Exploring the Role of AI Chatbots in K-12 Education: A Comparative An Experimental Study of Socratic and vs. Non-Socratic Approaches and the Role of Step-by-Step Reasoning

[redacted]

This version: June 03, 2025

[1] [redacted]

[2] [redacted]

* Correspondence: [redacted]>

Highlights

- An experiment was conducted where K-12 students (N=122) used AI chatbots to solve school-related problems.
- Two ~~AI chatbot approaches~~ interventions were tested: ~~one provided incremental guidance to encourage critical thinking (“Socratic”) , while the other offered immediate solutions (“~~(1) comparing AI-generated step-by-step explanations to predictions alone, and (2) contrasting AI chatbot support inspired by the Socratic method (guided questioning) versus non-Socratic ~~”) .~~ (direct answers) support.
- ~~Results showed that AI-generated explanationsimproved students’ performance over solutions without them, highlighting the value of AI-generated guidance.~~
- ~~Participants~~ Participants performed better under step-by-step explanations. They also engaged more frequently with the Socratic AI, though this approach did not result in improved performance and retain of knowledge compared to the non-Socratic AI.
- While students found interactions with the AI useful, they perceived the Socratic AI as less helpful than non-Socratic overall.
- The findings highlight challenges in designing AI tutors that effectively foster critical thinking while maintaining student satisfaction, raising concerns about their adoption in educational settings.

Abstract

How does integrating large language models (LLMs) into classroom activities influence students' learning? To ~~address~~ investigate this question, we conducted a randomized experiment ~~comparing two distinct chatbot approaches: one designed to encourage~~ with students aged 14-16 (N=122), testing two interventions: (i) comparing AI-generated solutions with step-by-step explanations versus solutions alone and (ii) Socratic AI chatbots promoting critical thinking through incremental guidance ~~("Socratic AI") and another providing immediate solutions ("~~versus non-Socratic AI ~~"). The study involved students aged 14 to 16, who engaged in various~~ offering direct solutions. Students completed school-related tasks under these ~~experimental conditions. Students' attitudes were measured~~ conditions. We collected students' performance and interaction logs data, as well as their attitudes through self-reported surveys. Results indicated that AI-generated explanations significantly improved students' performance over solutions ~~provided without such explanations~~alone, highlighting the benefits of step-by-step guidance. The Socratic AI approach fostered significantly greater engagement and interaction. However, it did not achieve significant improvements in learning, and a higher fraction of students perceived it as less helpful. Furthermore, despite ~~students generally perceiving AI assistanceas beneficial, they~~ overall positive perceptions of AI assistance, students exhibited limited retention, failing to apply learned concepts to new situations ~~when we removed AIassistance~~without the aid of AI. These findings ~~contribute to the ongoing debate on integrating LLM-powered chatbots in education, highlighting key challenges in designing AI tutors that effectively foster critical thinking while maintaining student satisfaction, raising concerns about their adoption in educational settings~~underscore both the potential and limitations of LLM-based chatbots in K-12 education, especially in balancing the development of critical thinking with user satisfaction.

Keywords: artificial intelligence; large language models; learning; education policy; experiment~~,~~ ; k-12

1 Introduction

Integrating artificial intelligence (AI) into educational settings has been debated for over three decades (?). Recent advances in generative AI and large language model (LLM) chatbots have opened even more possibilities for improvements in educational practices (?; ?; ?; ?). Although LLMs mimic human intelligence, their applications extend beyond imitation and include a range of ~~educational~~ academic tasks: facilitating problem-solving (?), providing personalised assistance (?), language learning (?; ?), ~~or evaluation of~~ and evaluating students' work (?). However, the impact of these applications on students' learning remains controversial, with proponents highlighting potential benefits and critics cautioning about risks and drawbacks ~~. This study addresses two critical aspects of AI integration in education: the effectiveness of AI-generated explanations and the optimal modes of interaction between AI and students, specifically whether Socratic or alternative approaches yield better outcomes.~~ (?; ?; ?).

A key advancement over previous AI systems is that LLMs can ~~generate~~ provide step-by-step ~~solutions to accompany their responses to~~ reasoning to answer students' queries ~~. This study investigates how this ability influences students' interactions and helps them enhance their problem-solving skills. Specifically, we first examine the impact of AI explanations on students'performance in a numerical estimation task and their assessment of the accuracy of AI-generated predictions. Then, we test two approaches to access these explanations through AI chatbots: one approach that encourages critical thinking through incremental guidance ("Socratic AI")and another providing immediate solutions ("non-Socratic AI").~~

~~Previous research has shown that~~ (?) and that by adding "Let's think step by step' ', LLMs can improve their performance significantly (?). Although evidence suggests that such explanations can foster children's ~~causal~~ learning and contribute to the development of their scientific reasoning (?; ?; ?)~~. Similarly, explanations generated by LLMs may guide studentsin their~~ , there is limited empirical research examining how such AI-generated explanations impact students' learning processes. Most prior studies have not isolated the specific role of detailed AI-generated reasoning in enhancing problem-solving skills. To fill this gap, we conducted a randomised experiment with k-12 students (n = 122) who were engaged in a simple estimation task – guessing the value of coins in a jar. Half of the participants received only an AI-generated prediction, while the other half were additionally provided with AI-generated explanations detailing how to derive such estimation.

This setup enabled to address our first research question (RQ):

RQ1: Do AI-generated explanations enhance students' problem-solving performance in school tasks, specifically in terms of estimation accuracy?

The distinction between AI-generated *explanations* and *solutions* further allows us to evaluate whether providing explanations influences students' trust and judgment of AI-generated content. Indeed, ~~for example, by decomposing a numerical estimation task into more manageable steps, such as considering heuristics or approximations. However,~~ LLMs can ~~also generate inaccuracies~~ generate inaccurate solutions and flawed reasoning (?; ?; ?)~~, misleading students if followed without scrutiny. Understanding this dual role of AI explanations as a valuable educational tool and a potential source of misinformation leads to our first research question(RQ)~~. However, there is evidence that providing explanations increases users' trust and acceptance of AI systems, especially when the reasoning behind the output is clear (?; ?). In an educational setting, this may affect learning outcomes in a complex manner. Students may interpret well-written explanations as evidence that an answer is correct, even when it is not. Conversely, they may notice minor errors or inconsistencies in an explanation that reduce their trust, even if the overall solution is accurate. This situation motivates our second research question:

~~RQ-1~~RQ2: ~~Do~~ How do AI-generated explanations ~~enhance learning by providing insights, or do they risk impairing~~ influence students' ~~critical thinking by presenting inaccurate reasoning as authoritative~~perceived credibility of AI-generated solutions?

~~The study also aims to explore how different modes of student-AI interaction more broadly influence students' performance and perceptions, of which providing explanations is one key element. Specifically, we contrast two fundamental approaches: the “Socratic ” and~~ Another critical technical progress of contemporary LLMs over previous AI systems is their ability to follow specific instructions when engaged in conversations. Building on this capability, another goal of this study is to examine the impact of two fundamental modes of interaction on students' learning outcomes: one that fosters critical thinking through incremental guidance (“Socratic AI”) and another that offers immediate solutions (“non-Socratic ~~” methods. The Socratic approach, inspired by the Socratic teaching method~~ AI”).

Although classical pedagogical methods, such as the Socratic Method, have long been valued for promoting critical thinking and a deeper understanding (?; ?~~), engages students through an~~

~~argumentative dialogue where the AI tutor asks thought-provoking questions to stimulate critical thinking and self-reflection (?). This~~ ; ?; ?), there is a notable gap in the AI education literature regarding how to integrate these methods into AI-driven solutions. Existing AI tools, such as ChatGPT, often focus on providing direct solutions or feedback, which can potentially limit opportunities for active student engagement and reflection. Our study addresses this gap by conducting a parallel experimental intervention that compares Socratic-style AI interactions with more conventional, solution-focused AI interactions in a K-12 setting. This comparison contributes novel evidence on how different AI interaction modes can influence student learning processes and outcomes, informing the design of more pedagogically sound AI educational systems.

More specifically, building on previous literature suggesting that the Socratic method encourages students to think critically and arrive at their answers rather than relying on the AI-generated solution, thus helping students gain more confidence in their reasoning (?)~~. In contrast, the~~ , we test the following research questions:

RQ3: Do student-AI interactions guided by the Socratic method promote better performance?

RQ4: Do student-AI interactions guided by the Socratic method enhance students' confidence in their answers?

In addition to educational outcomes, exploring the trade-offs between Socratic and non-Socratic ~~approach directly answers students ' queries without necessarily engaging them in a dialogue or exploration.~~ AI raises an important question about student *satisfaction*. If students find interacting with Socratic AI chatbots unengaging or frustrating, such perceptions might drive students away from Socratic AI to seek other (non-Socratic) tools. Indeed, research on conversation length with LLM-powered chatbots has shown that people tend to have mixed reactions when chatbots are instructed to engage in more extended conversations (?). Therefore, aligning educational benefits with user satisfaction is crucial for ensuring the long-term adoption of AI learning tools. This consideration leads to our last research question:

~~The Socratic approach~~**RQ5: How do students perceive the helpfulness of Socratic AI compared to non-Socratic AI?**

Addressing these questions aims to contribute to a growing debate on using pedagogical principles and techniques, like the Socratic approach, to integrate AI in education, which has recently gained considerable attention~~in the context of AI tutoring in education. Notably~~. For instance, Khan Academy has developed its AI tutor, Khanmigo, based on ~~these~~ Socratic principles, aiming to foster critical thinking and engagement in learners.[1] ~~.~~ Several commentators have argued that the Socratic method is particularly effective for designing AI tutors for children, as it enhances critical thinking and discourages students from copying AI-generated responses without scrutiny (?).[2] ~~This debate raises the following research question:~~ Beyond education, Socratic AI systems are also being developed to foster critical thinking among the general public, particularly in efforts to combat disinformation (?).

~~**RQ-2: Do student-AI interactions guided by the Socratic method promote deeper critical thinking and enhance students' confidence in their answers?**~~

1.1 Literature Review

~~Another objective of comparing Socratic~~ A growing body of research highlights the tremendous potential of LLMs to improve student learning outcomes. A recent scoping review identifies 53 distinct use cases for LLMs in education (?), ranging from automatic grading and personalised feedback to teaching support and content generation, reflecting the high versatility and scalability of LLMs. Within this expanding literature, this study centres on *AI chatbots*, conversational agents powered by LLMs, and ~~non-Socratic AI is to assess the extent of "user satisfaction" among students. Regardless of its academic benefits, Socratic AI could find limited application if students find interacting with Socratic AI chatbots unengaging or frustrating. Negative interactions or perceptions might drive students away in favour of more straightforward (non-Socratic)tools that provide direct answers, such as ChatGPT. Therefore, aligning educational benefits with user satisfaction is crucial for ensuring the long-term adoption of AI learning tools. This leads to our last research question:~~ their educational applications, reflecting their increasing prominence and potential impact in this field.

~~**RQ-3: How do students perceive the helpfulness of Socratic AI compared to non-Socratic AI?**~~ Early investigations into the use of AI chatbots, such as ChatGPT, in educational settings have identified both opportunities and challenges (?; ?). On the one hand,

[1] https://docsbot.ai/prompts/education/khanmigo-lite-socratic-tutor

[2] See also https://research.gatech.edu/ai-oral-assessment-tool-uses-socratic-method-test-students-knowledge

ChatGPT can serve as a personalised tutor, assist in drafting assignments, or generate creative prompts for classroom activities. On the other hand, concerns have been raised around academic integrity, overreliance on AI-generated content, and the need for critical thinking skills when interpreting chatbot outputs.

~~By focusing on the impact of AI explanations and the Socratic AI approach on students' critical thinking, we aim to contribute to a growing literature about integrating AI into education. Recent research on AI chatbots and LLMs indicates significant potential for improving student learning outcomes. A meta-analysis of~~ A recent review of over 70 studies provides a comprehensive overview of how chatbots are used in various educational contexts (?). It classifies chatbot applications by roles, including tutoring, giving feedback, and helping with administrative tasks, and examines how they are designed and utilised. The review finds that AI chatbots generally meet their goals. For example, a quasi-experimental study with 320 middle schoolers found that an AI teaching chatbot significantly improved math knowledge (?). Similarly, a meta-analysis of 24 ~~randomised studies highlights the positive impact of chatbots, especially on student productivity and learning engagement (?). However, several studies are more cautious about the benefits, partly due to concerns that AI might encourage cheating, over-reliance, or superficial understanding, which are challenging to measure in short-term studies (?). Additional studies on LLMs' impact on academic integrity show that students and teachers struggle to differentiate between human and AI-generated text, raising risks of increased academic dishonesty and eroded trust (?; ?). Furthermore, as LLMscontinue to evolve, students and educators may lack the skills to leverage them effectively, which could result in missed educational opportunities~~ RCTs shows that AI chatbots significantly enhance student outcomes, with stronger effects in higher education (~~?).~~

~~The ethical and societal implications of LLMs in education are also complex. Issues such as bias, privacy, surveillance, and autonomy persist, especially in K-12 contexts, where the impact of algorithmic decisions on young learners' agency and privacy can be profound (?). LLMs introduce unique challenges, notably in alignment, where the system's goals may not align fully with educational objectives, despite companies' efforts to mitigate such risks through fine-tuning and reinforcement learning (?). Concerns also extend to students' digital well-being, as AI in education could inadvertently contribute to unhealthy digital habits and reliance on AI assistance over critical thinking (?). Given these multidimensional impacts, further research is needed to optimise AI's role in education, balancing~~ ?). However, most studies are short-term~~learning gains~~

~~with long-term skill development and ethical considerations~~, may be subject to publication bias, and often lack a strong theoretical foundation.

~~Research in the field of human-AI combinations does not indicate consensus in terms of its effectiveness. For example, using a~~ Despite promising results, key limitations remain. For instance, Er et al. (?) found that students in a programming class performed better with instructor feedback, which was seen as more useful and fair than AI-generated feedback, highlighting the current gap between AI and human instruction. Similarly, evidence of performance boosts from non-education settings is mixed. A meta-analysis ~~of studies published between 2020 and 2023,~~ by Vaccaro, Almaatouq, and Malone (?) ~~finds that , on average,~~ found that human-AI ~~combinations performed significantly worse than the best of~~ teams often performed worse than either humans or AI alone~~. On the other hand,~~ , suggesting a more nuanced relationship. In education, additional concerns are that integrating AI might encourage cheating, over-reliance, or superficial understanding (?; ? ; ?; ?), a risk that is difficult to evaluate without longitudinal studies.

The effectiveness of AI in education often depends on students' attitudes and perceived usefulness of AI tools, understanding which can guide curriculum design. For instance, Li et al. (?) ~~uses a sample of 200 Chinese studentsto show significant correlations between human-computer interactions, perceived usefulness and students' skills. Understanding these associations can help institutions design better AI courses and more effective integration of AI in education~~found that K–12 students' views on AI tools were linked to factors like readiness, confidence, and anxiety. Similar concerns have been noted from teachers' perspectives as well (?). Our study builds on this by exploring how interaction styles, such as Socratic vs. non-Socratic, shape students' perceptions of AI helpfulness.

This study advances ~~this~~ the literature on human-AI collaboration in educational contexts by examining the role of step-by-step reasoning and the impact of different modes of AI interaction. Additionally, it contributes to ongoing research exploring the factors that influence students' engagement with AI. Gaining insights into the value these combinations bring is crucial for developing practical guidelines for schools and educational institutions.

~~Results of our investigation also contribute to the current debate on the pedagogical principles for designing AI-driven systems and educational opportunities, the so-called *design for learning* (?). It underscores~~ Most existing research focuses on general AI chatbots, like ChatGPT, which are not specifically designed for students and often lack clear pedagogical foundations. Our study highlights the importance of ~~using a multi-disciplinary approach that combines traditional pedagogical insights with principles from human-computer interactions. Consequently, our work adds~~ purposeful design and pedagogical goals, contributing to the growing ~~efforts to tackle critical challenges in integrating AI into education , helping to automate and scale up tasks like providing feedback, grading, and making learning recommendations (~~field of AI chatbot design for education or design for learning (?). This research calls for a multidisciplinary approach combining pedagogy and human-computer interaction. Building on recent efforts (?; ?; ?)~~. ~~, we examine how k–12 students interact with different AI chatbot styles. By doing so, the study underscores both the potential and challenges of applying classical pedagogical methods, such as the Socratic approach, emphasizing the need to provide explanations rather than just answers.

2 Methods

We conducted a field experiment across two schools to investigate the impact of different AI tutoring strategies on students' critical thinking, self-confidence, and task performance. Participants engaged with a custom-designed web application, the *AI Tutor*, which randomly assigned students to interact with distinct versions of GPT-4-based chatbots during structured educational activities. The experimental design featured two primary manipulations: (1) the presence of step-by-step reasoning by the AI and (2) the use of Socratic versus non-Socratic dialogue styles. To assess the effects, we employed objective performance measures alongside pre- and post-task surveys capturing students' cognitive and affective responses. The following subsections detail the AI Tutor system, participant characteristics, and experimental procedures.

2.1 AI Tutor

To assess the impact of LLMs on students' critical thinking, we developed an *AI Tutor*, a web-based application built using Python with the Flask framework. The application integrates OpenAI APIs, allowing students to interact with the GPT-4.0 model while performing different educational tasks. This app's key features include:

- **User authentication**: A secure and anonymous login system for students.
- **Web Survey**: A web interface to ~~survey studentsand where they could perform simple tasksonline~~collect data from students, including answering survey questions and performing simple tasks, such as writing a short essay or ~~performing basic calculations~~solving simple math problems.
- **AI Chatbot**: A chatbot powered by GPT-4 provided students with personalised assistance and tutoring during the session. A new chatbot instance was associated with each question for each student, isolating each conversation from interactions in previous questions. The chatbot was available to students only for specific questions or tasks under the researchers' control.
- **Data logging**: A SQL database storing students' conversation logs with the AI chatbot, including texts, timestamps, and survey responses.
- **Randomised assignment**: A system to randomly assign registered students to different versions of the AI chatbot or treatment groups to explore the causal impact of various

configurations on students' interactions and performance.

2.2 Participants

We recruited N=122 students between 14 and 16 years old enrolled in secondary education from two schools: one in Brussels (Belgium) and one in Seville(Spain).located in Brussels, Belgium, and the other in Seville, Spain.[3] The experiment was conducted during school hours at the participating schools (Fig. ??). Both schools are bilingual institutionssecondary institutions that follow a European curriculum framework, with English as a primary language , ensuring participants had a high English proficiency. The the primary language of instruction across most subjects. The student populations at both schools are culturally and linguistically diverse, including a mix of local and expatriate families. Many students come from middle- to upper-middle-class socioeconomic backgrounds and are accustomed to using digital technologies in their academic work: both schools are equipped with computer laboratories and digital interactive whiteboards for classroom instruction.

All participants had appropriate levels of English proficiency, given the language policy of the schools. Even so, the AI Tutor was multilingual, allowing students to interact with either it either in English or in their native language. The experimental instructions were in English to ensure consistency across sites. However, if needed, students could translate the instructions using the browser's automatic translation service.

Recruiting students was conducted

2.2.1 Ethical Approval and Data Privacy

We recruited students after receiving ethical approval from the [redacted]internal Ethical Review Board, ensuring that the consent procedures, data protection requirements, and the experimental protocol complied with local laws and were safe for the participantsethical research standards. The data privacy protection protocol was approved by the [redacted] data protection officer. As an additional safeguard, given the minor age of the participants, we required their parents or legal guardians to provide us with written consent, allowing their children

[3] 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.

to participate in the study. The students also had to give their assent to participate by completing an online consent form.

2.3 Experimental sessions

About ten experimental sessions were conducted at the schools between November 11 and 12, 2023, in Brussels and February 8 and 9, 2024, in Seville. Each session took about two hours. In the first 45-60 minutes, students received a personal computer and were asked to perform ~~multiple tasks online~~ three main tasks using the AI tutor ~~, including answering a questionnaire~~ and answer a questionnaire without the AI tutor. In the following 45-60, there was a group discussion on how the students ~~found~~ judged the interactions with the AI tutor and a more general debate on how they perceived the potential benefits and drawbacks of integrating LLMs in the classroom. The results of the group discussions are not discussed in this paper; however, they were used as material for interpreting the results of the experimental study.

2.4 Experimental Conditions

Figure **??** illustrates the two randomised manipulations in our experimental design: (1) AI step-by-step reasoning and (2) Socratic vs non-Socratic AI. It is important to note that these were conducted as independent experiments, with different dependent and independent variables, rather than a factorial 2x2 design.

~~AI step-by-step reasoning~~

2.4.1 AI step-by-step reasoning

The first manipulation focuses on a task in which students are asked to estimate the value of coins in a jar (see Section **??** for the details).[4] Specifically, we used the coin jar from Steiner's experiment (**?**), aimed originally at assessing Internet users' guessing accuracy. For this intervention, we varied the AI's response by providing either a complete answer, which included both an estimated value and a step-by-step explanation generated by the AI tutor, or a partial answer that provided only the estimate without additional details. In both conditions, all participants received the same

[4]This is a common experimental activity in economics, especially in the context of auction theory (**?**). The task is ideal in our setting because it requires participants to guess based on limited information, thus creating a situation where AI-generated assistance could potentially influence their decisions.

estimated value ($213), but those in the full-answer condition also viewed an explanation outlining the step-by-step reasoning the AI used to arrive at this estimation based on the jar image.[5]

This setup allows us to examine how AI-provided explanations affect students' performance and their perceptions of the AI's accuracy. Specifically, we focus on three outcome variables: (1) the propensity to update their initial guess~~and size of their updates~~, (2) the accuracy of students' final estimations, measured as the (absolute) difference between their final guesses and the actual coin value, and (3) students' perceived accuracy of the AI, rated on a scale from low to high. We also asked participants (4) to rate the perceived accuracy of the mean guess among 600 people ($596), which exaggerated the correct value, as reported in the original article (**?**).

~~**Socratic vs Non-Socratic AI**~~

2.4.2 Socratic vs Non-Socratic AI

The second manipulation involved randomly assigning students to one of two different types of AI tutors: Socratic or non-Socratic AI tutors. Both tutors were powered by the same underlying large language model (GPT-4). Still, each was instructed with different "system messages" to create different behaviours, as illustrated in Table **??**.[6] For the Socratic tutor, the system message asked the model to engage students with open-ended, thought-provoking questions, encouraging them to think critically about their responses. In contrast, the non-Socratic AI tutor was instructed to provide concise, direct answers without necessarily engaging in deeper dialogue or posing further questions.

This setup ~~allowed~~ allows us to investigate the impact of different pedagogical AI tutoring approaches on students' performance and perceptions. Specifically, we asked students to use the AI tutor while performing three tasks: guess an unknown quantity ("How much water in litres do students consume at our school each week?")~~;~~, express an opinion and write a short essay ("What is your opinion about the effect of social media on teenagers?"); respond to physics questions on how sound propagates in different media ("In which of the following materials does sound travel faster?"). These questions enabled us to assess the AI tutor's impact on students' learning and

[5] Using GPT 4.0, we uploaded the image of the coin jar asking for an estimate of the value of coins ~~for~~ ten times. We then selected the median response for the experiment.

[6] An AI's system message guides how the AI interprets the conversations by setting ~~paramters~~ parameters for interaction.

problem-solving abilities.

We focused on two primary metrics: (1) confidence in their answers ~~,~~ and (2) perceived usefulness of interacting with the AI tutor. Specifically, students were asked, “How confident are you that the answer you provided is accurate?” on a five-point scale ranging from “not confident at all” to “very confident.” They were also asked, “How helpful was it to interact with the AI tutor?” This was also rated on a five-point scale, from “Not at all helpful” to “Very helpful.” Only for task 3, we had an additional metric, which was the correctness of their answers.

2.5 ~~Background Measures~~Student Characteristics

~~As an initial step in our analysis, and to~~ To examine student interactions with AI and their potential impact on learning and critical thinking, we collected a range of self-reported and behavioural data. The full questionnaire is provided in the Appendix (Section **??**).

2.5.1 AI Attitudes and Usage

To better understand students' ~~awareness of and~~ general attitudes toward AI, we ~~developed a brief four-question “AI in Education Attitudes Scale.” This scale was adapted from the broader AI Attitudes Scale proposed by~~ selected four questions from Schepman and Rodway's AI Attitudes Scale (**?**). These questions ~~included: Do you agree or disagree that society will benefit from a future of AI? Do you agree or disagree that AI is dangerous? Do you agree or disagree that AI will foster students' learning in the future? Do you agree or disagree that AI is often misused by students ?~~ explored students; beliefs about AI's societal benefits, potential dangers, capacity to support learning, and the likelihood of misuse by students (Appendix, Section **??**). We also asked students about their direct experience using ChatGPT for homework — the most common chatbot at the time of the study — and their beliefs about how many of their peers use ChatGPT for their homework, thereby capturing both individual behaviour and the perceived social norm around AI usage. These measures help establish baseline orientations that may moderate how students engage with and evaluate AI explanations (related to RQ-1) and how they perceive different AI interaction styles (related to RQ-3).

~~Finally, to consider participants' academic performance and skills, participants were also asked about their school grades and~~

2.5.2 AI Academic Habits and Skills

Academic skills or habits may moderate the effect of AI interactions on problem-solving activities. Accordingly, to account for participants' heterogeneity in academic skills, we asked students to report their average school grades within five categories. We also asked students about their academic habits or the challenges they face at school, such as how often they complete their homework assignments on time and what factors affect their ability to do so. ~~Self-reported experience using ChatGPT (the main LLM available at the time of the study) was also asked, including a question on their estimation of how many of their peers use ChatGPT for their homework.~~

~~Section ?? presents the complete questionnaire.~~

2.5.3 Student-AI Interaction Metrics

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 behavioural metric, it does not capture the quality of cognitive engagement or helpfulness, limiting its direct relevance to RQ-1 and RQ-2. 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 RQ-1) and whether Socratic dialogue promoted critical thinking or confidence (for RQ-2).

3 Results

3.1 Overview of Student Demographics

A total of 122 students participated in the study, 64 in Brussels and 58 in Seville. The sample was gender balanced. Over half of the students reported B+ grades and prior use of ChatGPT, with boys more likely to report ChatGPT experience. Around 50% reported to always complete homework on time, with a minority reporting less than always, with factors like time management (17% in Brussels, 25% in Seville) and material comprehension (56% in Brussels, 28% in Seville) affecting completion. Students' self-efficacy varied by task, with 30% of the students feeling they could "easily" write a technology essay (task 2) but only 6% and 11% feeling confident about explaining sound propagation (task 3) or solving a numerical estimation task such as guessing the litres of water in a pool (task 1). Thus, the sample was substantially homogeneous in terms of demographics but diverse in terms of self-efficacy.

Table **??** illustrates the distribution of student characteristics ~~according to the treatment assignment for the Socratic versus Non-Socratic AI tutor. (A similar table for the AI step-by-step reasoning assignmentcan be shown.)~~ across treatment groups alongside the results of separate Fisher's Exact Tests on the association with treatment assignment. The sample characteristics were generally balanced across ~~treatments~~treatment groups. Only one variable out of ten ~~associations with the treatment assignment was statistically significant for~~ – i.e., self-reported difficulties in homework – was statistically associated ($p = 0.05$) with the Socratic/Non-Socratic ~~conditions, and~~ assignment, while no significant associations were found for the ~~AI Step-by-step reasoning~~ AI-reasoning assignment. Thus, only one out of twenty Fisher's exact tests showed a significant result~~($p < 0.05$)~~, indicating a good balance across treatments.

3.2 Attitudes Towards AI

As illustrated in Figure **??**, students expressed conflicting views on the role of AI in education. On the one hand, a majority felt that AI is often misused by students (65%) and potentially dangerous (57%). On the other hand, most students also anticipated that AI would enhance student learning in the future (59%) and contribute positively to society (65%). These findings suggest that most students in our sample are aware of the risks of misuse and safety with a prevailing optimism that

AI could support educational growth and societal progress.

3.3 The Impact of AI Step-by-step Reasoning Exposure

~~We tested whether showing students~~ Nearly all students revised their initial estimates (110 out of 122), with no significant difference between treatment groups (Fisher's test, p = 0.9). However, presenting students with AI-generated step-by-step reasoning ~~affected how they~~ may have influenced the accuracy of their revised estimates.

To address RQ1, we examined whether the treatment—showing step-by-step AI reasoning—affected how students used the AI's prediction, compared to ~~just~~ seeing the prediction ~~without any reasoning. We measured students' accuracy using the absolute difference between each~~ alone without any accompanying explanation. For each student $i = 1, \cdots, 122$, we computed the absolute error as the absolute difference between student i's ~~prediction~~ guess, Guess_i, and the actual value of coins:

$$\text{absolute error}\,\text{Absolute Error}_i = |\text{guess}|\ \text{Guess}_i - \text{actual value}|\text{Actual Value Coins}\ |,$$

~~for $i = 1, \cdots, N$,~~ where a smaller absolute error indicated greater accuracy.

~~The absolute error distribution in our sample~~ Figure **??** 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.[7] This procedure involved resampling participants' absolute errors (n = 1,~~complicating the testing for mean differences. Consequently, we shifted our analysis towards treatment differences in the median absolute error, as illustrated in Figure **??**. We applied bootstrap resampling~~ 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 ~~estimate the median difference across treatment groups and the corresponding confidence intervals.~~[8] ~~Results indicated that, although both groups received the same exact AI~~

[7] We detected one outlier in the data, which was removed from the bootstrap analysis: one student provided an unusual and notably high guess of 6969. However, this value deviated substantially from both the overall distribution and the student's initial guess, suggesting a mistake rather than an estimate.

[8] ~~Outliers were detected in the data and removed from the bootstrap analysis. One student provided a notably high guess of 6969. However, this value deviated substantially from both the overall distribution and the student's~~

~~estimate of $213, students who also received~~ 70, indicating a greater accuracy (lower error) for the students exposed to AI-generated ~~step-by-step reasoning demonstrated a significantly higher accuracy , reflected by a lower median of the absolute error($p = 0.08$). This evidence indicates that exposure to AI reasoning contributed to an improvement in student performance.~~ reasoning (one-sided, $p < 0.05$).

~~We subsequently examined students' perceptions of the AI 's prediction accuracy on a~~ To examine the impact of AI explanations on students' perceived credibility of AI predictions (RQ2), we analysed students' five-point ~~rating~~ratings of the perceived accuracy of the AI estimated value of coins. We noticed that 53% of students who did not receive step-by-step reasoning considered the AI estimate as either "good" (39%) or "very good" (14%) compared to 43% of students exposed to the AI reasoning (19% and 24%, respectively). This gap of ten percentage points suggests that students viewed the AI as more accurate when the AI-generated guess was presented as a "black box." However, we didn't have enough observations to reach a statistically significant association (Fisher's test, $p = .27$), and even regression analysis, controlling for individual characteristics, showed no significant association (Figure **??**).

Conversely, exposure to the AI step-by-step reasoning ~~significantly influenced students' perceptions of the~~ seemed associated with students' perception of the accuracy of the "human" ~~estimate of $596. This value was~~ guess — the average guess ~~of~~ from 600 participants ~~in the initial~~ reported in the original study (**?**)~~, which exceeded the correct value ($379.54) by about the same amount as the AI guess underestimated it ($213).Notably,~~ .[8] While all groups received the same human and AI estimates, ~~but while~~ 64% of students in the control ~~group~~ rated the human estimate as "poor" or "very poor~~," only~~" against 45% of students exposed to AI reasoning~~did so~~. Although this difference was ~~marginally significant~~ insignificant (Fisher's exact test, $p = 0.15$), regression analysis, controlling for ~~student's location, gender, their initial estimate~~ fixed differences across schools and across individuals, such as their gender, the initial guess of the value of coins, and experience with ChatGPT, ~~confirmed that AI exposure led students to rate the human estimate about half a point more favourably~~ showed a significant effect ($p < 0.05$) of the assignment to AI reasoning on students' perceptions of the human guess, as shown in Figure **??**. This finding underscores the complex relationship between AI and learning. It suggests that students exposed to the AI's

~~initial guess, suggesting an error rather than an informed estimate.~~

[8] Notably, this guess was $596, exceeding the correct value ($379.54) by about the same amount as the AI guess underestimated it ($213).

step-by-step reasoning may have identified minor errors, thus increasing their expectations of human accuracy, and those who viewed AI as a "black box" tended to underrate human estimates.

3.4 A Comparison of Socratic vs Non-Socratic AI

~~**Student-AI Interactions**~~

3.4.1 Student-AI Interactions

Figure **??** shows the students' message length and frequency in the Socratic AI and Non-Socratic AI groups, allowing us to compare the student-AI interactions across treatment groups. As shown in the figure, Socratic AI students exchanged a median of 20 messages, significantly more than the eight messages by non-Socratic AI students (Wilcoxon test, $p < 0.01$). In addition, the Socratic AI's messages were significantly shorter, with a median of 42 words, compared to 123 for the non-Socratic tutor (Wilcoxon test, $p < 0.01$). Socratic students also used fewer words, with two peaks: one at ten and the other at one word. This evidence supports the hypothesis that the Socratic tutor encouraged more engagement and interaction, resembling a relatively more genuine student-tutor talk.

~~**Students' Confidence in Their Responses**~~

3.4.2 Students' Confidence in Their Responses

~~We hypothesised Socratic AI would promote a deeper understanding of the task, raising~~ To test whether Socratic AI increased students' confidence in their answers ~~. To test this hypothesis~~(RQ4), we examined students' self-reported confidence ~~using a five-point scale~~ levels at the end of each task. As shown in Figure **??**, the observed differences in confidence ~~levels~~ between the treatment groups were minimal. ~~Among Socratic students~~Specifically, 25% of Socratic students reported feeling "very confident," and 33% ~~felt~~ reported feeling "confident" compared to 21% and 38%, respectively, among non-Socratic students. These differences were not statistically significant, even with regression analysis controlling for ~~student-task differences~~ variation across students and tasks (Figure **??**). Thus, our analysis found no evidence of treatment effects on students' confidence ~~levels~~in their answers.

Further explorative regression analysis to examine potential treatment interactions shows that

Socratic AI students with prior ChatGPT experience reported significantly less confidence (p < 0.1) than Non-Socratic AI students (see Figure ??). This explorative finding suggests that experienced users may perceive new or unconventional AI tutoring methods as less effective or even counterproductive, indicating that encouraging the use of such AI tutoring tools among experienced ChatGPT students can be challenging.

~~Perceived Helpfulness of AI Tutor~~

3.4.3 Perceived Helpfulness of AI Tutor

~~We examined~~ To test differences in students' perceptions of the AI tutor's helpfulness (~~How helpful was it to interact with the AI tutor?)~~ RQ5), we examined students' ratings of helpfulness of the AI interaction at the end of each task. These ratings were on a five-point scale, from "Not at all helpful" to "Very helpful."[9]

In both treatment groups, many students found interacting with the AI tutor "very helpful" or "helpful", with 56% of the non-Socratic and 44% of the Socratic students, as shown in Figure **??**. However, the Socratic treatment group showed a bimodal distribution, with a substantial fraction of students (21%) finding the interaction "not at all helpful." The association between perceived helpfulness and treatment assignment was statistically significant (Fisher's exact test, $p = 0.025$), providing evidence that Socratic AI was less helpful to certain students. This association, however, was stronger for Task 1 and Taks 3 than Task 2, suggesting an association between the perceived AI's helpfulness and the type of task. Ordinal logistic regression accounting for individual student and task characteristics, including self-efficacy per task, revealed a statistically significant negative difference that corresponds to a drop of approximately 13 percentage points in perceived helpfulness associated with the Socratic AI, as illustrated in Figure **??**. See Section **??** for more details.

~~Learning Outcomes and Knowledge Retention~~

[9] Due to a coding issue, the scale used in Brussels and Seville differed slightly in the labels: Seville's students saw "Extremely helpful" whereas Brussels students saw "Very helpful." However, this issue has a minimal impact on the overall results, as results focus on the negative labels ("not at all helpful" or "not helpful") and they remain the same even after controlling for location effects in a regression. Additionally, open discussions held with students after the experimental session confirmed that they found the Socratic AI less helpful. This feedback, combined with the observed differences in negative labels, reinforces that the coding discrepancy has a minimal impact on the overall findings.

3.4.4 Learning Outcomes and Knowledge Retention

To evaluate the impact of learning (RQ3), we compared the correctness of responses to Task 3, which focused on sound propagation, and assessed knowledge retention using a follow-up question on the same topic without AI assistance. As shown in Figure **??**, the use of AI significantly enhanced students' response accuracy. Before interacting with the AI tutor, only 32% of students correctly answered that sound travels faster in water than air due to water's higher density. After AI interaction, this percentage nearly doubled increasing to 68%. However, there was no significant difference in learning outcomes between the Socratic and non-Socratic AI approaches, as illustrated in the figure. Moreover, when presented with a follow-up question (asking in which medium sound travels fastest among gold, rubber, warm air, cold air, and water), only 18% of students responded correctly by selecting the denser material (i.e., gold), again with no significant difference across treatments. This result suggests two implications. Firstly, we found no evidence of Socratic AI improving learning. Secondly, our results ~~confirmed~~ seems to suggest a problem of limited retention of the insights obtained with AI assistance or difficulty in applying such learning to novel scenarios.[10]

[10] 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 students' outcomes.

4 Discussion

~~The current results indicate several important implications~~This study contributes to our understanding of the impact of pedagogically-aligned configurations of an LLM-based tool on the learning and attitudes of high school students, with a special focus on their critical thinking.

4.1 Overview of the findings

With the rapid adoption of AI in education, the focus on students' competencies and skills, such as critical thinking, has become particularly urgent (?; ?). Despite the rapidly increasing body of research on the impact of LLMs on students' learning, most studies focus on participants in higher education (e.g., ?), which may not generate insights applicable to younger students with different cognitive maturity. In addition, despite the increasing research interest regarding the impact of LLMs on students' learning, there is little consensus among researchers, mainly due to students' over-reliance on AI tools (?).

The results of this study focus on the role of two interventions in configuring an LLM chatbot for educational support: (i) step-by-step explanations and (ii) the Socratic method with guiding questions. First, we found that students significantly benefit from AI-generated step-by-step reasoning accompanying solutions, mainly when performing open-ended tasks like estimating unknown quantities (RQ1). This finding aligns with expectations of current research on the use of LLM chatbots and their ability to provide explanations for learning purposes (?). It also aligns with previous research demonstrating that LLMs can enhance student performance ~~and that explanations help children develop critical thinking~~(?). However, our results underscore that the key mechanism driving AI improvements is the step-by-step reasoning provided by the AI, which allows students to understand better and engage with problem-solving. ~~This insight suggests that teachers should focus on educating students to enhance their ability to evaluate and judge the correctness of AI-generated reasoning.~~

Furthermore, our study reveals that step-by-step reasoning not only helps students in solving problems but also enhances students' ability to evaluate AI-generated information critically ~~. This~~ (RQ2). Critical thinking is indeed a complex cognitive process that involves the evaluation and analysis of a given information (?). The mobilisation of students' critical thinking in the specific task was evidenced by the more positive evaluations of human-generated guesses compared to AI-generated solutions, suggesting that students could better assess and challenge AI predictions. This contribution extends the existing literature by showing that AI can foster critical thinking and analytical skills when coupled with transparent reasoning ~~, instead of~~ rather than being presented as a “black box~~.”~~” (?).

In addition, we compared various ways to structure student-AI interactions and, more specifically, the effectiveness of Socratic AI (an interactive, questioning-based AI) with non-Socratic AI. Our results showed that Socratic AI was more engaging and promoted greater interaction, aligning with the notion that AI-student interactions should be dynamic and dialogical (?). However, contrary to our expectations, we found no significant differences in students’ self-reported confidence in the accuracy of their answers or in the correctness of their responses despite the more frequent interactions with the Socratic AI (RQ4). Additionally, the Socratic AI’s perceived helpfulness was rated lower compared to the non-Socratic AI (RQ5). These results cast some doubts on the effectiveness of Socratic AI in short-term tasks or, more broadly, the effect of certain kinds of AI-student interactions. As such, existing pedagogical practices extensively used in human-human interaction might not always work in human-AI interaction. ~~This underscores that new pedagogical paradigms are needed to integrate AI~~ In addition, while recent research attempts to understand the quality of Socratic LLMs for critical thinking (e.g., ?), these studies use simulations without experiments with human participants. Our results underscore the need for adapting existing pedagogical paradigms or proposing new ones for integrating LLM tools into pedagogical practices effectively.

~~Our~~ Interestingly, our findings indicate that simply interacting with AI cannot promote meaningful and lasting learning (RQ3). In our initial test on sound propagation, approximately half of the students could revise their initial answers based on AI interactions, improving their response accuracy from 30% to 70%. However, in a verification task where students had no access to AI, the majority failed to identify the correct answer, mistakenly claiming that sound propagates faster in water than in gold. Most students exhibited this misunderstanding, which supports key concerns that AI-generated answers alone ~~may not facilitate effective learning , and that~~ do not scaffold effective learning and the mere implementation of Socratic AI does not mitigate this risk.

Our results suggest that AI has great potential as an educational tool, but its implementation requires careful consideration. First, students with prior experience in using AI tools may not readily adopt new pedagogical approaches, especially if they perceive them as less effective than commercially available alternatives. Second, the effectiveness of the pedagogical approach may vary depending on the nature of the task, complicating the design of a one-size-fits-all solution for AI-assisted learning.

4.2 Limitations of the study and future work

Several limitations of our study should be acknowledged. The small sample size limits the generalizability of our findings, though the controlled environment and the use of multiple tasks help mitigate potential noise in the data. Also, our experiment focused on short-term results. Although ~~short term~~ short-term results are important to foster adoption, it is unclear the effectiveness of AI in the ~~long-term.~~ long term. 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. Additionally, our study combined objective performance metrics with self-assessed ratings of helpfulness and confidence. However, individual perceptions do not necessarily reflect actual learning (?). Further research is needed to validate our findings using objective learning measures.

Finally, we conducted our study with a cohort of students possessing strong English proficiency and a clear understanding of the limitations of AI, which may not be representative of the general student population. Additionally, we focused on only two schools, which allowed us to control for school-specific fixed effects. However, we recognise that the impact of the treatments may differ across schools, and a broader investigation involving many more institutions would be necessary to explore this variability. Furthermore, the experiment was carried out at school and in a secure and anonymous digital environment, with a robust protocol developed to ensure the safe and ethical handling of AI-based interactions in experimental settings. However, it remains to be seen if the results of our analysis will remain when students use AI in the field.

Our future work includes a large-scale study with a representative sample in a country in Europe, where we aim to address the above-mentioned limitations.

4.3 Concluding remarks

While our study focused on comparing a fairly general pedagogical approach—Socratic AI, future research should explore ~~the effectiveness of~~ this pedagogical method in a more nuanced way and consider alternative pedagogical approaches ~~and~~ to facilitate the scalability of AI tools across diverse tasks and educational contexts. Yet, our findings have important implications for designing AI tutors, highlighting the importance of providing transparent AI-generated step-by-step reasoning and the challenges of fostering learning through guided AI-student interactions. Therefore, ~~our study suggests that~~ AI systems must engage students interactively and ~~foster~~ provide learning opportunities for students critical thinking and problem-solving skills to maximise their educational value. Future developments should focus on refining the integration of AI-generated reasoning and ensuring that AI tools are adaptable to various learning tasks and students' needs.

References

Bearman, Margaret, and Rola Ajjawi. 2023. "Learning to Work with the Black Box: Pedagogy for a World with Artificial Intelligence." *British Journal of Educational Technology* 54 (5): 1160–73.

Cleveland, Julie. 2015. *Beyond Standardization: Fostering Critical Thinking in a Fourth Grade Classroom Through Comprehensive Socratic Circles*. Arizona State University.

Dai, Wei, Jionghao Lin, Hua Jin, Tongguang Li, Yi-Shan Tsai, Dragan Gašević, and Guanliang Chen. 2023. "Can Large Language Models Provide Feedback to Students? A Case Study on ChatGPT." In *2023 IEEE International Conference on Advanced Learning Technologies (ICALT)*, 323–25. IEEE.

Dalim, Siti Fairuz, Aina Sakinah Ishak, and Lina Mursyidah Hamzah. 2022. "Promoting Students' Critical Thinking Through Socratic Method: The Views and Challenges." *Asian Journal of University Education* 18 (4): 1034–47.

Danovitch, Judith H, Candice M Mills, Kaitlin R Sands, and Allison J Williams. 2021. "Mind the Gap: How Incomplete Explanations Influence Children's Interest and Learning Behaviors." *Cognitive Psychology* 130: 101421.

Darvishi, Ali, Hassan Khosravi, Shazia Sadiq, Dragan Gašević, and George Siemens. 2024. "Impact of AI Assistance on Student Agency." *Computers & Education* 210: 104967.

Debets, Tim, Seyyed Kazem Banihashem, Desirée Joosten-Ten Brinke, Tanja EJ Vos, Gideon Maillette de Buy Wenniger, and Gino Camp. 2025. "Chatbots in Education: A Systematic Review of Objectives, Underlying Technology and Theory, Evaluation Criteria, and Impacts." *Computers & Education*, 105323.

DeJong, Gerald, and Raymond Mooney. 1986. "Explanation-Based Learning: An Alternative View." *Machine Learning* 1: 145–76.

Derakhshan, Ali, and Farhad Ghiasvand. 2024. "Is ChatGPT an Evil or an Angel for Second Language Education and Research? A Phenomenographic Study of Research-Active EFL Teachers' Perceptions." *International Journal of Applied Linguistics*.

Duelen, Aline, Iris Jennes, and Wendy Van den Broeck. 2024. "Socratic AI Against Disinformation: Improving Critical Thinking to Recognize Disinformation Using Socratic AI." In *Proceedings of the 2024 ACM International Conference on Interactive Media Experiences*, 375–81.

Eke, Damian Okaibedi. 2023. “ChatGPT and the Rise of Generative AI: Threat to Academic Integrity?” *Journal of Responsible Technology* 13: 100060.

~~Gabriel, Iason. 2020.~~

Er, Erkan, Gökhan Akçapınar, Alper Bayazıt, Omid Noroozi, and Seyyed Kazem Banihashem. 2024. “~~Artificial Intelligence, Values, and Alignment.~~ Assessing Student Perceptions and Use of Instructor Versus AI-Generated Feedback.” *British Journal of Educational Technology.*

Farrokhnia, Mohammadreza, Seyyed Kazem Banihashem, Omid Noroozi, and Arjen Wals. 2024. “A SWOT Analysis of ChatGPT: Implications for Educational Practice and Research.” ~~*Minds*~~ *Innovations in Education and* ~~*Machines*~~*Teaching International* ~~30~~ 61 (3): ~~411–37~~460–74.

Favero, Lucile, Juan Antonio Pérez-Ortiz, Tanja Käser, and Nuria Oliver. 2024. “Enhancing Critical Thinking in Education by Means of a Socratic Chatbot.” *arXiv Preprint arXiv:2409.05511.*

Ferrario, Andrea, and Michele Loi. 2022. “How Explainability Contributes to Trust in AI.” In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 1457–66.

Gašević, Dragan, George Siemens, and Shazia Sadiq. 2023. “Empowering Learners for the Age of Artificial Intelligence.” *Computers and Education: Artificial Intelligence*. Elsevier.

Hendrycks, Dan, 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*.

Henkel, Owen, Libby Hills, Adam Boxer, Bill Roberts, and Zach Levonian. 2024. “Can Large Language Models Make the Grade? An Empirical Study Evaluating Llms Ability to Mark Short Answer Questions in k-12 Education.” In *Proceedings of the Eleventh ACM Conference on Learning@ Scale*, 300–304.

Holmes, Wayne, Fengchun Miao, et al. 2023. *Guidance for Generative AI in Education and Research*. UNESCO Publishing.

Huang, Shih-Hong, Ya-Fang Lin, Zeyu He, Chieh-Yang Huang, and Ting-Hao Kenneth Huang. 2024. “How Does Conversation Length Impact User’s Satisfaction? A Case Study of Length-Controlled Conversations with LLM-Powered Chatbots.” In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, 1–13.

Joksimovic, Srecko, Dirk Ifenthaler, Rebecca Marrone, Maarten De Laat, and George Siemens. 2023.

"Opportunities of Artificial Intelligence for Supporting Complex Problem-Solving: Findings from a Scoping Review." *Computers and Education: Artificial Intelligence* 4: 100138.

Kalyan, Ashwin, Abhinav Kumar, Arjun Chandrasekaran, Ashish Sabharwal, and Peter Clark. 2021. "How Much Coffee Was Consumed During EMNLP 2019? Fermi Problems: A New Reasoning Challenge for AI," October. https://arxiv.org/pdf/2110.14207.pdf.

Kaplan-Rakowski, Regina, Kimberly Grotewold, Peggy Hartwick, and Kevin Papin. 2023. "Generative AI and Teachers' Perspectives on Its Implementation in Education." *Journal of Interactive Learning Research* 34 (2): 313–38.

Kasneci, Enkelejda, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, et al. 2023. "ChatGPT for Good? On Opportunities and Challenges of Large Language Models for Education." *Learning and Individual Differences* 103: 102274.

Kojima, Takeshi, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. "Large Language Models Are Zero-Shot Reasoners." *Advances in Neural Information Processing Systems* 35: 22199–213.

Lai, Emily R. 2011. "Critical Thinking: A Literature Review." *Pearson's Research Reports* 6 (1): 40–41.

Lam, Faith. 2011. "The Socratic Method as an Approach to Learning and Its Benefits."

Lara, Francisco, and Jan Deckers. 2020. "Artificial Intelligence as a Socratic Assistant for Moral Enhancement." *Neuroethics* 13 (3): 275–87.

Lee, Victor R, Denise Pope, Sarah Miles, and Rosalıa C Zárate. 2024. "Cheating in the Age of Generative AI: A High School Survey Study of Cheating Behaviors Before and After the Release of ChatGPT." *Computers and Education: Artificial Intelligence* 7: 100253.

Legare, Cristine H. 2014. "The Contributions of Explanation and Exploration to Children's Scientific Reasoning." *Child Development Perspectives* 8 (2): 101–6.

Li, Wei, Xiaolin Zhang, Jing Li, Xiao Yang, Dong Li, and Yantong Liu. 2024. "An Explanatory Study of Factors Influencing Engagement in AI Education at the k-12 Level: An Extension of the Classic TAM Model." *Scientific Reports* 14 (1): 13922.

Noroozi, Omid, Maryam Alqassab, Nafiseh Taghizadeh Kerman, Seyyed Kazem Banihashem, and Ernesto Panadero. 2025. "Does Perception Mean Learning? Insights from an Online Peer Feedback Setting." *Assessment & Evaluation in Higher Education* 50 (1): 83–97.

Noroozi, Omid, Saba Soleimani, Mohammadreza Farrokhnia, and Seyyed Kazem Banihashem.

2024. "Generative AI in Education: Pedagogical, Theoretical, and Methodological Perspectives." *International Journal of Technology in Education* 7 (3): 373–85.

Rai, Arun. 2020. "Explainable AI: From Black Box to Glass Box." *Journal of the Academy of Marketing Science* 48: 137–41.

Roll, Ido, and Ruth Wylie. 2016. "Evolution and Revolution in Artificial Intelligence in Education." *International Journal of Artificial Intelligence in Education* 26: 582–99.

Saxton, David, Edward Grefenstette, Felix Hill, and Pushmeet Kohli. 2019. "Analysing Mathematical Reasoning Abilities of Neural Models." *arXiv Preprint arXiv:1904.01557*.

Schepman, Astrid, and Paul Rodway. 2020. "Initial Validation of the General Attitudes Towards Artificial Intelligence Scale." *Computers in Human Behavior Reports* 1: 100014.

Song, Cuiping, and Yanping Song. 2023. "Enhancing Academic Writing Skills and Motivation: Assessing the Efficacy of ChatGPT in AI-Assisted Language Learning for EFL Students." *Frontiers in Psychology* 14: 1260843.

Steiner, Erik. 2015. "Turns Out the Internet Is Bad at Guessing How Many Coins Are in a Jar." *WIRED*. https://www.wired.com/2015/01/coin-jar-crowd-wisdom-experiment-results/.

Suriano, Rossella, Alessio Plebe, Alessandro Acciai, and Rosa Angela Fabio. 2025. "Student Interaction with ChatGPT Can Promote Complex Critical Thinking Skills." *Learning and Instruction* 95: 102011.

Thaler, Richard H. 1988. "Anomalies: The Winner's Curse." *Journal of Economic Perspectives* 2 (1): 191–202.

Tlili, Ahmed, Boulus Shehata, Michael Agyemang Adarkwah, Aras Bozkurt, Daniel T Hickey, Ronghuai Huang, and Brighter Agyemang. 2023. "What If the Devil Is My Guardian Angel: ChatGPT as a Case Study of Using Chatbots in Education." *Smart Learning Environments* 10 (1): 1–24.

Urban, Marek, Filip Děchtěrenko, Jiřı Lukavskỳ, Veronika Hrabalová, Filip Svacha, Cyril Brom, and Kamila Urban. 2024. "ChatGPT Improves Creative Problem-Solving Performance in University Students: An Experimental Study." *Computers & Education* 215: 105031.

Vaccaro, Michelle, Abdullah Almaatouq, and Thomas Malone. 2024. "When Combinations of Humans and AI Are Useful: A Systematic Review and Meta-Analysis." *Nature Human Behaviour*, 1–11.

Walczak, Krzysztof, and Wojciech Cellary. 2023. "Challenges for Higher Education in the Era of

Widespread Access to Generative AI." Article. *Economics and Business Review* 9 (2): 71–100. https://doi.org/10.18559/ebr.2023.2.743.

Walter, Yoshija. 2024. "Embracing the Future of Artificial Intelligence in the Classroom: The Relevance of AI Literacy, Prompt Engineering, and Critical Thinking in Modern Education." *International Journal of Educational Technology in Higher Education* 21 (1): 15.

Wei, Jason, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models." *Advances in Neural Information Processing Systems* 35: 24824–37.

Weidinger, Laura, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, et al. 2022. "Taxonomy of Risks Posed by Language Models." In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 214–29.

Wilberding, Erick. 2021. *Socratic Methods in the Classroom: Encouraging Critical Thinking and Problem Solving Through Dialogue (Grades 8-12)*. Routledge.

~~Williams, Ryan Thomas. 2024. "The Ethical Implications of Using Generative Chatbots in Higher Education." In *Frontiers in Education*, 8:1331607. Frontiers Media SA.~~

Wu, Rong, and Zhonggen Yu. 2024. "Do AI Chatbots Improve Students Learning Outcomes? Evidence from a Meta-Analysis." *British Journal of Educational Technology* 55 (1): 10–33.

Wu, Yi. 2024. "Critical Thinking Pedagogics Design in an Era of ChatGPT and Other AI Tools—Shifting from Teaching 'What' to Teaching 'Why' and 'How'." *Journal of Education and Development* 8 (1): 1.

Xing, Wanli, Yukyeong Song, Chenglu Li, Zifeng Liu, Wangda Zhu, and Hyunju Oh. 2025. "Development of a Generative AI-Powered Teachable Agent for Middle School Mathematics Learning: A Design-Based Research Study." *British Journal of Educational Technology*.

Yan, Lixiang, Samuel Greiff, Ziwen Teuber, and Dragan Gašević. 2024. "Promises and Challenges of Generative Artificial Intelligence for Human Learning." *Nature Human Behaviour* 8 (10): 1839–50.

Yan, Lixiang, Lele Sha, Linxuan Zhao, Yuheng Li, Roberto Martinez-Maldonado, Guanliang Chen, Xinyu Li, Yueqiao Jin, and Dragan Gašević. 2024. "Practical and Ethical Challenges of Large Language Models in Education: A Systematic Scoping Review." *British Journal of Educational Technology* 55 (1): 90–112.

Yusuf, Abdullahi, Nasrin Pervin, and Marcos Román-González. 2024. "Generative AI and the

Future of Higher Education: A Threat to Academic Integrity or Reformation? Evidence from Multicultural Perspectives." *International Journal of Educational Technology in Higher Education* 21 (1): 21.

Zhai, Chunpeng, Santoso Wibowo, and Lily D Li. 2024. "The Effects of over-Reliance on AI Dialogue Systems on Students' Cognitive Abilities: A Systematic Review." *Smart Learning Environments* 11 (1): 28.

A B

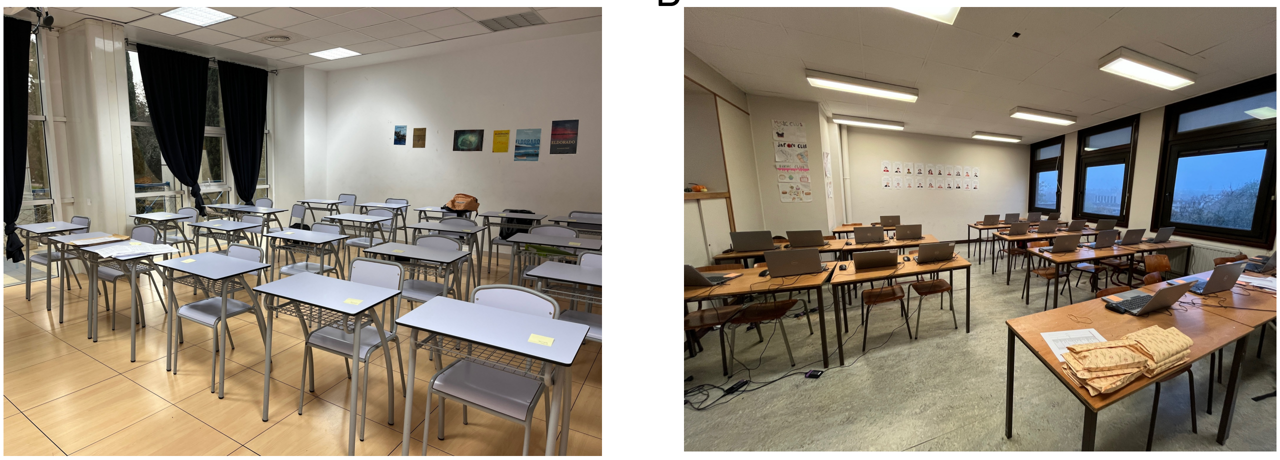


Figure 1: Classrooms used for the experiment in Seville (**A**) and in Brussels (**B**).

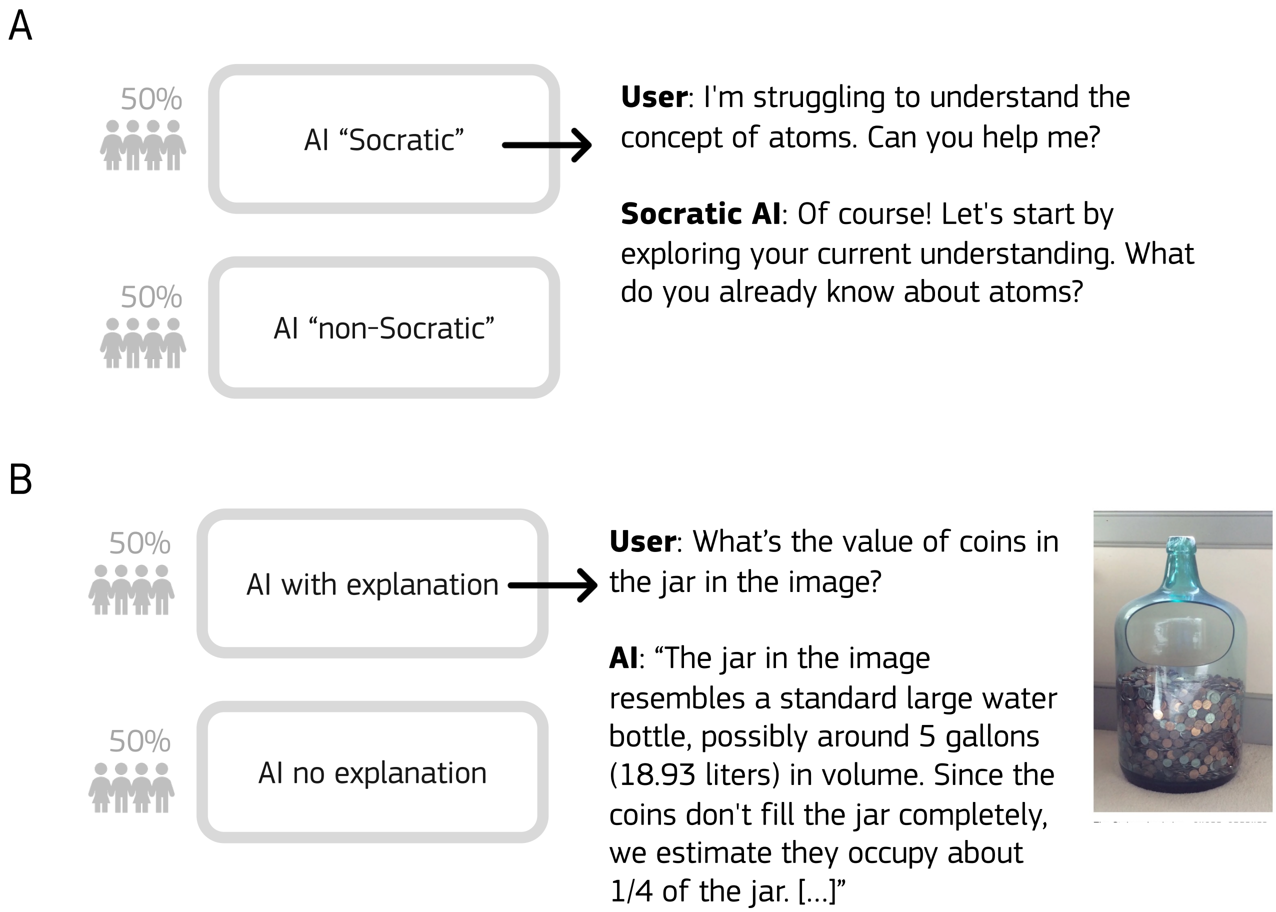


Figure 2: The experimental design involves two manipulations: **A.** Socratic vs Non-Socratic; **B.** AI solution with explanation vs AI solution (without explanation)

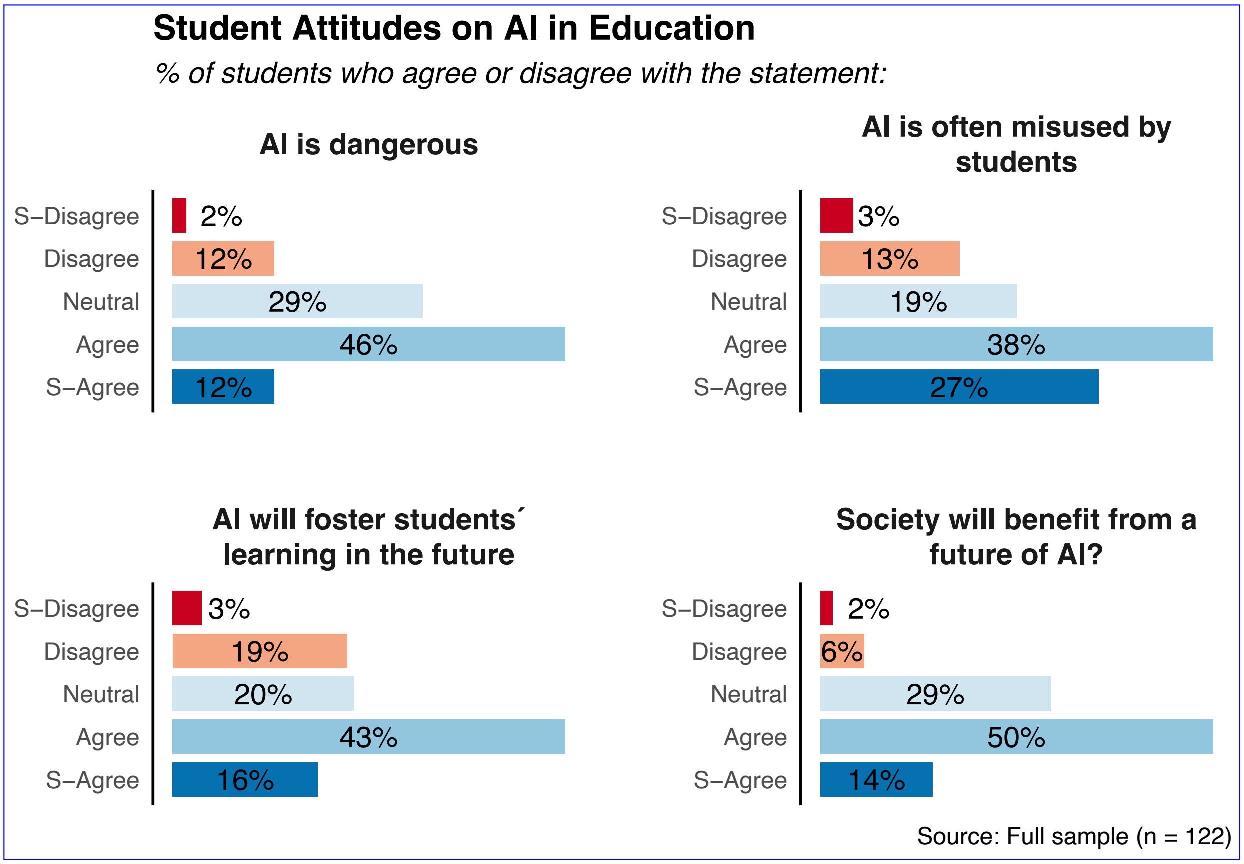


Figure 3: Student attitudes on AI in education

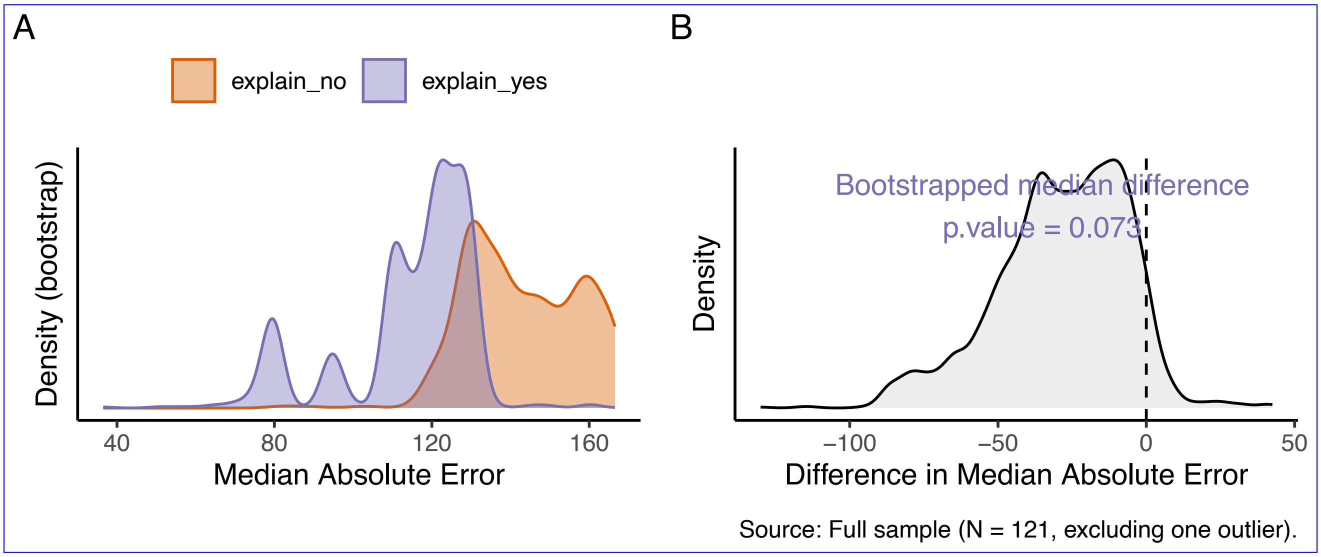


Figure 4: Comparison of Students' Absolute Error in Estimating Coin Jar Value with and without AI Explanations. All students received the AI-generated estimate of $213 (the correct value was $379.54). Still, those in the AI reasoning treatment also viewed the AI-generated step-by-step explanation for the estimation. Exposure to the AI-generated explanation significantly reduced the students' median absolute error.

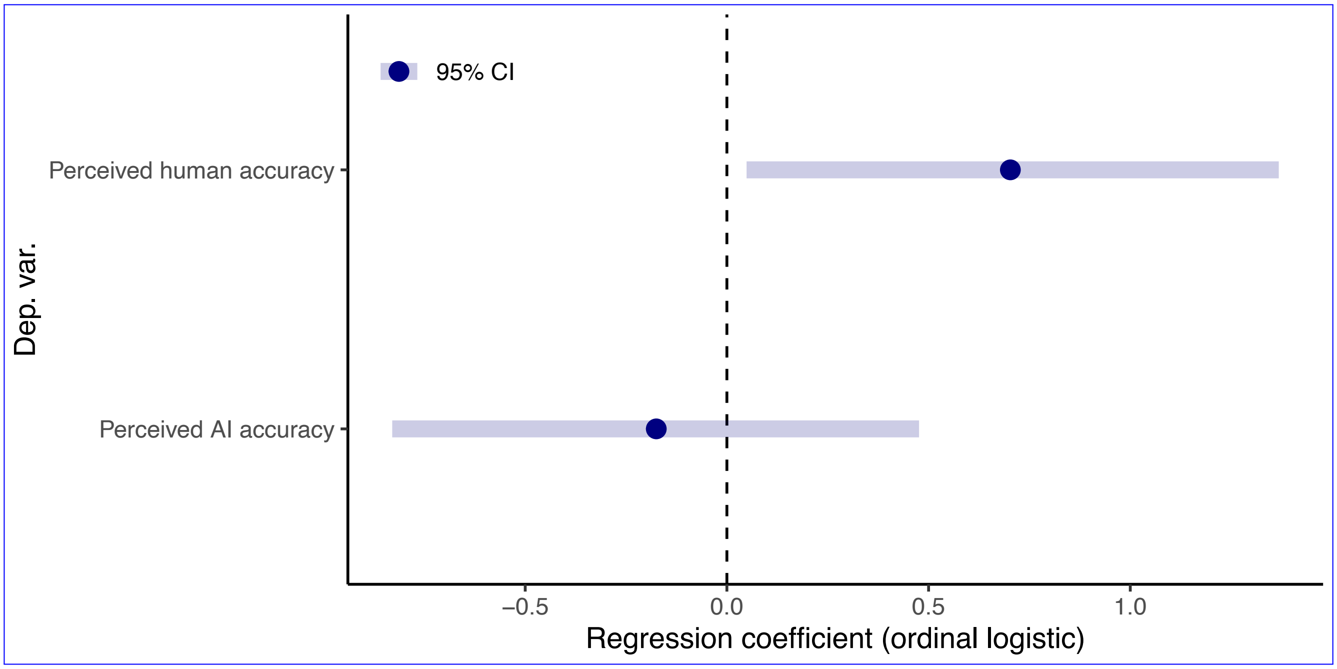


Figure 5: Comparison of students' perceived accuracy of AI and human estimates across treatments on a five-point scale. Coefficients from separate ordinal logistic regressions controlling for students' location, gender, and prior experience with ChatGPT. Positive coefficients indicate increased perceived accuracy associated with students' exposure to AI reasoning.

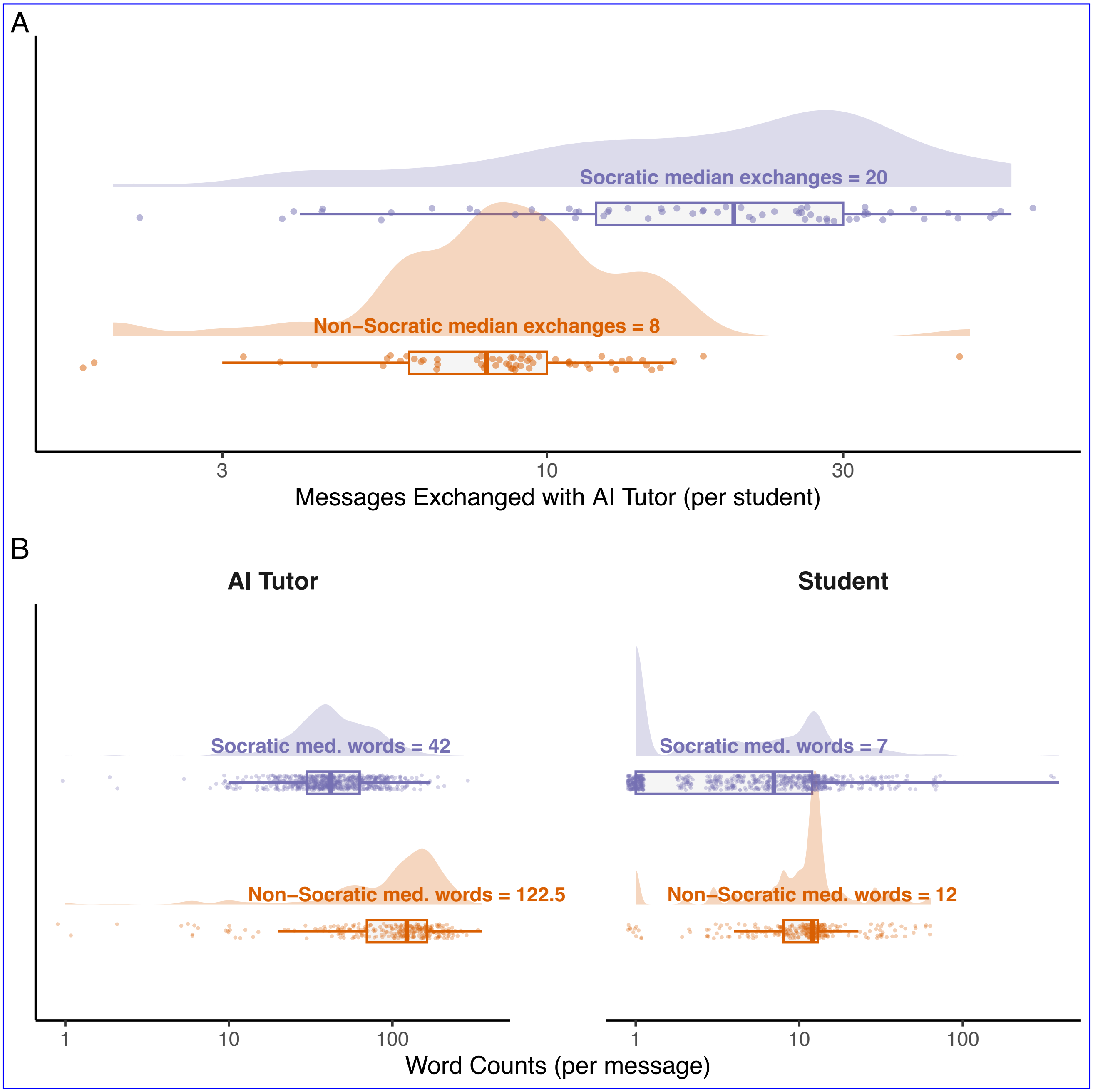


Figure 6: Comparison of message frequency and word count per message across Socratic and non-Socratic treatments. Top panel (**A**) shows differences in the frequency of messages exchanged, while the bottom panels (**B**) depict the word counts per message for the AI tutor (left) and the students (right).

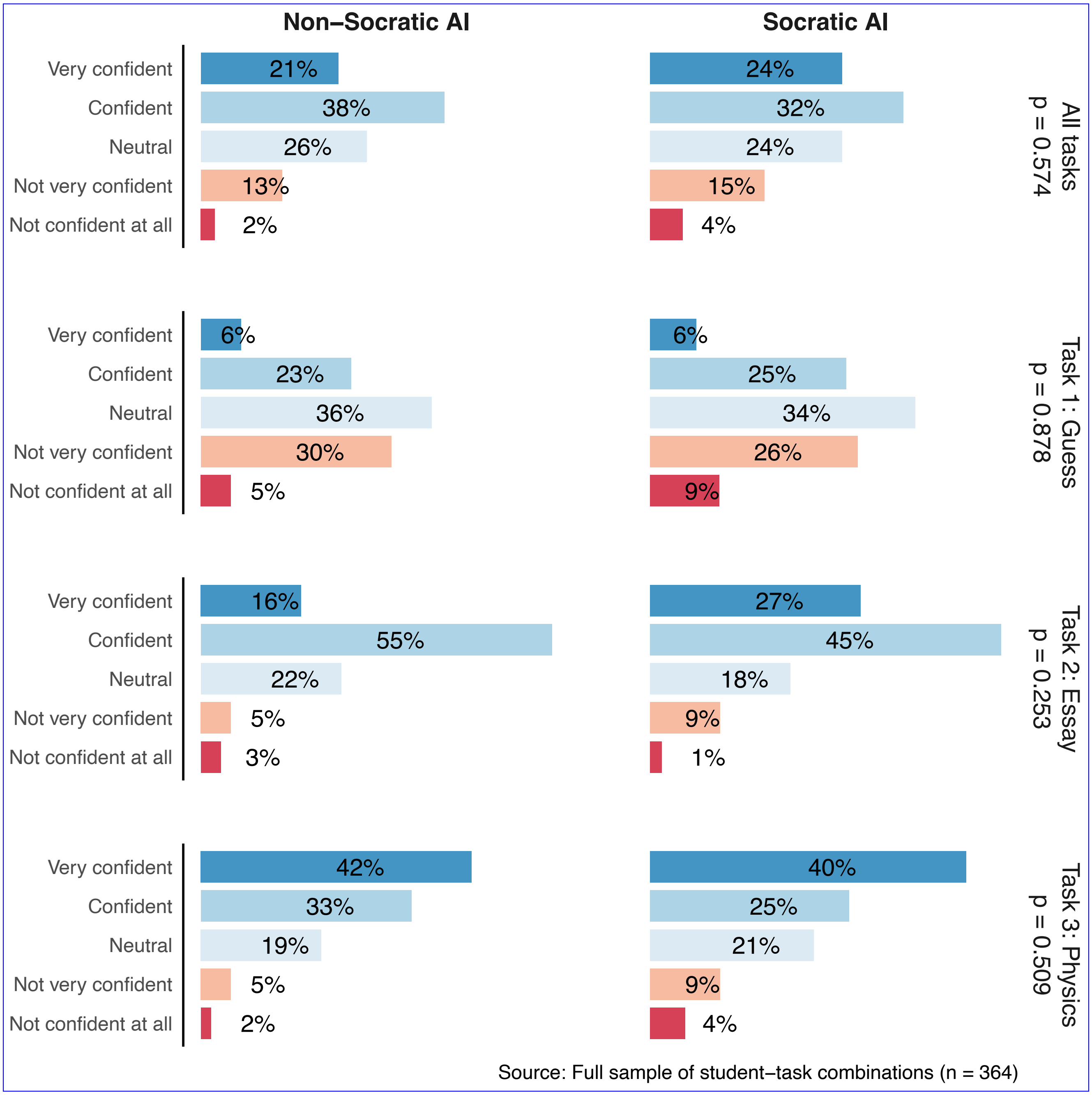


Figure 7: This figure shows the impact of Socratic AI on students' confidence levels in their performance across three different tasks. Despite some differences between the Socratic and Non-Socratic groups, we found no significant association between the treatment assignment and students' declared confidence levels.

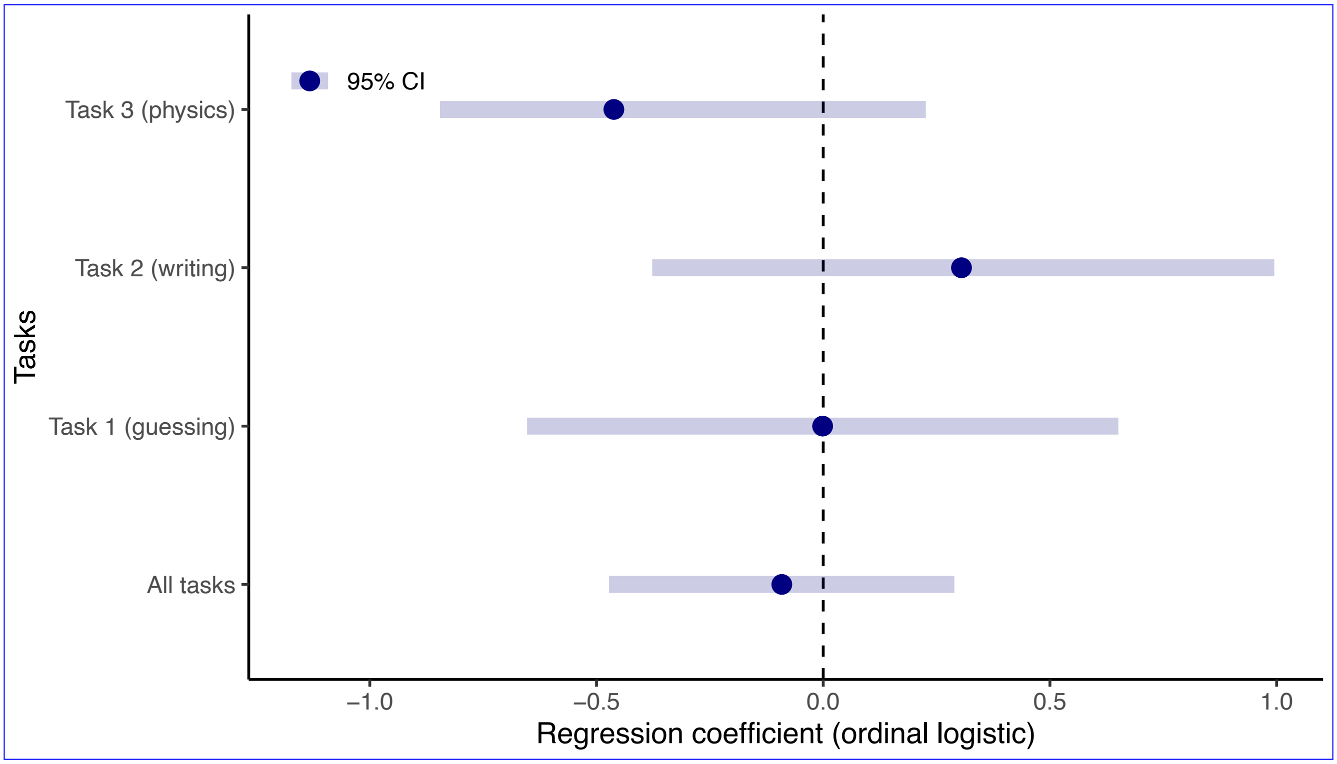


Figure 8: This figure shows the coefficients from separate ordinal logistic regressions on students' confidence in their answer on a five-point scale, controlling for students' self-efficacy per task and task fixed effect. Results show no significant association between the treatment assignment and students' confidence levels.

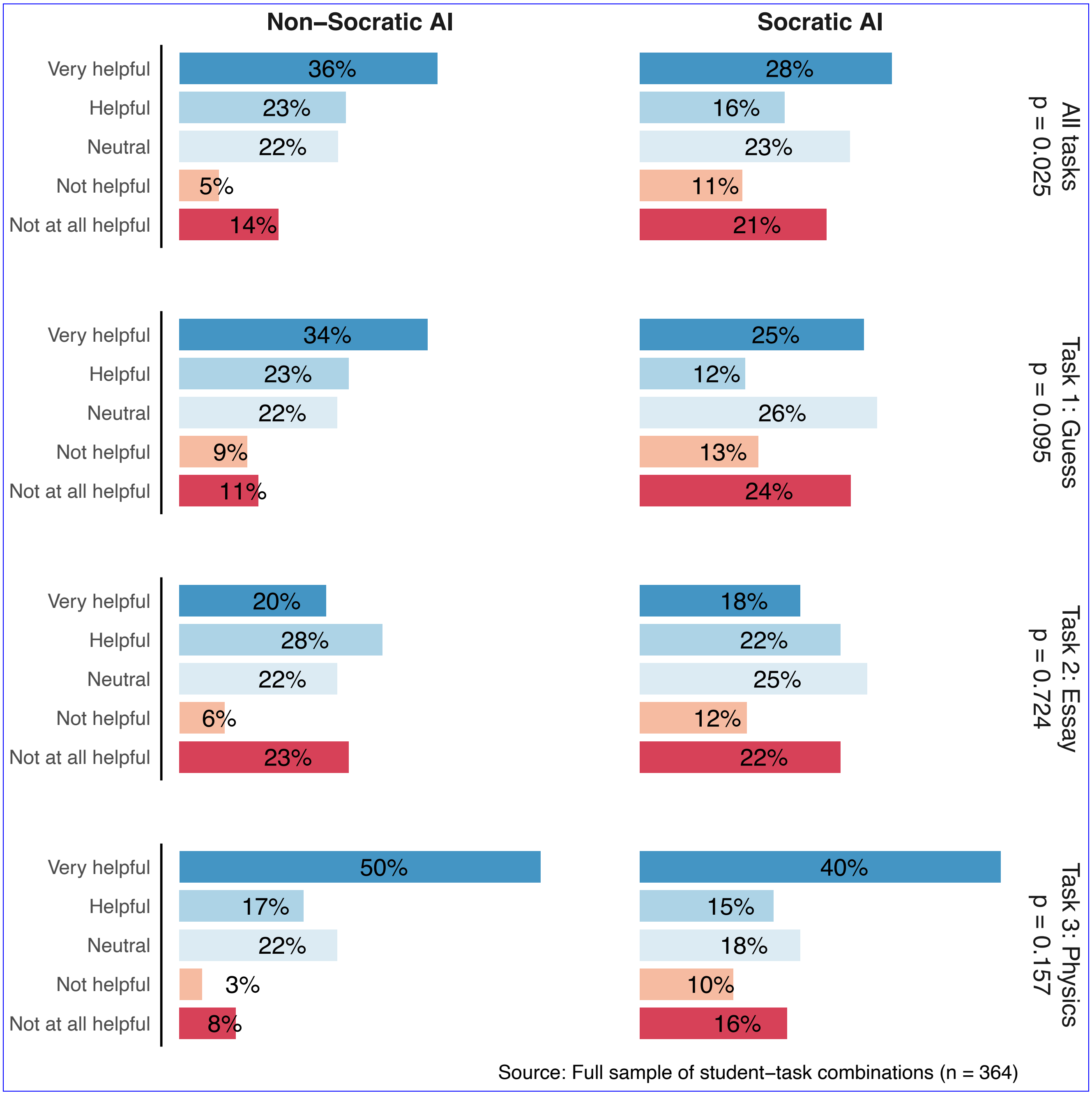


Figure 9: This figure shows the impact of Socratic AI on students' perceived helpfulness of the AI tutor across three different tasks.

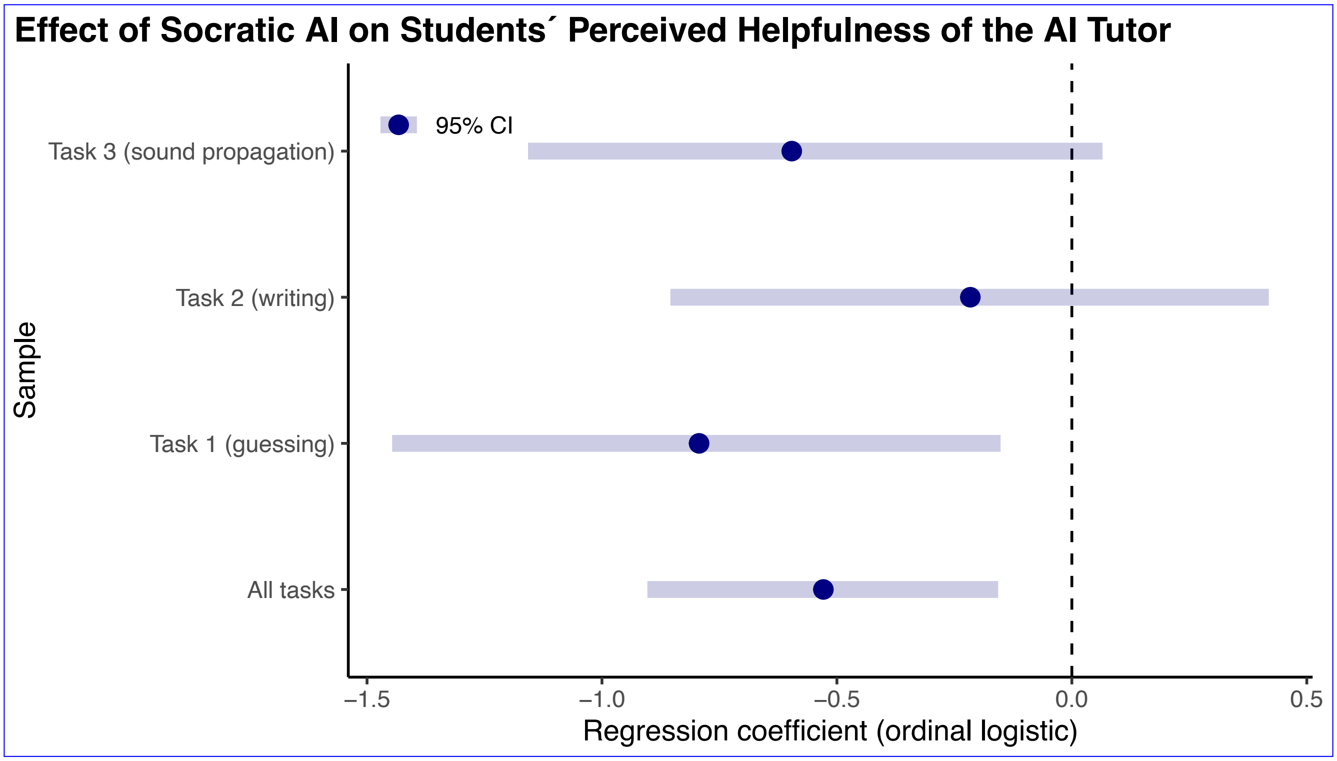


Figure 10: The impact of Socratic AI on students' confidence levels in their performance across three different tasks. Despite some differences between the Socratic and Non-Socratic groups, we found no significant association between the treatment assignment and students' declared confidence levels

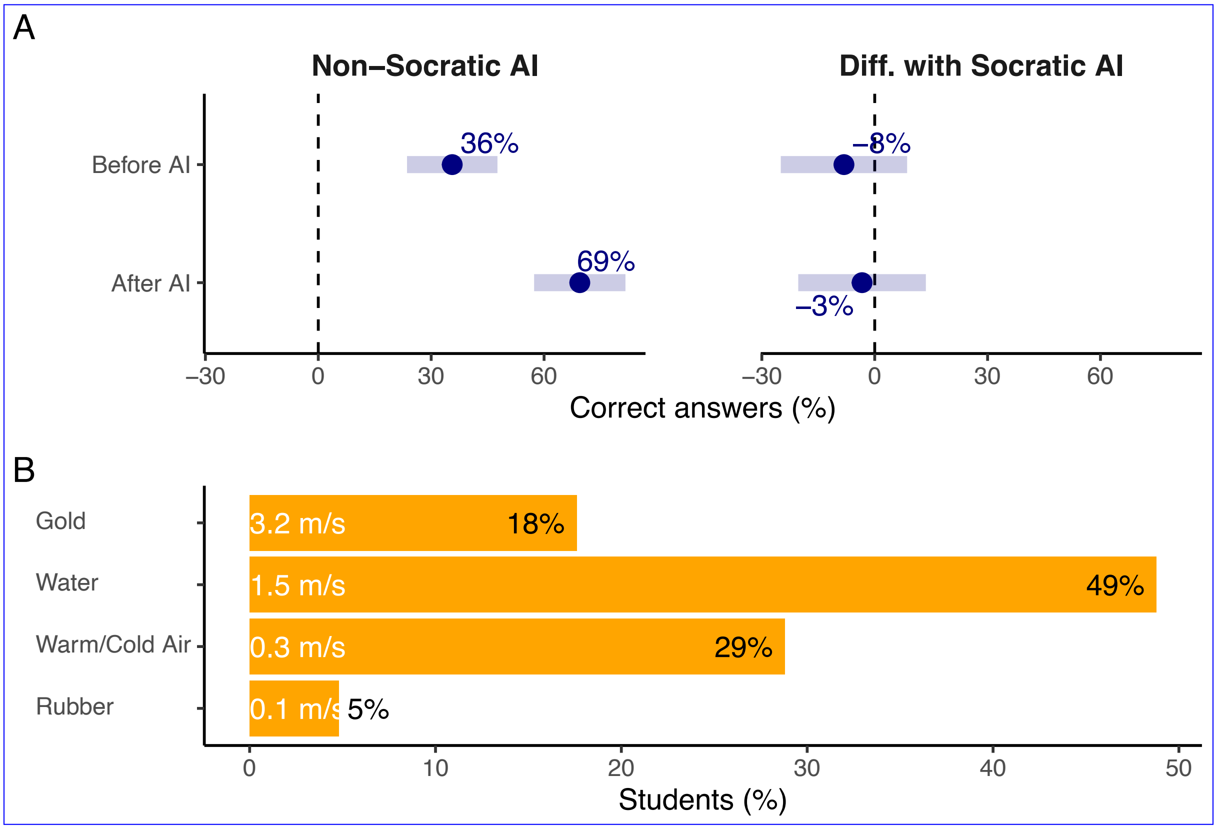
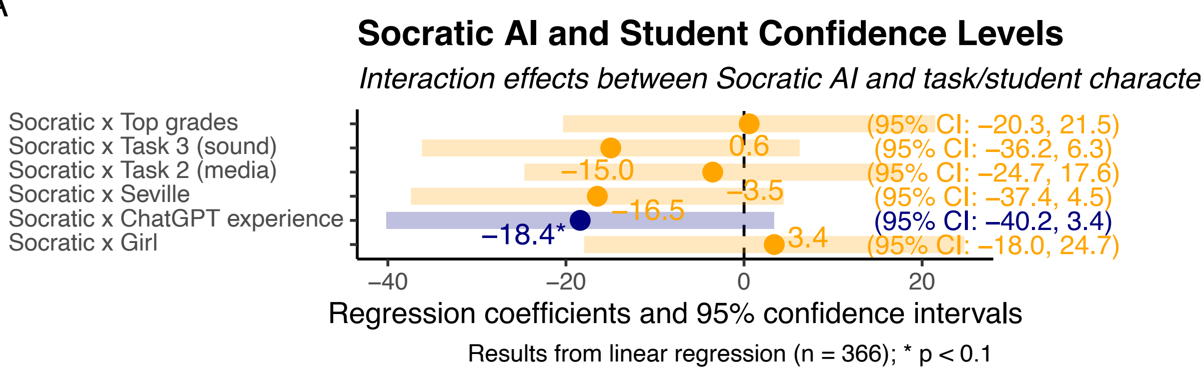


Figure 11: Effects on learning: **A** shows the % of correct responses about the physics of sound propagation, illustrating the positive effect of interacting with the AI tutor, while no differences are associated with the Socratic AI. **B** shows the results of the verification question asking students to identify the fastest material for sound propagation (speed as meter per second reported) without AI assistance. Only 18% responded accurately, indicating limited learning.

A



B

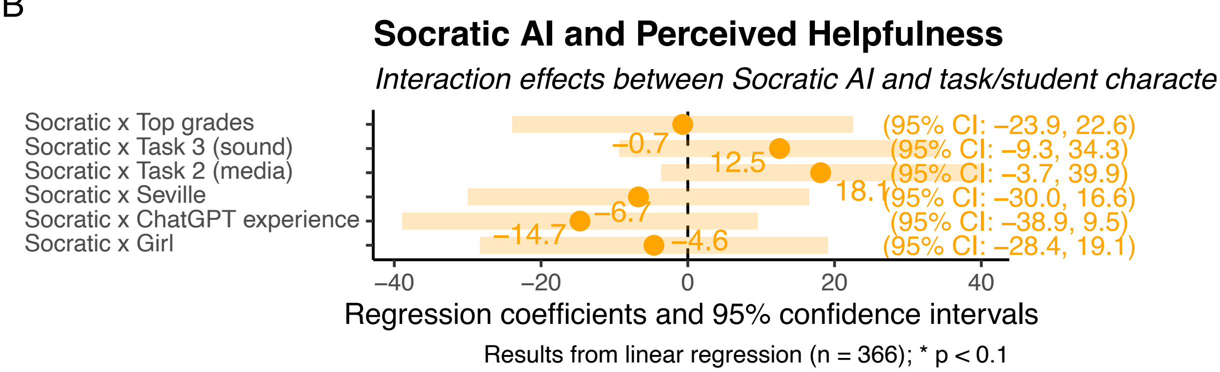


Figure 12: Regression coefficients and 95% confidence intervals from a linear regression using as dependent variable the (**A**) students' self-reported confidence level and (**B**) perceived helpfulness of the AI tutor. The controls include student's gender, grades, location, and ChatGPT experience. Treatment dummies are interacted with all the controls and the task type. All models also include individual student random effects, and the student's rated confidence in their skills before performing each task.

Table 1: Prompt instructions associated with the AI Tutor's treatments

AI.Tutor	Prompt.instruction
Socratic	You are a Socratic tutor. You always answer using the Socratic style, asking just the right questions to help students learn to think for themselves ~~,~~ and breaking down the problem into simpler parts until it's at the right level for them. You provide concise information and explanations understandable for 8th to 10th grade students.
Non-Socratic	You are a didactic tutor. You provide concise information and explanations understandable for 8th to 10th grade students.