based on participants' actual use of the learning environments. This reassignment reflects a more intuitive assignment process, which deviates from a strictly randomized approach. Additionally, the necessity of reassignment could indicate that participants gravitated toward their preferred learning environment rather than engaging with both environments as intended. This preference may have influenced their subjective ratings.

Although our study was primarily limited by its sample size, it highlights promising potential for future research. Subsequent studies could explore AI tutor performance in contexts where participants have limited prior knowledge, particularly within higher education. There, the rapid advancement of knowledge is crucial to enhance educational outcomes, which might be accelerated by AI-assisted learning. Future research might also investigate less technical tasks, as these could explore differences of AI tutors in broader learning environments. This is particularly relevant as our study focused on regular expressions, a task domain in which GPT models presumably perform particularly well.

Due to the limitations of our study, we cannot rule out that students might perform differently with the varying learning environments, although our results show no difference in actual performance. Subjective evaluations suggest that learning with the AI tutor is favored by students. Some literature suggests a potentially beneficial

combination of AI tutors and video resources. Immediate and individualized feedback, as provided by the AI tutor, is a crucial factor that contributes to the long-term retention of the material students are expected to learn (Srinivasan & Centea, 2019). Additionally, guidance in the form of a video that is aligned with the curriculum can maximize the effectiveness and efficiency of independent learning (Saunders & Wong, 2020). This curriculum-oriented support helps ensure that students remain focused on relevant content and objectives, thereby enhancing the learning process. Therefore, further research is needed to determine whether or not there is an advantage of combining these learning resources.

Future research could also include an analysis of the intensity of AI tutor usage and a comparison of time differences between different AI tutor implementations (e.g., chat interfaces like in our study vs. programming copilots with AI-driven autocompletion functionality) to provide a comprehensive understanding of their relative efficacy. Additionally, investigating the combination of AI tutoring systems with human tutors would provide valuable insights into how integrative educational approaches would benefit from combining the two resources' relative strengths and weaknesses. However, it is crucial to recognize that integrating AI into educational environments presents significant ethical challenges, particularly concerning data privacy. Collecting and analyzing extensive personal data – such as students' learning styles, abilities, and progress – is necessary not only for research, but also for model training to improve existing AI tutors and develop increasingly powerful and advanced LLMs. Ensuring the protection of this sensitive information is vital, as any misuse could jeopardize students' privacy (Saaida, 2023). When ethical boundaries are respected, the potential to explore and exploit the transformative impact of AI on higher education is immense.

In the words of Kamalov et al. (2023): "Ultimately, we find that the only way forward is to embrace the new technology, while implementing guardrails to prevent its abuse." (p. 1). Put differently, the only way forward necessitates further empirical investigation into the effects of AI tutors, particularly those attempting to capitalize the highly transformative potential of generative AI, such as LLMs. Our quasi-experimental results show no significant increase in efficiency regarding performance

and time for AI tutor use, but they suggest that educational research should further explore the potential benefits and pitfalls of AI tutors. This is particularly relevant and promising given that our findings provide additional evidence for students' subjective preference for them.

Acknowledgements

We would like to thank Clément Préau for his help with scoring the R tasks. This research was funded by a teaching innovation grant awarded to Frank Papenmeier, Tobias Rebholz, and Tjark Müller by the University of Tübingen.

References

4 in 10 college students are using ChatGPT on assignments. (2024, February 27). Intelligent. https://www.intelligent.com/4-in-10-college-students-are-using-chatgpt-on-assignments/

Amdan, M. A. B., Janius, N., Kasdiah, M. A. H. B., Amdan, M. A. B., Janius, N., & Kasdiah, M. A. H. B. (2024). Concept paper: Efficiency of Artificial Intelligence (AI) tools for STEM education in Malaysia. *International Journal of Science and Research Archive*, *12*(2), 553–559. https://doi.org/10.30574/ijsra.2024.12.2.1273

Baillifard, A., Gabella, M., Lavenex, P., & Martarelli, C. (2024). Effective learning with a personal AI tutor: A case study. *Education and Information Technologies*, 1–16. https://doi.org/10.1007/s10639-024-12888-5

Bastani, H., Bastani, O., Sungu, A., Ge, H., Kabakcı, Ö., & Mariman, R. (2024). Generative AI can harm learning. *The Wharton School Research*. https://doi.org/10.2139/ssrn.4895486

Bilyalova, A. A., Salimova, D. A., & Zelenina, T. I. (2020). Digital transformation in education. In T. Antipova (Ed.), *Integrated Science in Digital Age* (pp. 265–276). Springer International Publishing. https://doi.org/10.1007/978-3-030-22493-6_24

Buchanan, J., & Hickman, W. (2024). Do people trust humans more than ChatGPT? *Journal of Behavioral and Experimental Economics*, *112*, 102239. https://doi.org/10.1016/j.socec.2024.102239

Cowen, T., & Tabarrok, A. T. (2023). How to learn and teach economics with Large Language Models, including GPT. *GMU Working Paper in Economics*, *23*(18). https:///dx.doi.org/10.2139/ssrn.4391863

Frankford, E., Sauerwein, C., Bassner, P., Krusche, S., & Breu, R. (2024). AI-tutoring in software engineering education. *Proceedings of the 46th International Conference on Software Engineering: Software Engineering Education and Training*, 309–319. https://doi.org/10.1145/3639474.3640061

Isaac, O., Aldholay, A., Abdullah, Z., & Ramayah, T. (2019). Online learning usage within Yemeni higher education: The role of compatibility and task-technology fit as mediating variables in the IS success model. *Computers & Education*, *136*, 113–129. https://doi.org/10.1016/j.compedu.2019.02.012

Kamalov, F., Santandreu Calonge, D., & Gurrib, I. (2023). New era of Artificial Intelligence in education: Towards a sustainable multifaceted revolution. *Sustainability*, *15*(16), 12451. https://doi.org/10.3390/su151612451

Kim, J., Merrill, K., Xu, K., & Sellnow, D. D. (2020). My teacher is a machine: Understanding students' perceptions of AI teaching assistants in online education. *International Journal of Human-Computer Interaction*, *36*(20), 1902–1911. https://doi.org/10.1080/10447318.2020.1801227

Lin, C.-C., Huang, A. Y. Q., & Lu, O. H. T. (2023). Artificial intelligence in intelligent tutoring systems toward sustainable education: A systematic review. *Smart Learning Environments*, *10*(1), 41. https://doi.org/10.1186/s40561-023-00260-y

McClain, C. (2024, March 26). Americans' use of ChatGPT is ticking up, but few trust its election information. *Pew Research Center*. https://www.pewresearch.org/short-reads/2024/03/26/americans-use-of-chatgpt-is-ticking-up-but-few-trust-its-election-information/

Paul, P., Bhimali, A., Kalishankar, T., Aithal, P. S., & Rajesh, R. (2018). Digital education and learning: The growing trend in academic and business spaces – an international overview. *International Journal on Recent Researches in Science, Engineering & Technology (IJRRSET)*, *6*(5), 11–18.

R Core Team. (2024). *R: A Language and Environment for Statistical Computing* [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/

Saaida, M. B. (2023). AI-Driven transformations in higher education: Opportunities and challenges. *International Journal of Educational Research and Studies*, *5*(1), 29–36.

Saunders, L., & Wong, M. A. (2020). *Learning theories: Understanding how people learn*. https://iopn.library.illinois.edu/pressbooks/instructioninlibraries/chapter/learning-theories-understanding-how-people-learn/

Shaengchart, Y. (2023). A conceptual review of TAM and ChatGPT usage intentions among higher education students. *Advance Knowledge for Executives*, *2*(3), 1–7. https://ssrn.com/abstract=4581231

Srinivasan, S., & Centea, D. (2019). An active learning strategy for programming courses. In M. E. Auer & T. Tsiatsos (Eds.), *Mobile Technologies and Applications for the Internet of Things* (pp. 327–336). Springer International Publishing. https://doi.org/10.1007/978-3-030-11434-3_36

Tian, H., Lu, W., Li, T. O., Tang, X., Cheung, S.-C., Klein, J., & Bissyandé, T. F. (2023). *Is ChatGPT the ultimate programming assistant – How far is it?* (No. arXiv:2304.11938). arXiv. https://doi.org/10.48550/arXiv.2304.11938

Vargas-Murillo, A. R., Pari-Bedoya, I. N. M. de la A., & Guevara-Soto, F. de J. (2023). Challenges and opportunities of AI-assisted learning: A systematic literature review on the impact of ChatGPT usage in higher education. *International Journal of Learning, Teaching and Educational Research*, *22*(7), Article 7.

Zografos, G., & Moussiades, L. (2023). A GPT-based vocabulary tutor. In C. Frasson, P. Mylonas, & C. Troussas (Eds.), *Augmented Intelligence and Intelligent Tutoring Systems* (pp. 270–280). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-32883-1_23