# Rebuttal to Reviewer Comments – Round 1

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.**

### Comments

**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. We have now expanded the literature review into a dedicated section (p. xxx). We have elaborated on the evolving role of LLMs in education, focusing on the literature speficic to AI chatbots. The section discusses the main findings and points out to a few 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 are 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.

**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 section begins with a summary of our research design and setup before presenting the details in the relevant subsections.
* We now clearly justify the recruitment of the two schools (see footnote XXX), as in Reviewer 1’s point discussed above.
* We also expanded the “Background Measures” sub-section into a new section “Student characteristics” tthat provides a detailed description of all the variables collected, including our questions on students’ AI attitudes and experience, their academic habits and grades, and we have included a new sub-section discussing how we collected AI-student interaction metircs.
* Finally, in the Results section, we have revised Table 2 (p. XXX) to illustrate the distribution of participants characteristics across treatments, which provides a 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://urldefense.com/v3/https://doi.org/10.1080/02602938.2024.2345669;!!DOxrgLBm!DU2xLNvI0Ke45t43rGQAI3BYa8bni0Io2JesBwINodtn8kzLHvspBE2my62mGZlAoKzFv6RFxgUv6FWyBQIbCg9ObQu4SA$ ). 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 conclusions [TBA] “solely on students’ perceptions of their learning experiences may not accurately reflect their actual learning”

**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.**

Many thanks we have …

## 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.**

### Comments

**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 expanded the literature review to more thoroughly justify the formulation of each research question.
* 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 that we have addressed by updated the Results section with 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.**

Specifically:

* Regarding the missing information about the other experimental intervention, Table 2 now shows descriptives 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. We report this analysis in footnote XXXX at p. xXXX.
* Regarding the missing information in the Methods section, we now report the student-AI interaction metrics – logs and number of words – in a dedicated sub-section titled “Student-AI Interaction Metrics” in the Methods section (p. xxxx).

**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 with “marginally significant (p = 0.15)” to “insignificant (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 @ref(fig:boot-accuracy) 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).*

* In the Supporting Information section, we have clarified the information about the regression models [TBC]

**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.**

TBA

**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 revised the statement on retention effects. We appreciate the suggestion that is considered in the footnot.

*foonote here*

**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.**

OK

**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 discussed above, we have greatly improved the reporting and we thank Reviewer 2 for his work and detailed comments in the annotated PDF that have been extremely valuable for this revision.