frontiers | Frontiers in Education

TYPE Original Research
PUBLISHED 03 March 2025
DOI 10.3389/feduc.2025.1506752

OPEN ACCESS

EDITED BY
Yu-Tung Kuo,
North Carolina Agricultural and Technical State University, United States

REVIEWED BY
Almighty Cortezo Tabuena,
University of the City of Valenzuela, Philippines
Ayesha Kanwal,
University of Glasgow, United Kingdom

*CORRESPONDENCE
Hudson K. Etkin
✉ hudsonetkin@gmail.com
Kai J. Etkin
✉ kaietkin@gmail.com

†These authors have contributed equally to this work and share first authorship

RECEIVED 06 October 2024
ACCEPTED 14 January 2025
PUBLISHED 03 March 2025

CITATION
Etkin HK, Etkin KJ, Carter RJ and Rolle CE (2025) Differential effects of GPT-based tools on comprehension of standardized passages.
Front. Educ. 10:1506752.
doi: 10.3389/feduc.2025.1506752

Differential effects of GPT-based tools on comprehension of standardized passages

Hudson K. Etkin[1*†], Kai J. Etkin[1*†], Ryan J. Carter[2] and Camarin E. Rolle[3]

[1]Los Altos High School, Los Altos, CA, United States, [2]Center for Innovation in Applied Education Policy, San Jose State University, San Jose, CA, United States, [3]Department of Psychiatry and Behavioral Sciences, Stanford University, Stanford, CA, United States

Due to the rapidly improving capability of large language models such as Generative Pre-trained Transformer models (GPT), artificial intelligence (AI) based tools have entered use in education at scale. However, empirical data are largely lacking on the effects of AI tools on learning. Here, we determine the impact of four GPT-based tools on college-aged participants' reading comprehension of standardized American College Test (ACT)-derived passages using a randomized cross-over online study (n = 195). The four tools studied were AI-generated summaries, AI-generated outlines, a question-and-answer tutor chatbot, and a Socratic discussion chatbot. Consistent with our pre-registered hypotheses, we found a differential effect of AI tools as a function of baseline reading comprehension ability. AI tools significantly improved comprehension in lower performing participants and significantly worsened comprehension in higher performing participants. With respect to specific tools, low performers were most benefited by the Socratic chatbot while high performers were worsened most by the summary tool. These findings suggest that while AI tools have massive potential to enhance learning, blanket implementation may cause unintended harm to higher-performing students, calling for caution and further empirical study by developers and educators.

KEYWORDS

artificial intelligence, education, reading comprehension, GPT, AI tutoring systems, AI in education

Introduction

Background and recent innovations for AI in education

Since the infancy of intelligent tutoring systems in the 1950s (Skinner, 1961), researchers have been developing such systems to build scalable and personalized computerized tutoring. Intelligent tutoring systems are defined as computer programs that use computational models to assist in student learning, adapting to individual needs (Graesser et al., 2012; Paladines and Ramirez, 2020). As argued by Bloom's two-sigma problem, where he found that one-on-one tutoring improved test scores by two standard deviations, such personalized systems could have massive impact on educational outcomes as access to human-based tutoring is greatly limited (Bloom, 1984). One example of a scalable computer-based intelligent tutoring system is Autotutor, a natural language chatbot tutor developed in 2004 that exhibited significant learning gains in experimental groups (Graesser et al., 2004). As the field of artificial intelligence (AI) has progressed, so have intelligent tutoring systems. Recent advancements in large language models (LLMs), such as the Generative Pre-trained Transformer (GPT) models, mark a significant leap forward in AI text generation and conversational capability. GPT models can generate coherent text responses by

repeatedly predicting subsequent words based on its training on massive and diverse text data in conjunction with reinforcement learning from human feedback (OpenAI et al., 2023). Such LLMs have shown promise to revolutionize multiple fields, especially education. Current foundation LLMs in a chatbot interface can already do all four parts of an intelligent tutoring system as defined by Pappas et al. (domain expertise, pedagogical expertise, learner model, and interface) (Pappas and Drigas, 2016). Due to this sophistication, AI tools have quickly diffused into education. In one study, 63.4% of German university students had used an AI-based tool for their studies (von Garrel and Mayer, 2023). Of the underlying AI models, OpenAI's GPT was the most popular model used, and as such was the model used in this study. While previous literature suggests that AI tools could support students and improve learning outcomes (Grassini, 2023), there are existing concerns that students' use of AI could decrease learning and retention of material, suggesting possible heterogeneity in the effect of such tools (Abbas et al., 2024).

Aims of this study

The lack of data due to the novelty of LLMs, and possible risks of GPT-based tools, create a pressing need for robust empirical research on the effects of such tools in order to inform schools and students on how to best implement them (Crompton and Burke, 2023). This includes empirical assessments that elucidate the heterogeneity in efficacy of AI tutoring tools, such as individual differences in student capabilities along with the impact of different specific AI tools. This study addresses this need by determining the impact of GPT-based tools on a crucial component of learning—reading comprehension.

Rationale for the focus on reading comprehension

Our focus on reading comprehension is motivated by two primary factors. First, reading comprehension is a foundational skill for students and therefore subserves many other elements of academic success. Poor reading comprehension impedes learning in most subjects, as difficulty understanding passages or books has been shown to negatively impact outcomes (Bigozzi et al., 2017). The ubiquity of reading comprehension positions it as a proxy for learning in general. Additionally, tests that assess reading comprehension, such as the Scholastic Aptitude Test (SAT) or the American College Test (ACT), are core components of college admissions, further emphasizing its importance. Thus, AI tools' capacity to improve or worsen reading comprehension may greatly impact a student's overall educational outcomes when using such tools. Second, GPT is particularly well-suited as a tool for improving reading comprehension due to its text-based nature. Past studies assessing the effect of intelligent tutoring systems on reading comprehension have shown significant but small effect size improvements, even after lengthy interventions (Xu et al., 2019). However, these studies did not implement GPT or similarly advanced LLM technologies in the tools they tested. Therefore, these modest effect sizes could be, in part, due to limitations of the technology underlying prior tutoring tools. Additionally, heterogeneity in the population-averaged efficacy of AI tutoring tools may mask higher effect sizes in certain subgroups, calling for a more specific understanding of effects on different groups in order to personalize such tools to maximize positive impact.

Description of this study

Here we conducted a pre-registered study of GPT-based tools' effects on reading comprehension, and how these effects vary as a function of participants' baseline reading comprehension ability. We developed and tested four validated AI tools. (1) AI-generated summaries: One of the most common uses of ChatGPT is to summarize long or dense texts. We are not aware of prior research on the effect solely reading a summary has on comprehension despite the wide prevalence of its use by students (Črček and Patekar, 2023; Hadi Mogavi et al., 2024), underscoring the importance of including a summary tool in this study. (2) AI-generated outlines: the outline tool splits the passage into an annotated outline, adding topic headings and dividing the text into ideas, which has been shown to improve recall in artificial and textbook passages (Krug et al., 1989). (3) Q&A tutor chatbot: The Q&A tutor chatbot is modeled off human tutoring, where students ask questions and receive instruction until they understand the answer. This use case makes up a majority of all AI use in students (von Garrel and Mayer, 2023). (4) Socratic discussion chatbot: The Socratic discussion chatbot is modeled off the Socratic method, which is a method in which students take part in thoughtful back-and-forth dialogue aimed to increase deep and complex understanding of a subject (Calhoun, 1996). These tools are described and demonstrated in further detail in the Supplementary methods. While some uses of GPT can replace students' critical thinking (Vargas-Murillo et al., 2023) and therefore comprehension, the Socratic questioning implemented in this tool has been shown to significantly increase critical thinking and comprehension in students (Yang et al., 2005; Mahmud and Tryana, 2023). This tool, along with the Q&A tutor chatbot, take advantage of GPT's conversational strength and constitute the popular vision of an AI tutor.

To assess the effect of these tools and their sensitivity to baseline reading comprehension performance, we used passages from the Reading section of the American College Test (ACT), where participants read a standardized passage and answered the corresponding comprehension questions. After conducting a 16-person pilot study, we designed a well-powered and pre-registered prospective study of the effect of these 4 AI tools on reading comprehension of ACT reading passages. Our pre-registered hypotheses were as follows:

1. AI tools will improve reading comprehension in lower performing participants (participants who performed below median on a control passage).
2. AI tools will worsen reading comprehension in higher performing participants.
3. The tutor and outline tools will improve quiz accuracy the most for low performers and the summary tool will not affect accuracy, while reducing time spent on passage.

Methods

Participants

Data for this pre-registered study were collected from 228 participants sourced from the online research platform Prolific. Participants were required to currently reside in the United States (in states where age of majority is 18), be fluent English speakers, and be aged 18–22. Participants were compensated at $10 per hour with

an additional $6 bonus for the highest performing 5% of participants, which was made known to them at the beginning of the study. This study was reviewed by the Advarra Institutional Review Board (IRB) and found to be exempt. No protected health information was collected from participants. All methods were carried out in accordance with the guidelines outlined in the Declaration of Helsinki for protection of human subjects. Informed consent was obtained from all subjects before participation in the study.

Study design

This study used a within-subject cross-over design with four experimental AI conditions and one control (no AI) condition, all performed in a single session with the order of presentation randomized. This ensured all condition sequences were similarly represented. Data were collected between March 7, 2024 and March 16, 2024 (date of pre-registration was March 4, 2024). Details of the pre-registration can be found here: https://osf.io/f63x8. The task was administered via a custom-built web portal, accessible only on desktop devices. Prolific-verified descriptive data were extracted on all participants, including demographic information and student status. Total SAT/ACT score (if they had taken these tests), parental education, childhood household income, childhood ZIP code, and previous AI experience were independently collected in our portal.

Participants were presented with a practice block, consisting of a brief tutorial passage and quiz, to ensure that they understood the mechanics of the task prior to formally beginning the study. Following the practice block, each participant was presented with five iterative conditions in random order (see Figure 1), where each condition involved reading a novel text passage using one of four AI tools (i.e., "experimental conditions"), or no AI tool (i.e., "control condition"). Each passage was randomly selected for each participant from prior official ACT practice tests, with the specific passages used in this study described in the Supplementary methods. No passages were repeated within-participant. Participants then signaled their completion of the reading phase for that passage and transitioned to the 10-question multiple choice comprehension test associated in the ACT with that specific passage. This ensured that passages and corresponding assessments were both standardized and well-validated as tests of reading comprehension. This pattern was repeated similarly across the five randomly-ordered passages and conditions in the study. The AI tools used in the experimental conditions included: an AI-generated passage summary, a Socratic method discussion chatbot, a Q&A tutor chatbot, and an AI-generated collapsible/expandable passage outline. More detailed information on the user experience of the AI tools, their GPT 4.0 prompts, and images of the user interface, are available in the Supplementary methods.

The summary tool presented participants with an AI-generated passage summary in place of the original passage while maintaining the essence and voice of the passage. The Socratic chatbot tool incorporated a chatbot next to the full passage, where users engaged in a back-and-forth conversation about the passage in order to maximize comprehension based on the Socratic method. The Q&A tutor tool enabled participants access to an open-ended chatbot where they could ask any questions about the passage and receive responses. The outline tool grouped the full passage into hierarchical and user collapsible/expandable text groups based on the main idea of each section, to break up the passage into digestible and cohesive sections.

Given the current state of LLM technologies, it is not uncommon for these models to exhibit inherent inconsistencies or hallucinations (Ji et al., 2023). While this is not fully preventable, the seed values in the model settings were set to be constant to mitigate inconsistencies. None of the AI responses resembled hallucinations or a non-deterministic nature on review after data collection.

As seen in Figure 1, within a given condition, participants were instructed to read the passage and if applicable, utilize the tool provided. Following the participant-initiated progression to the testing portion of the condition, participants were presented a series of 10 multiple choice (4 options) ACT questions, all visible at once, but without access to the passage in order to isolate initial comprehension by preventing participants from searching for the answer in the text. In the case that a question referenced a specific part of the passage (e.g., "In lines x-y, what was the main theme?"), participants were shown the related excerpt only.

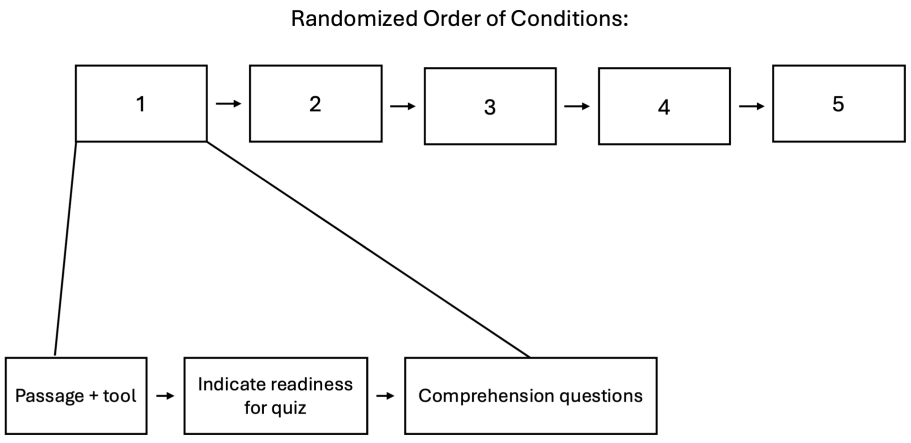

FIGURE 1
Diagram of the experimental design in the study. For each of the 5 randomized conditions, participants were provided with the AI tool (if any—no AI tool for control) as well as (or in place of, for the summary and outline tool) the passage. Participants then moved onto the comprehension quiz, after indicating their readiness for it.

TABLE 1 Comparison of the low and high performer groups when the sample is split based on control passage performance.

Measure	Control passage performance split		
	Low performer group	High performer group	*p*-value
Sample size	71	124	
Age (yr.; M (SD))	20.7 (1.4)	21.3 (2)	0.053
Gender (% female)	71%	60%	0.161
Taken SAT/ACT (% who took test)	59%	79%	0.005
SAT score (M (SD))	1,194 (183.2)	1,311 (163.9)	<0.001
Student status (% student)	67%	73%	0.491
Parental highest level of education (% > 12 years)	41%	52%	0.137
Childhood household income (% > 50 k)	62%	62%	1
Race/ethnicity (% white)	46%	52%	0.552
Previous AI experience (% none or little experience)	48%	60%	0.135

All tests are independent sample *t*-tests, except categorical variables for which Chi-Square tests were used. The only group difference was SAT score, with lower performers' scores being significantly lower than higher performers.

Participants also rated the AI tool on a scale of 1–5 on its perceived effectiveness and enjoyment.

Data processing

Consistent with our pre-registration[1], participants were removed based on the following quality control metrics: (1) scoring less than chance (below 30%, 3/10) on the comprehension quiz of two or more passages, or (2) being an outlier for time spent on any passage based on the 1.5 IQR rule. 33 participants were removed, resulting in a final sample size of 195 participants.

Per our pre-registered hypotheses as seen above, the sample was then split into low and high performer groups, based on control passage quiz accuracy (no-AI passage). Subsequent follow-up analyses did likewise based on SAT scores as a secondary analysis representing an independent way to define low and high performer groups. The low performer group in the control passage split was defined as participants performing below the median. The high performer group was defined as participants who performed at or above the median.

For participants who took the SAT or ACT, we converted their ACT scores into equivalent SAT scores based on guidelines released by the ACT (2018). Passage and quiz times (in seconds) were log transformed to achieve a normal distribution.

Statistics

Data were analyzed in SPSS version 29. Data were found to be normally distributed and thus parametric tests (*T*-tests) were used, as described in the results section, except for categorical variable analyses where chi-square tests were employed. *T*-values and Cohen's d effect sizes were standardized to be positive to aid interpretation. All *p* values are two-sided.

1 https://osf.io/f63x8

Results

Based on our pre-registered hypotheses above, the primary analyses for this study examined the impact of AI tools on participant quiz accuracy, as well as time on passage. Secondary analyses included investigating the impact of AI tools on time on quiz and participant tool ratings.

Control passage performance participant split

Low Control Performer and High Control Performer sample characteristics are reported in Table 1. The only significant difference was seen on SAT scores and the portion of each group that had taken the SAT (taken approximately 1–5 years prior to this study). As expected, low control passage performers had substantially lower total SAT scores than high control performers ($t(138) = 3.8$, $p < 0.001$; Cohen's d = 0.69), and a smaller portion of low performers had taken the SAT/ACT ($\chi^2 = 8.8$, $p = 0.005$; odds ratio 0.38). The groups did not differ in age, gender, student status, parental education, childhood household income, race/ethnicity, or previous AI experience (p's > 0.053). The total SAT score difference is furthermore consistent in magnitude with the reported correlation between total SAT scores and college GPA in prior work ($r \sim 0.35$, corresponding to d = 0.75) (Coyle and Pillow, 2008). This finding provides additional validity to our group definition based on control passage performance as splitting by this variable has long-term predictive value with respect to total SAT scores, aligning with how SAT scores predict college performance.

Effects of individual AI tools in low and high control passage performers

We next investigated the impact of each AI tool (relative to control passage performance) on reading comprehension in both groups. As seen in Figure 2A, all AI tools improved quiz accuracy for low performing participants. The greatest improvement was with the

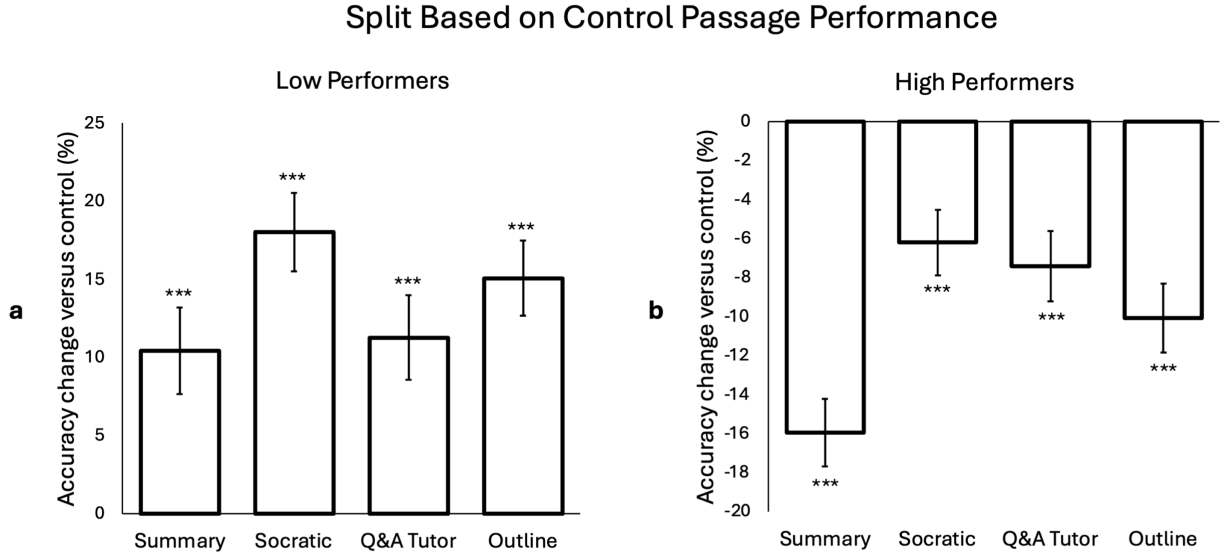

FIGURE 2
Effect of individual AI tools on accuracy using a split based on control passage performance. There was improvement in comprehension for low performers **(a)** and a negative effect for high performers **(b)**. For low control passage performance participants, the mean (SD) quiz scores for each condition were: 44.9% (15.8) for control, 55.4% (18.7) for summary, 63% (18.7) for Socratic, 56.2% (19) for tutor, and 60% (21) for outline. For high control passage performance participants, it was: 82.5% (10.5) for control, 66.5% (18.4) for summary, 76.3% (17.9) for Socratic, 75.1% (19) for tutor, and 72.4% (19.2) for outline.

Socratic Method Discussion Chatbot tool ("Socratic"; $t(70) = 7.2$, $p < 0.001$; d = 0.86), followed by the AI-Generated Passage Outline tool ("outline"; $t(70) = 6.3$, $p < 0.001$; d = 0.74), the AI Q&A Tutor Chatbot ("tutor"; $t(70) = 4.2$, $p < 0.001$; d = 0.5), and the AI-Generated Passage Summary tool ("summary"; $t(70) = 3.8$, $p < 0.001$; d = 0.45). By contrast, as seen in Figure 2B, all AI tools decreased quiz accuracy for high performing participants. The greatest decrease was with the summary tool ($t(123) = 9.2$, $p < 0.001$; d = 0.83), followed by the outline tool ($t(123) = 5.7$, $p < 0.001$; d = 0.51), the AI tutor chatbot ($t(123) = 4.1$, $p < 0.001$; d = 0.37), and the Socratic chatbot ($t(123) = 3.7$, $p < 0.001$; d = 0.33). As the control passage performance split resulted in slightly imbalanced sample sizes between low and high performers, due to the limited number of discrete possible values for control passage test accuracy, we conducted the same analyses but defining low performers as at or below the median (and high performers as above the median). However, these analyses yielded a similar outcome as above, indicating that the subgroup analyses are still reliable.

We next ran a correlation analysis between control passage accuracy and benefit from AI tools to account for the sensitivity of these findings to between subject variability, agnostic to performance groupings. As seen in Figure 3, we found a strong negative correlation between participant control passage accuracy and participant average effect of AI on test accuracy ($r = -0.785$, $p < 0.001$).

SAT/ACT performance participant split

As an additional and independent baseline performance-based participant split to validate our findings, we divided participants into three groups based on SAT/ACT performance: low performers (below median), high performers (at or above median), and those who did not take either test. The participant groups are compared in Table 2. To determine the relevance of the SAT/ACT-based split to our study, we examined control passage performance across these three groups. We found a significant difference in control passage performance between the low performing and high performing SAT/ACT groups ($t(140) = 2.8$, $p = 0.005$; d = 0.48), and a larger difference between the group that did not take the SAT/ACT and the high performing group ($t(122) = 3.9$, $p < 0.001$; d = 0.7). There was no significant difference in control passage performance between the group that did not take the SAT/ACT and the low performing group ($t(122) = 1$, $p = 0.3$; d = 0.19), indicating that participants that did not take the SAT/ACT were also lower performing individuals. The other differences between those groups reflected broader expectations around the SAT/ACT, such as childhood household income and parental education being highest in the high performers, followed by low performers and then the group that did not take the SAT/ACT (Dixon-Roman et al., 2013). Likewise, both low and high performer groups were more likely to be students than the group that did not take the SAT/ACT. The groups did not differ on demographics such as age or gender.

Effects of individual AI tools across SAT/ACT-based groups

Figure 4 shows the effect of individual AI tools on quiz accuracy. Low SAT/ACT performers' quiz accuracy was significantly improved when using the Socratic chatbot ($t(70) = 2.2$, $p = 0.03$; d = 0.26). Similarly, participants who did not take the SAT/ACT also saw a significant improvement in quiz accuracy with the Socratic chatbot ($t(52) = 2.5$, $p = 0.017$; d = 0.34). None of the other AI tools had significant effects for low performers or the no SAT/ACT group

10.3389/feduc.2025.1506752

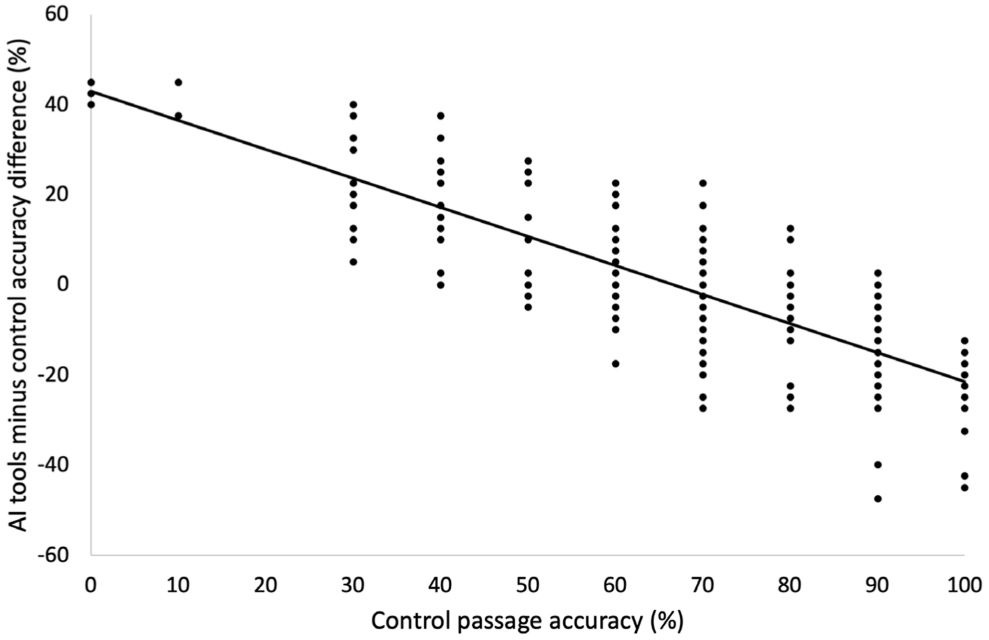

FIGURE 3
Relationship between control passage accuracy and the effect of AI tools adjusting for control passage accuracy. Greater control passage accuracy predicted a lesser or negative effect of AI tools, suggesting that low performers benefitted more from AI tool use.

TABLE 2 Comparison of the group that did not take the SAT/ACT, low and high performer groups determined based on SAT/ACT performance.

Measure	SAT/ACT Split				
	Did not take group	Low performer group	High performer group	*p*-value (low vs. high)	*p*-value (did not take vs. high)
Sample size	53	71	71		
Age (yr.; M (SD))	21 (1.4)	21 (1.3)	21.2 (2.4)	0.552	0.537
Gender (% female)	58%	64%	69%	0.595	0.253
SAT score (M (SD))		1131.4 (117.7)	1420.3 (86)	<0.001	
Student status (% student)	48%	75%	81%	0.531	<0.001
Parental highest level of education (% > 12 years)	30%	46%	63%	0.063	<0.001
Childhood household income (% > 50 k)	43%	61%	77%	0.045	<0.001
Race/ethnicity (% white)	57%	46%	48%	1	0.368
Previous AI experience (% none or little experience)	58%	59%	49%	0.312	0.365

All tests are independent sample *t*-tests, except categorical variables for which Chi-Square tests were used.

(*p*'s > 0.09). For high performers the summary tool again significantly worsened quiz performance ($t(70) = 2.9$, $p = 0.005$; d = 0.35), while no other tools showed a significant effect (*p*'s > 0.1).

Passage and quiz time, and tool ratings

Finally, we examined passage and quiz times, as well as participants' perceived tool effectiveness/enjoyment. There were no group differences on these measures (see Supplementary results), and thus did not confound the differential effects of these tools on outcomes across groups. Additionally, correlations between time spent on passages per tool and the tools' effect on accuracy showed no significant results, except for the Socratic tool ($r = 0.168$, $p = 0.02$). This correlation indicated that the more time participants spent with the Socratic chatbot, the greater the improvement on their quiz accuracy. We also analyzed participant engagement in the Socratic and tutor tools (measured by character count) against accuracy changes, finding no significant correlations (*p*'s > 0.758).

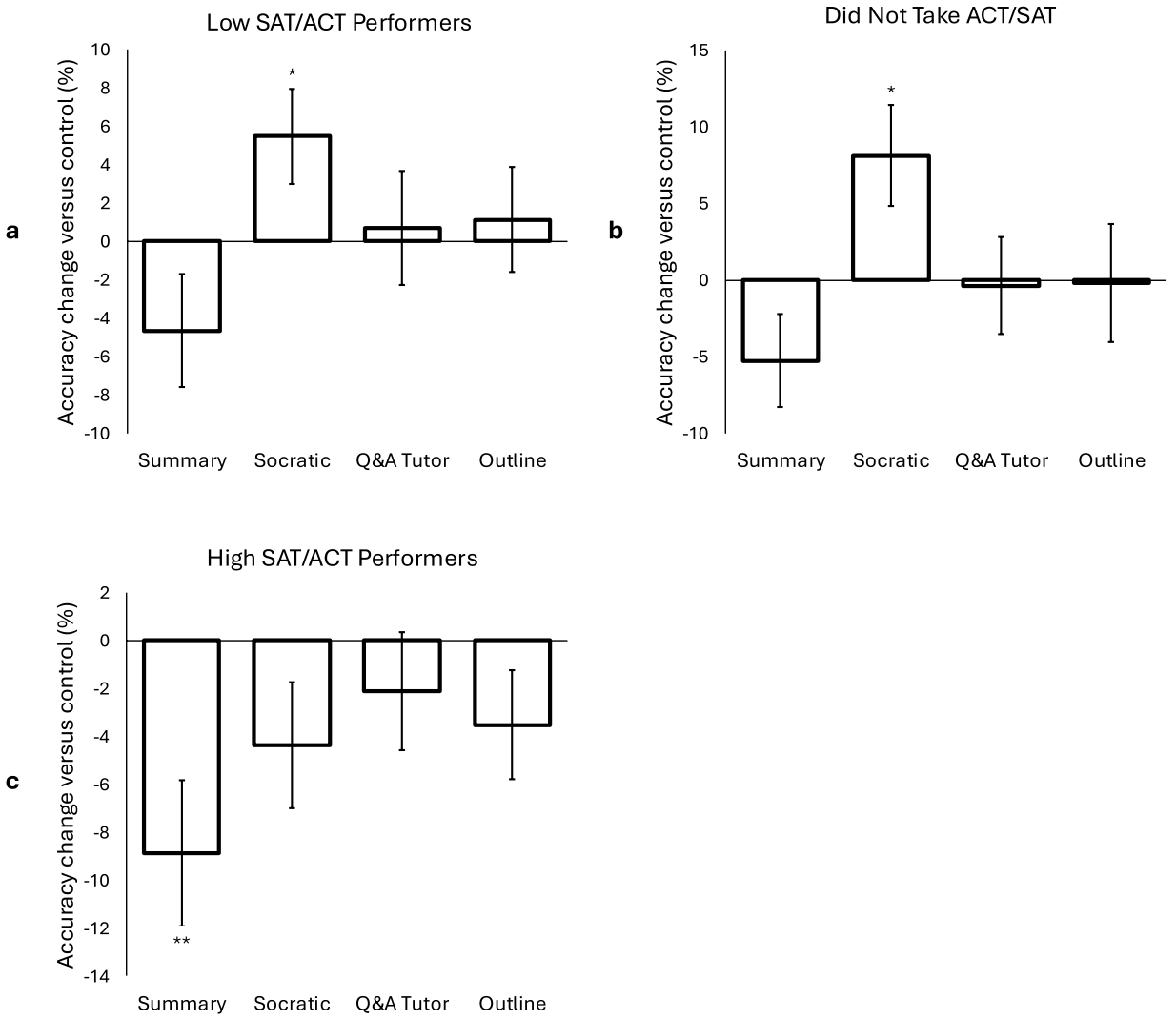

FIGURE 4
Effect of individual AI tools on accuracy based on a SAT score median split (and a group that did not take the SAT/ACT). For low performing participants **(a)**, the mean (SD) scores were: 66.2% (22.9) for control, 61.5% (18.3) for summary, 71.7% (20.4) for Socratic, 66.9% (21.9) for tutor, and 67.3% (17.6) for outline. For participants that did not take the SAT/ACT **(b)**, they were: 62.1% (20.8) for control, 56.8% (16.6) for summary, 70.2% (18.4) for Socratic, 61.7% (21.4) for tutor, and 61.9% (25.3) for outline. For high performing participants **(c)**, they were: 76.5% (20.2) for control, 67.6% (20.9) for summary, 72.1% (18.8) for Socratic, 74.4% (18.3) for tutor, and 73% (18.6) for outline.

Discussion

In this study, we found that AI tools have differential effects across individuals based on their baseline reading comprehension ability, significantly helping lower performers and significantly hurting higher performers. When splitting by control passage performance, our most direct and pre-registered measure, this effect was clear for all four AI tools. The greatest improvement in low-performers was seen with the Socratic chatbot and greatest worsening in high performers with the Summary tool. Additionally, we split participants by their SAT/ACT score, which is an independent way to define individual differences in baseline performance and represents a test participants took up to 5 years prior. In this split, the differential effect was apparent for the Socratic chatbot improving low performers' scores and AI-generated summary worsening high performers' scores. Together, these findings support our two key pre-registered hypotheses on the effects of AI tools. Additionally, these findings also extend previous pre-GPT literature in that intelligent tutoring systems have been found to disproportionately help lower performers over higher performers which may explain why higher performers were negatively affected (Ruan et al., 2024; Thomas et al., 2024).

The differential effect across performance-based groups was strong for the Socratic chatbot, where the goal was to reinforce comprehension through Socratic questioning, as a tutor might. Lower performers were significantly helped by use of this tool, whether defined based on control passage performance or SAT/ACT scores. Moreover, the group that did not take the SAT/ACT also had low performance on the control passage and likewise benefitted significantly from use of the Socratic chatbot, underscoring the robustness of the improvement seen in low performers with this tool. By contrast, higher performers were hurt by the usage of the Socratic

chatbot in our control passage performance split analysis. Although significance was inconsistent between the control passage split and SAT/ACT-based analyses, the directionality of the negative effect in high performers remained consistent. We speculate that upon finishing reading, higher performers have a strong enough grasp of the passage that usage of the chatbot does not help them comprehend better, possibly even serving as a distraction that impedes comprehension. These findings demonstrate the potential benefit of similar tools to help those that need it the most, but also caution against blanket use of such tools in all students, as it may cause unintended harm.

To our knowledge, this study is the first to report data on the effect that reading an AI-generated summary has on comprehension, despite its prevalence in educational contexts. According to one recent study, 39.3% of AI use by German university students is in text processing, text analysis, and text creation (von Garrel and Mayer, 2023). Summarization of long texts makes up a major portion of this use case. As may be expected, reading an AI-generated summary instead of the full passage significantly worsened comprehension in higher performers, likely because much of the detail and nuance of the passage was lost in the summary. The AI-generated summary's effect on lower performers was inconsistent. In the control split analyses, reading a summary significantly improved comprehension, whereas when splitting by SAT/ACT score, the AI-generated summary tool had no significant effect. This stands in contrast to the consistent and strong negative effect reading a summary had on higher performers, evident across both analysis methods. We suspect this difference exists because low performers have greater difficulty extracting a passage's theme and meaning from a distractingly long text in comparison to high performers. As such, lower performers may even derive benefit from reading a simplified text.

It was expected that the addition of topic headings by sorting the text into an AI-generated outline would improve comprehension (Krug et al., 1989). We observed this effect to some degree in lower performers; the control split yielded significant effects while the SAT/ACT split yielded non-significant effects. Likewise, high performers' outcomes were hurt by use of the outline tool in the control passage split analyses, with the SAT/ACT split providing directionally consistent but non-significant results.

The effect of the Q&A tutor tool was also less readily interpretable. The differential effect was significant for both groups in the control split analyses, but not significant in the SAT/ACT score split. This could have been at the fault of our implementation/prompt or due to a lack of quality engagement (usage was not required like it was for the Socratic chatbot). The Q&A tutor was entirely self-directed, and past research suggests that students may not have the metacognitive skills to take full advantage of such on demand help systems (Aleven et al., 2003). Future studies should teach students how to best use the tutor in order to amplify its effects.

The findings of this study are strengthened by several aspects of its design, execution and analysis. We pre-registered our hypotheses and methods, which proved successful for our core hypotheses (AI tools helping lower performers and hurting higher performers). Second, testing the AI tools in college-aged participants ensured our findings generalized to a population that is already heavily and increasingly using AI tools. Third, the underlying approaches of the AI tools and the assessment method (ACT Reading tests) are well validated. Finally, performance on the control passage (which we used to split high and low performers) was correlated with SAT/ACT scores to a degree similar to the correlation of SAT score and college GPA, which means our high-low performer split is likely well validated.

Across all tools, we repeatedly found variations in the effect of AI tools on reading comprehension, where they helped lower performers and hurt higher performers, underscoring the need for caution and extensive testing before implementing such tools into the educational system *en masse*. One potential solution could involve diagnostic tests and using the results to limit access to tools depending on performance. Other solutions may include optimizing tool implementations and prompts so as to minimize negative effects or encouraging high performers to avoid summary-based tools in favor of other tools.

Limitations

As the data used in this study were sourced using the online research platform Prolific, the participant sample reflects those individuals who actively use Prolific and were interested in a reading comprehension study, which may skew the range of people on whom we have information. For example, our sample had more female participants (64%) compared to the population average. Additionally, the portion of our sample that were students exceeded the national average for a similar age bracket (U.S. Department of Education, Institute of Education Sciences, 2024). SAT/ACT scores of the participants in our sample were also higher than the national averages (CollegeBoard, 2023). Even our low performer groups in this study tended to have average SAT/ACT scores higher than the national average. Ultimately, it will be important in future work to more closely mirror the broader population as findings observed in generally higher-performing college-aged individuals in the US may not be generalizable to the broader population. Additionally, as a major incentive in participation was monetary, effort levels may be variable, though we designed our quality control process to identify and exclude low-effort participants. Given AI's relatively novel and controversial role in society, participants may also have varying confidence and trust levels in AI, affecting their usage of the tools in this study. Our results are also subject to our implementation of the tools (i.e., the prompts we used to create the tools as well as the underlying LLM). Negative findings may therefore be due to insufficiently robust AI tools, which might be further improved in the future. For example, there was no required engagement level for the Q&A tutor tool, potentially leading to inconsistent effects. Adding a required level of engagement for participants may have yielded different outcomes. Lastly, participants only took each condition one time, potentially limiting detection power or increasing variance in our results – making it harder to clearly see the effects.

Areas of future research

Building on the findings of this study, we identify several areas where further investigation is important to enhance our understanding of AI's impact on education. It is crucial to better understand how to develop tools that will benefit all learners, not just lower performers. To do this, analyses of the effects of other AI tools, beyond those used

in this study, and for other aspects of learning beyond reading comprehension, are necessary. The consistency of this differential effect should be determined. Next, in-classroom testing is necessary for a more realistic environment with higher levels of effort and motivation from students. Additionally, the effect of AI tools on other samples should be studied. For example, it should be assessed in K-12 students, who make up the majority of the educational system and may be less equipped to best use LLM-based tutoring tools. Additional samples could include participants from different countries or with varying languages or learning ability. The effect of AI tools on participants in an international setting should also be examined. As mentioned above, the low performers in this study still had higher SAT/ACT scores than the US average SAT score, potentially indicating the presence of an additional group of low performers below those of our study. The effect should be studied in this group as they may have more potential to benefit from the AI tools. Studying these tools in individuals with below-average SAT/ACT scores or from educationally disadvantaged communities may provide an opportunity to explore AI's impact on a wider range of learners. Additionally, future work should examine the effects of AI tools that vary in their implementation or prompts compared to those used in this study. Finally, longitudinal research over a longer period of time should be conducted to reliably test the effect of AI-based tools on learning in the long term, as participants in our study interacted with each condition once.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

The studies involving humans were approved by Advarra Institutional Review Board (the protocol was determined to be exempt). The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

Author contributions

HE: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. KE: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. RC: Supervision, Conceptualization, Writing – review & editing, Validation, Investigation, Methodology. CR: Formal analysis, Writing – original draft, Writing – review & editing, Conceptualization, Supervision, Validation, Methodology.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

Acknowledgments

We thank Joshua Jordan for his comments on the manuscript.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The authors declare that no Gen AI was used in the creation of this manuscript.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/feduc.2025.1506752/full#supplementary-material

References

Abbas, M., Jam, F. A., and Khan, T. I. (2024). Is it harmful or helpful? Examining the causes and consequences of generative AI usage among university students. *Int. J. Educ. Technol. High. Educ.* 21:10. doi: 10.1186/s41239-024-00444-7

ACT (2018) ACT SAT concordance tables. ACT, Inc. Available at: https://www.act.org/content/dam/act/unsecured/documents/ACT-SAT-Concordance-Tables.pdf (accessed June 14, 2024).

Aleven, V., Stahl, E., Schworm, S., Fischer, F., and Wallace, R. (2003). Help seeking and help Design in Interactive Learning Environments. *Rev. Educ. Res.* 73, 277–320. doi: 10.3102/00346543073003277

Bigozzi, L., Tarchi, C., Vagnoli, L., Valente, E., and Pinto, G. (2017). Reading fluency as a predictor of school outcomes across grades 4–9. *Front. Psychol.* 8:200. doi: 10.3389/fpsyg.2017.00200

Bloom, B. S. (1984). The 2 sigma problem: the search for methods of group instruction as effective as one-to-one tutoring. *Educ. Res.* 13, 4–16. doi: 10.2307/1175554

Calhoun, D. H. (1996) 'Which "Socratic Method"? Models of Education in Plato's Dialogues', in K. Lehrer et al. (eds) Knowledge, Teaching and Wisdom. Dordrecht: Springer Netherlands, 49–70.

CollegeBoard (2023) SAT suite of assessments annual report. CollegeBoard. Available at: https://reports.collegeboard.org/media/pdf/2023-total-group-sat-suite-of-assessments-annual-report%20ADA.pdf (accessed June 14, 2024).

Coyle, T. R., and Pillow, D. R. (2008). SAT and ACT predict college GPA after removing g. *Intelligence* 36, 719–729. doi: 10.1016/j.intell.2008.05.001

Črček, N., and Patekar, J. (2023). Writing with AI: university students' use of ChatGPT'. *J. Lang. Educ.* 9, 128–138. doi: 10.17323/jle.2023.17379

Crompton, H., and Burke, D. (2023). Artificial intelligence in higher education: the state of the field. *Int. J. Educ. Technol. High. Educ.* 20:22. doi: 10.1186/s41239-023-00392-8

Dixon-Roman, E. J., Everson, H. T., and Mcardle, J. J. (2013). Race, poverty and SAT scores: modeling the influences of family income on black and white high school students' SAT performance'. *Teach. Coll. Rec.* 115, 1–33. doi: 10.1177/016146811311500406

Graesser, A. C., Conley, M. W., and Olney, A. (2012). "Intelligent tutoring systems" in APA educational psychology handbook, Vol 3: Application to learning and teaching (Washington, DC, US: American Psychological Association (APA handbooks in psychology®)), 451–473.

Graesser, A. C., Lu, S., Jackson, G. T., Mitchell, H. H., Ventura, M., Olney, A., et al. (2004). AutoTutor: a tutor with dialogue in natural language. *Behav. Res. Methods Instrum. Comput.* 36, 180–192. doi: 10.3758/BF03195563

Grassini, S. (2023). Shaping the future of education: exploring the potential and consequences of AI and ChatGPT in educational settings. *Educ. Sci.* 13:692. doi: 10.3390/educsci13070692

Hadi Mogavi, R., Deng, C., Juho Kim, J., Zhou, P., Kwon, Y. D., Hosny Saleh Metwally, A., et al. (2024). ChatGPT in education: a blessing or a curse? A qualitative study exploring early adopters' utilization and perceptions. *Comput. Human Behav.* 2:100027. doi: 10.1016/j.chbah.2023.100027

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., et al. (2023). Survey of hallucination in natural language generation. *ACM Comput. Surv.* 55, 1–38. doi: 10.1145/3571730

Krug, D., George, B., Hannon, S. A., and Glover, J. A. (1989). The effect of outlines and headings on readers' recall of text. *Contemp. Educ. Psychol.* 14, 111–123. doi: 10.1016/0361-476X(89)90029-5

Mahmud, L., and Tryana, T. (2023). Promoting Reading comprehension by using Socratic questioning. *Jurnal Onoma: Pendidikan, Bahasa, dan Sastra* 9, 218–226. doi: 10.30605/onoma.v9i1.2221

OpenAIAchiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., et al. (2023) GPT-4 technical report. arXiv.org. arXiv:2303.08774. Available at: https://arxiv.org/abs/2303.08774v6 (accessed June 13, 2024).

Paladines, J., and Ramirez, J. (2020). A systematic literature review of intelligent tutoring systems with dialogue in natural language. *IEEE Access* 8, 164246–164267. doi: 10.1109/ACCESS.2020.3021383

Pappas, M., and Drigas, A. (2016). Incorporation of artificial intelligence tutoring techniques in mathematics. *Int. J. Eng. Pedag.* 6, 12–16. doi: 10.3991/ijep.v6i4.6063

Ruan, S., Nie, A., Steenbergen, W., He, J., Zhang, J. Q., Guo, M., et al. (2024). Reinforcement learning tutor better supported lower performers in a math task. *Mach. Learn.* 113, 3023–3048. doi: 10.1007/s10994-023-06423-9

Skinner, B. F. (1961). Teaching Machines. *Sci. Am.* 205, 90–102. doi: 10.1038/scientificamerican1161-90

Thomas, D. R., Lin, J., Gatz, E., Gurung, A., Gupta, S., Norberg, K., et al. (2024). "Improving student learning with hybrid human-AI tutoring: a three-study quasi-experimental investigation" in Proceedings of the 14th learning analytics and knowledge conference (New York, NY, USA: Association for Computing Machinery (LAK '24)), 404–415.

U.S. Department of Education, Institute of Education Sciences. (2024) College enrollment rates. Available at: https://nces.ed.gov/programs/coe/indicator/cpb (accessed June 14, 2024).

Vargas-Murillo, A. R., Pari-Bedoya, I. N. M. d. l. A., and Guevara-Soto, F. d. J. (2023). 'Challenges and opportunities of AI-assisted learning: a systematic literature review on the impact of ChatGPT usage in higher education', international journal of learning. *Teach. Educ. Res.* 22, 122–135. doi: 10.26803/ijlter.22.7.7

von Garrel, J., and Mayer, J. (2023). Artificial intelligence in studies—use of ChatGPT and AI-based tools among students in Germany. *Humanit. Soc. Sci. Commun.* 10, 1–9. doi: 10.1057/s41599-023-02304-7

Xu, Z., Wijekumar, K., Ramirez, G., Hu, X., and Irey, R. (2019). The effectiveness of intelligent tutoring systems on K-12 students' reading comprehension: a meta-analysis. *Br. J. Educ. Technol.* 50, 3119–3137. doi: 10.1111/bjet.12758

Yang, Y.-T. C., Newby, T. J., and Bill, R. L. (2005). Using Socratic questioning to promote critical thinking skills through asynchronous discussion forums in distance learning environments. *Am. J. Dist. Educ.* 19, 163–181. doi: 10.1207/s15389286ajde1903_4