Exploring the Role of AI Chatbots in K-12 Education: A Comparative Study of Socratic and Non-Socratic Approaches

Point-by-Point Rebuttal to Reviewer Comments

Thank you for the constructive feedback. Below is our detailed, point-by-point response to the reviewers' comments. Reviewer comments are in **bold**, and our responses follow in regular text with manuscript changes noted where applicable.

Reviewer 1

This paper investigates whether and how the integration of large language models (LLMs) into classroom activities influence students' learning. Although the topic, questions, and results are crucial, the manuscript requires revision before publication. Therefore, I recommend a major revision, and I am willing to review the revised manuscript. I have provided a list of comments below to assist the authors in transforming their paper into a manuscript worthy of publication.

The abstract sound great.

Thanks for this nice comment.

The literature review provides a broad context for AI's potential; however, it lacks a strong connection to recent studies that address AI's evolving role in educational contexts, critical thinking, and argumentative writing. The paper could enhance its theoretical foundation by engaging more deeply with contemporary research on AI's impact on students' critical thinking, argumentation and reasoning. There are much more recent studies by scholars such as Kazem Banihashem and that you can consult with.

We thank the reviewer for this comment and suggestion. We have now expanded the literature review into a dedicated section (pp. 6-8). We have elaborated on the evolving role of LLMs in education, focusing on the literature specific to AI chatbots (drawing many insights from Kazem Banihashem and co-authors's work). The section discusses the main findings and points out to the gaps in the literature that we aim to fill with our study. At the same time, given the novelty of the LLMs applications in education, the theoretical foundation for our work is necessarily limited.

Please justify why the participants come from Brussels (Belgium) and one in Seville (Spain)? Why these two countries and these two cities? What is the argument here?

We recruited schools in different countries to increase the external validity of our study. The country choice was by convenience as the authors lived in Brussels and Seville at the time of the experiment. Furthermore, the selected schools had similar curricula and backgrounds —they were both international with students from comparable socio-economic backgrounds — which we deemed ideal for conducting the experiment. We clarified this point in footnote 3 (p. 10).

More information on the method section, the context, the design, and the participant characteristics are needed to give the readers a clear description of the study. Please elaborate on the background information of the characteristics of the participants that could influence the way learners engage with AI and critical thinking and argumentation. Acknowledging these limitations can guide future research in this area.

We have revised the manuscript expanding the Methods section to improve the clarity of our experimental design and provide more details about the context. Specifically:

- The revised Methods (pp. 9-14) section begins with a summary of our research design and setup to address the need to give the readers a clear description of the study, before presenting the details in the relevant subsections.
- We now clearly justify the recruitment of the two schools (see footnote 3), as in Reviewer 1's point discussed above.
- We also expanded the "Background Measures" sub-section into a new section "Student characteristics" that provides a detailed description of all the background variables collected, including our questions on students' AI attitudes and experience, their academic habits and grades. We have further included a new sub-section discussing how we collected AI-student interaction metrics.

- Finally, in the Results section, we have revised Table 2 to illustrate the distribution of participants characteristics across treatments, which provides more precise information on the demographics and academic skills of our participants.

The results show promising outcomes. However, the findings of this study are derived from self-reported data collected from participants through surveys. However, it is important to note that the literature suggests caution in heavily relying on such self-reported data, as perception does not always equate to actions (see: https://www.tandfonline.com/doi/full/10.1080/02602938.2024.2345669). Therefore, this must be acknowledged as a limitation of the study, along with suggestions for future research to address this issue.

Many thanks for this important point. We have discussed this limitation in the Discussion section, with the following paragraph (p. 21):

"*Additionally, our study combined objective performance metrics with self-assessed ratings of helpfulness and confidence. However, individual perceptions do not necessarily reflect actual learning (Noroozi et al. 2025). Further research is needed to validate our findings using objective learning measures.*"

The depth and width of the discussion section could be improved by incorporating relevant and recent literature. This would provide more theoretical insights and add value to the study by linking the findings to the broader literature.

We thank the reviewer for this comment. We have now expanded the Discussion section to include all relevant supporting literature and reflect on how the existing literature supports our findings as well as how our findings contribute to the current gaps in the literature.

Reviewer 2

The article deals with a current and pressing topic. Since AI-based Large Language Models have become freely accessible, the way in which students learn has changed rapidly. There are new opportunitiesbut also risks for didactic methods. The article addresses the use of an AI tutor to support students' learning and understanding. An experimental setup is used to investigate the conditions under which an AI tutor enhances learning, how an AI tutor influences students' confidence in their own performance, and whether an AI tutor is perceived as helpful by students.

The article presents three research questions. The first question should be expressed with greater

precision. Of particular note, the second part of the research question is not addressed by the selected experiments. The second research question, which comprises two distinct inquiries, should be subdivided to enhance its clarity and comprehensibility. The research questions are derived from existing literature. However, this section is rather brief. Notably, no hypotheses have been formulated or derived from the literature. Only one hypothesis has been stated explicitly. However, this has been done in the results section only, contrary to scientific conventions.

Many thanks for the opportunity to clarify this important point. In response, we have modified the Introduction section to improve the clarity, precision, and coherence of our research questions. Specifically:

- We have refined and split the original Research Question 1 into two separate and more focused questions (RQ1 and RQ2) to better reflect the distinct aspects under investigation. This addresses the reviewer's concern about the lack of precision and the mismatch between the research question and the experiments.
- Similarly, we have divided the original Research Question 2 into two distinct questions (RQ3 and RQ4) to improve clarity and comprehensibility, as suggested.
- To strengthen the theoretical foundation of our study, we have clarified what justifies the formulation of each research question more thoroughly and have expanded the literature review into a new subsection (pp. 6-8).
- In line with scientific conventions, we have now explicitly stated all hypotheses in the Introduction, rather than introducing them only in the Results section. And the Results section has also been updated to link the analysis to each of the five research questions (see next point).

The revised research questions are as follows:

- RQ1: Do AI-generated explanations enhance students' problem-solving performance in school tasks?
- RQ2: How do AI-generated explanations influence students' perceived credibility of AI- generated solutions?
- RQ3: Do student-AI interactions guided by the Socratic method promote better performance?
- RQ4: Do student-AI interactions guided by the Socratic method increase students' confidence in their answers?

- RQ5: Do students perceive Socratic student-AI interactions more helpful than non-Socratic AI interactions?

In the Methods section and in the Supplementary Materials, the experimental procedure is largely described in a comprehensive and reproducible manner. However, the methods/experiments/questionnaire items are not related to the previously stated research questions or hypotheses. This makes it unnecessarily difficult for the reader to assess the validity of the approach chosen to address the research questions. To enhance comprehensibility, it would be beneficial to explicitly label the independent and dependent variables, as well as the control variables collected for each research question/hypothesis.

We fully agree with this point. In the Results section, we now make a direct reference linking the analysis to each of the five research questions.

The results section first describes the group of subjects. For one experimental intervention (socratic AI tutor), it is reported whether the experimental group differed significantly from the control group. However, this information is missing for the other experimental intervention (AI explanation). In the remainder of the section, the analyses of the dependent variables are presented in a largely scientific manner. Notably, for the first experimental manipulation, a dependent variable has been listed in the methods section (propensity to update initial guess). However, no results are reported for this variable, and I do not understand what was done with this data or why it wasn't used. In the second experimental manipulation (socratic tutor), variables are reported that have not been described in the methods section. These new variables that measure the interaction with the AI tutor are only (and very briefly) motivated here. They should be motivated in the Introduction and included in the Methods section.

We thank the reviewer for these suggestions that greatly improved the clarity of our methods. Specifically:

- Regarding the descriptive information about the other experimental intervention, Table 2 now shows summary statistics for both treatment groups. It also shows p-values from separate statistical tests of independence between variable and treatmnet assignment.
- Regarding the propensity to update the guess, we found no evidence that the treatment had an effect on this propensity. Accordingly, we have included the following sentence in the Results section (p. 16): "*Regarding the propensity to update the guess, we found no evidence that the treatment had an effect on this propensity: Nearly all students revised their*

initial estimates (110 out of 122), with no significant difference between treatment groups (Fisher's test, p = 0.9)."

- About the student-AI interaction metrics for the Socratic intervention that have not been fully described in the Methods section, we now report this information in a dedicated sub-section titled "Student-AI Interaction Metrics" in the Methods section (p. 14). We report the main paragraph below: "*To quantify engagement, we analyzed students' interaction logs with the AI Tutor, recording the number of turns and the word counts as a proxy for interaction intensity. While this provides a basic behavioral metric, it does not capture the quality of cognitive engagement or helpfulness,limiting its direct relevance to RQ1 and RQ2. Therefore, we also asked students about how useful they found the AI interactions and how confident they were in their answers to selected tasks. These self-reported metrics are needed to assess whether explanations were helpful to students (for RQ1) and whether Socratic dialogue promoted critical thinking or confidence (for RQ2).*"

Contrary to usual conventions, some results with a significance level greater than 0.1 are reported as significant. In addition, in some cases, the language used is not specific/detailed enough to correctly understand the analyses (Which analysis uses exactly which data or answers from the questionnaire?). In some cases, the methods used are only briefly named and may therefore not be reproducible (in particular bootstrap method, in some cases also regression models - which variables exactly have been included?).

- In accordance to convention, we changed the sentence from "marginally significant (p = 0.15)" to "in-significant (p = 0.15)".
- We have also clarified the bootstrap resampling approach used to compute the difference in media accuracy between the treatmetn groups. Specifically: "*Figure 4 illustrates that students' prediction accuracy was positively skewed in both treatment groups, with a greater accuracy for students exposed to AI reasoning. To estimate confidence intervals for the median difference in accuracy betweeen the groups, we used a nonparametric bootstrap approach. This procedure involved resampling participants' absolute errors (n = 1,999) to build a distribution of medians for each group. The 5th and 95th percentiles of the resulting distribution were used to construct a 90% confidence interval for the difference in accuracy between groups. The interval ranged from 0 to 70, indicating a greater accuracy (lower error) for the students exposed to AI-generated reasoning (one-sided, p < 0.05).*"
- Regarding the lack of clarity on the additional regression models, as pointed out by Reviewer 1 in

the annotated document, we have revised the results section to include the results in a new figure (Figure 12). We have also clarified the analysis in the supporting information (p. 46). Specifically: "*The regression specification described above assumes homogeneous treatment effects across individuals and tasks. To relax this assumption, we employed a more flexible model that allows the average treatment effect, γ, to vary based on individual-level and task-specific factors. Specifically, we included interaction terms between the treatment indicator and variables such as the student's gender, academic performance (as measured by school grades), prior experience with ChatGPT, and school location. Additionally, we accounted for heterogeneity in treatment effects across tasks by estimating separate regressions for each task-specific subset of the data. The results are shown in Figure 8, Figure 10, and Figure 12.*"

The discussion addresses the research findings and identifies some methodological limitations. Since there was little specific literature on the specific research questions in the Introduction, the discussion of the results in the context of the existing literature is also very brief. However, it is particularly striking that the entire discussion does not contain a single reference/citation. In my opinion, this would be essential for a scientific publication.

We thank the reviewer for this observation. We have now expanded the discussion with supporting literature and reflect on how the existing literature supports our findings as well as how our findings contribute to the current gaps in the literature.

In my opinion, some interpretations of the results go too far. In particular, the data on learning and retention effects can only give first indications, because each type of task was completed exactly once by the students. The database is therefore too small to provide generalizable results. Nevertheless, the article contains interesting results, especially on the interaction/acceptance of a socratic AI tutor by people who had prior experience with chatGPT.

We have now revised the interpretation on retention effects, weakening it (see changes in the "Learning Outcomes and Knowledge Retention" section p. 19, and the Discussion section pp. 20-22). Furthermore, following the Reviewer's suggestion in the annotated file, we also acknowledge possible confounding factors in the analysis of retention in the footnote 10 (p. 19): "*While it is true that gold is significantly denser than water or air, it is possible that some students did not fully understand this fact or were unsure how to compare the densities of the materials. If that were the case, even if the AI effectively conveyed that sound travels faster in denser media, students lacking this knowledge may still have been unable to select the correct answer. However, this*

explanation seems unlikely, as the relative densities of air, water, and metals, like gold, are commonly taught and conceptually straightforward, suggesting that other factors have contributed to studens' outcomes."

The methodological discussion contains important points but could be more detailed. In particular, the validity of the methods used should be discussed (was the question about transferability of knowledge appropriate (sound/density)?). Further, it is not discussed whether the setting at different schools in two different countries could have influenced the results.

We have addressed these important points in the discussion section. Specifically:

- We added the following sentence (p. 21) on the limits of the methods for transferability of knowledge: "*Another limitation is that learning retention was tested using only one specific physics question. Although we controlled for prior knowledge and carefully designed a simple task to fit within a 40-minute intervention, additional questions would be needed to rule out confounding factors.*"
- In the discussion, we also clarified the possible limitations of focusing on two schools (p. 21): "*Additionally, we focused on only two schools, which allowed us to control for school-specific fixed effects. However, we recognize that the impact of the treatment may differ across schools, and a broader investigation involving many more institutions would be necessary to explore this variability.*"

Overall, the article contains good and important research approaches and first interesting results, but the scientific reporting should be significantly improved. (If you wish, you can find more detailed comments in the annotated PDF.)

As illustrated by the points discussed above, we have significantly improved the reporting, and we sincerely thank Reviewer 2 for the thorough and insightful comments provided in the annotated PDF, which guided this revision.