ABSTRACT

Title of Thesis: EFFECTS OF ARTIFICIAL INTELLIGENCE GENERATED STUDY TOOLS ON LEANRING AND METACOMPREHENSION

Isabella Rose Sky Klopukh
Master of Science, 2026

Thesis Directed by: Professor Michael Dougherty
Department of Psychology

Recent advances in artificial intelligence technologies such as ChatGPT have provided new tools for students to implement when studying text materials. However, little empirical research has been conducted to evaluate whether use of such technologies is actually beneficial for learning and metacognitive outcomes. The present study connects the fields of artificial intelligence and metacognition to evaluate the effects of artificial intelligence-generated study materials on memory, comprehension, and metacognitive monitoring. Participants read a series of text passages about the natural and social sciences, made metacomprehension judgments, studied the text passages either by re-reading, reviewing a non-interactive ChatGPT-generated study guide, or reviewing an interactive ChatGPT-generated study guide with fill-in-the-blank components, and finally took a multiple-choice test assessing their memory and comprehension of the texts. While reviewing ChatGPT study guides did not improve overall, memory, or comprehension test performance above and beyond re-reading, participants reported greater enjoyment and engagement with reviewing

interactive and non-interactive ChatGPT study guides as compared to re-reading text passages. For participants who reviewed interactive ChatGPT study guides, there was a boost in test performance on questions which matched fill-in-the-blank components and a reduction in overconfidence for metacognitive judgments, resulting in better absolute and relative metacomprehension accuracy compared to participants who re-read text passages or reviewed non-interactive ChatGPT study guides. Overall, results suggest that artificial intelligence generated study materials do not improve test performance relative to re-reading, but do provide some benefits for metacomprehension accuracy.

EFFECTS OF ARTIFICIAL INTELLIGENCE GENERATED STUDY TOOLS ON LEARNING AND METACOMPREHENSION

by

Isabella Rose Sky Klopukh

Thesis submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Master of Science
2026

Advisory Committee:
Professor Michael Dougherty, Chair
Professor Paul Hanges
Dr. Marissa Hartwig

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

Since the release of OpenAI's ChatGPT in November of 2022, artificial intelligence (AI) has rapidly infiltrated the every-day lives of individuals all around the world. From integration with major search engines like Google, to taking over entry-level and administrative tasks in the workplace, it is nearly impossible to avoid AI. Advances in AI technologies have also provided a plethora of new tools for students to implement in their academic work. Therefore, it is a crucial time to investigate the benefits (or detriments) of AI usage in academic settings. The goal of the present study is to empirically assess the usefulness of AI tools for learning and studying text-based materials.

Artificial Intelligence in Education

Usage Among Students

Several large, global surveys have reported remarkably high rates of AI usage among students. Amoah and colleagues (2025) found that over 70% of college students across 109 developed and developing nations reported using ChatGPT, with this percentage even higher (74%) for developed nations specifically. Similarly, 74% of college students in Brazil, India, Japan, the United Kingdom, and the United States reported a willingness to use ChatGPT for academic purposes despite potential ethical implications, with students most commonly citing that they believe it improves their skills and saves time (Ibrahim et al., 2023). Another survey of 516 Taiwanese undergraduate students found that ChatGPT was viewed as a "tool" and "tutor" for academic work, with over 50% of students reporting using it at least once a week (Yu et al., 2024). Ibrahim and colleagues (2023), in addition to surveying student and educator attitudes towards

ChatGPT, compared the AI's ability to answer sample test questions created by instructors for a variety of university courses. Performance of ChatGPT on these questions was comparable to that of real students in disciplines such as Biopsychology, Management, Public Policy, Data Science, and Civil Engineering (Ibrahim et al., 2023). While this finding suggests that AI can access sufficient knowledge to assist students and educators across many domains, it also contributes to concerns that students can use AI tools like ChatGPT in lieu of doing their own work (and succeed).

Benefits for Learning

As AI has become increasingly capable and popular among students, researchers have sought to better understand its potential applications in education. Many studies over the last few years have investigated the effects of AI-based interventions on student learning outcomes and perceptions, often with conflicting findings. In fields such as computer science, AI-powered chatbots have been found to be as good or better than human teaching assistants at helping students solve coding problems and build on their knowledge (e.g., Lee et al., 2023; Liu et al., 2024). Meanwhile, other studies have found no difference in task performance between students who used ChatGPT for assistance with programming assignments and those who did not (e.g., Johnson et al., 2024; Kosar et al., 2024). Similarly, conflicting findings have been obtained with regard to students' perceptions of their own learning; Johnson and colleagues (2024) reported significantly lower programming self-efficacy scores for students who used ChatGPT, while Urban and colleagues (2024) reported significantly higher problem-solving self-efficacy scores for students who used ChatGPT.

A recent meta-analysis by Wang and Fan (2025) comprised 51 experimental and quasi-experimental studies on ChatGPT as a learning tool in educational settings. Overall, there was a strong, positive effect of ChatGPT usage on students' learning performance, learning perceptions,

and higher-order thinking. Several moderating factors, including the role of ChatGPT (as an intelligent tutor, learning tool, conversational partner), type of course (STEM, language learning/writing, skills and competencies), and the learning model used (tailored to students' needs, solving authentic real-world problems, engaging in extended projects) were not found to impact the effectiveness of ChatGPT on enhancing learning outcomes or perceptions. However, Wang and Fan's (2025) meta-analysis also highlights a number of limiting factors. First, many of the included studies are severely underpowered, with some sample sizes as low as just nine participants per experimental group. Second, most of the studies take place in real educational settings, which provide high levels of external validity, but also results in a lack of control as to how participants are actually using ChatGPT throughout the study. Finally, studies both within and beyond this meta-analysis primarily focus on the application of ChatGPT in skill- and rule-based learning contexts such as programming, language learning, academic writing, and teaching. There has been little to no exploration of AI tools for coursework in fields like the social sciences, which often involve less project-based or problem-solving outcomes and more passive learning of text-based materials for evaluation of memory and understanding on traditional tests. The present study seeks to address some of these gaps by integrating the study of AI tools for learning with the field of metacomprehension.

Metacomprehension of Text Materials

Preparing for exams often requires college students to read and study text-based materials (e.g., course textbooks). It is important that students are able to not only remember key information and draw inferences from the material, but also to accurately assess their own learning in order to make informed decisions about how to study. This is referred to as metacomprehension, a form of metacognitive monitoring specifically pertaining to the understanding of text materials (Maki

& Berry, 1984). Several studies have found that students use metacomprehension to guide their learning by regulating study decisions (Dunlosky et al., 2005; Thiede et al., 2003). However, metacomprehension accuracy, the degree to which students' judgments of their understanding are predictive of their actual performance on some evaluation of understanding (i.e., a comprehension test) for text materials, is often very poor (Maki, 1998). Therefore, researchers have investigated various methods to try to improve students' metacomprehension accuracy.

The Rereading Effect

In a typical metacomprehension task, participants read a series of text passages, make judgments of their understanding, and then take a test. One of the simplest interventions that has been implemented to improve metacomprehension is re-reading. Rawson and colleagues (2000) assigned half of their participants to read each text passage twice prior to rating their own comprehension, followed by a multiple-choice test. Participants who read text passages twice compared to participants who read text passages just once displayed significantly improved metacomprehension accuracy in the form of the correlation between comprehension ratings and test performance for each individual (referred to as relative accuracy). This finding, which Rawson and colleagues (2000) termed "the rereading effect", was successfully replicated in a follow-up study with slightly different procedures and materials.

The theoretical underpinnings of the rereading effect stem from the intersection of theories of metacognitive monitoring and theories of text comprehension. The construction-integration theory of comprehension put forth by Kintsch (1988) proposes that mental representations of text materials are built across three levels: the lexical level (surface features and syntactic information), the textbase level (explicit semantic meaning), and the situation model (the integration of meaning

with one's own prior knowledge and experiences). Performance on tests of comprehension of text materials has been found to be largely related to the quality of the situation model constructed (Kintsch, 1994). When initially reading a text passage, cognitive resources are primarily dedicated to processing information at the lexical and textbase levels. At the time of the second reading, the individual can access previously constructed basic representations of the text and devote more cognitive resources to construction of the situation model, thereby improving comprehension.

Theories of metacognitive monitoring propose that metacomprehension judgments are inferential in nature, based on a variety of cues such as reading fluency, prior knowledge, and accessibility of information (Dunlosky, 2005; Rawson et al., 2000). The degree to which these cues are aligned with evaluations of comprehension determines the accuracy of metacomprehension judgments. One cue which has been shown to be influential when making metacomprehension judgments is the amount of disruptions in comprehension processes that occur when reading a text (Rawson & Dunlosky, 2002). Disruptions in processing can occur at any level of the construction-integration model, but are most accurate as cues when processing text at the situational level due to its importance for comprehension. Therefore, the disruption cues to metacomprehension that may be accessed at the time of initial reading are less strongly associated with comprehension, while the disruption cues at the time of re-reading are more likely to come from the situational level and are therefore more strongly associated with comprehension, thus forming the rereading effect (Dunlosky, 2005; Rawson & Dunlosky, 2002; Rawson et al., 2000).

While several studies have successfully replicated the rereading effect (e.g., Chiang et al., 2010; Dunlosky, 2005; Griffin et al., 2008), other recent re-evaluations have produced inconsistent results. For example, Margolin and Snyder (2018) found that while relative metacomprehension

accuracy increased with re-reading for texts with added negations (e.g., no, not, etc. which has been found to make texts more difficult to read), this was not the case for texts in their original form; in particular, the correlations between metacomprehension judgments and test performance for standard text passages read twice were not significantly different from zero, despite being significant for standard text passages read only once. Wong and Moss (2022) failed to find a rereading effect at all, with no significant difference in metacomprehension accuracy between participants who read passages once and participants who read passages twice. One possible explanation for this discrepancy is in the stimuli used, with several studies that found a rereading effect using the exact same text materials (e.g., Chiang et al., 2010; Dunlosky, 2005; Rawson et al., 2000), while more recent studies that failed to find a rereading effect such as those by Margolin and Snyder (2018) and Wong and Moss (2022) using new stimuli. If text passages are easy enough to comprehend on the first read, then the metacognitive cues may already be based on the situation level of comprehension processing, adding no benefit when reading a second time. Inversely, if text passages are more difficult, then cognitive resources may still be allocated to processing at the lexical and textbase levels even during the second reading, resulting in metacognitive cues at these levels which are less predictive of performance.

Other Interventions for Improving Metacomprehension

The rereading effect is not the only way in which researchers have been able to (at least some of the time) improve metacomprehension accuracy. Griffin and colleagues (2008) assigned one group of participants to a self-explanation condition in which they received additional instructions prior to reading text passages for the second time to reflect on the meaning and contributions of each paragraph to the text passages as a whole. The self-explanation instruction improved

metacomprehension accuracy above and beyond simply re-reading text passages (Griffin et al., 2008). Manipulating participants' expectations of the type of test they will be taking also affects the accuracy of their metacomprehension judgments. Across multiple experiments, Thiede and Griffin (2011) and Griffin and colleagues (2019) found that students who were told they would be taking tests that assessed their ability to make inferences about information in the text passages, and provided with practice test items, showed improved metacognitive accuracy for inference-based tests compared to memory-based tests (with the opposite being true for students instructed that they would receive memory-based tests regarding explicit details from the text passages).

Other interventions aimed at improving metacomprehension accuracy have focused on more active, generation-based processes. For example, studies have found that when students write summaries of text passages prior to making metacomprehension judgments, metacomprehension accuracy is improved (Anderson & Thiede, 2008; Thiede & Anderson, 2003). Further research suggests that writing full summaries are not required to reap this benefit; generating keywords that capture the essence of text passages prior to making metacomprehension judgments can also improve their accuracy (Thiede et al., 2003, 2005). Finally, asking participants to explicitly practice recall of information from text passages prior to making metacognitive judgments has been found to increase the accuracy of those judgments (Dunlosky et al., 2005).

Facilitating Learning of Text Materials

Findings from Metacomprehension Research

While improving metacomprehension accuracy can be important, improving actual learning of text materials and test performance is also of great interest to researchers. In many studies, re-

reading text passages has been found to not only improve metacomprehension accuracy, but also to enhance performance on tests of text materials (e.g., Margolin & Snyder, 2018; Rawson & Kintsch, 2005; Rawson et al., 2000; Wong & Moss, 2022). It should be noted, though, that this is not always the case; other studies have found no difference in test performance between participants who read text passages once and participants who read text passages twice (e.g., Dunlosky, 2005; Griffin et al., 2008), suggesting that the benefits of re-reading text passages for test performance may be moderated by other factors such as the specific text materials used and sample characteristics like comprehension ability or working memory capacity.

The Testing Effect

Independent of metacomprehension, researchers have examined the effects of various interventions aimed at improving students' memory for and comprehension of text materials. One prominent method of improving long-term retention for information from text passages is the testing effect, in which taking tests or practicing retrieval of information while learning text materials results in better test performance later on than simply reviewing those materials (Roediger & Karpicke, 2006). The testing effect has been found across a variety of material types, test types, and contexts, including with real learning materials in classroom settings beyond the laboratory (Dirkx et al., 2014; McDaniel et al., 2007). In an example of particular relevance to the present study, Hinze and Wiley (2011) found some success with using fill-in-the-blank tests as retrieval practice for text materials. However, this study also highlights a key limitation of the testing effect, which is that it often only applies when identical items are used for retrieval practice and actual tests. Additionally, some evidence suggests that the testing effect dissipates when the complexity of text materials is very high (see van Gog & Sweller, 2015, for a review and meta-analysis).

Outlining and The Adjunct Display Effect

Another line of work which is highly relevant to the present study focuses on the outlining of text materials. Outlines are succinct, hierarchically-structured and linearly organized documents which summarize the key points of text materials. The two most common forms of outlining research examine the effects of generating outlines of text materials as part of the learning process and the effects of presenting outlines alongside text materials (also referred to as an "adjunct display"). Learner-generated outlines are thought to activate a multitude of cognitive processes including selecting, organizing, and integrating information from text passages in order to promote effective learning and test performance. Instructor-provided outlines are thought to engage similar processes but in a passive manner, with learners being directed to important information and given the organizational structure. In a comprehensive meta-analysis of the outlining literature, Ponce and colleagues (2023) found that both learner-generated and instructor-provided outlines enhanced learning outcomes (including both memory and comprehension test performance) in college students compared to only reviewing text materials. Given these benefits, tools that can efficiently generate outlines may provide a useful resource for students to support their own learning.

The Present Study

Popular AI chatbots like Google's NotebookLM have built-in features that can quickly create various types of outlines (commonly referred to as "study guides") for text-based source materials. Students are already leveraging such tools to help break-down complex material and evaluate their own understanding, but it is unknown to both students and teachers alike whether using AI is to the benefit or detriment of students' learning and, of course, their performance on a test. In the present

study, we sought to bring together research on AI and metacomprehension to evaluate the effects of AI-generated study guides on students' memory, comprehension, and metacognitive accuracy. Participants read a series of text passages on various topics in the natural and social sciences, made metacomprehension judgments, studied the text passages either by re-reading or reviewing a ChatGPT-generated study guide, made a second round of metacomprehension judgments, and finally took a multiple choice test comprised of both memory-based and comprehension- or inference-based questions. Two forms of study guides were generated using AI: non-interactive study guides, which were simple, outline-style, hierarchically organized documents based on text passages; and interactive study guides, which were adaptations of non-interactive study guides with embedded multiple-choice fill-in-the-blank components in place of key words and phrases. These study guides represent two different ways students might choose to approach AI-assisted studying, either passively by reviewing some output or actively by having to engage with that output.

To address our research questions, we analyzed participants' overall test performance, memory-specific test performance, comprehension-specific test performance, relative and absolute metacognitive accuracy, trust in artificial intelligence, and use of artificial intelligence. We hypothesized that overall test performance (including both comprehension and memory test items), comprehension-specific test performance, and memory-specific test performance would be better for participants who reviewed either interactive or non-interactive ChatGPT study guides compared to participants who only re-read text passages. Study guides are thought to help participants identify the most relevant information from texts and provide an organizational structure for that information, both of which have been previously shown to improve test performance as per the adjunct display effect (Ponce et al., 2023). We also expected there to be an improvement in test

performance for participants who engaged with interactive ChatGPT study guides compared to those who reviewed non-interactive ChatGPT study guides due to retrieval practice instigated by completing interactive components, creating something akin to a testing effect (Hinze & Wiley, 2011; Roediger & Karpicke, 2006), or the act of having to complete interactive components drawing more attentional resources to the study guide itself.

However, it is also possible that participants may focus solely on the interactive components of the study guides and ignore the remainder of the content, resulting in a decrement in performance compared to participants who review non-interactive study guides. To investigate whether ChatGPT-generated study guides are only beneficial for learning or remembering the exact information they contain, as opposed to facilitating learning for the entirety of the material, we tested the following series of research questions: Are there differences in memory-specific test performance for questions whose answers are versus are not explicitly stated in the ChatGPT study guides? Are there differences in memory-specific test performance for questions whose answers are versus are not included in interactive components of the ChatGPT study guides? And, does this effect depend on whether the interactive components corresponding to those questions were answered correctly during the study phase?

With regard to metacognitive accuracy, we hypothesized that the accuracy of metacomprehension judgments would improve between the initial reading phase and the study phase for all three study conditions. This prediction is based on an extension of the re-reading effect, in which the second presentation of text passages either on their own or alongside study guides would allow for readers to construct a better situation model and thereby access metacognitive cues that are more predictive of later test performance (Dunlosky, 2005; Kintsch, 1988; Rawson & Dunlosky, 2002;

Rawson et al., 2000). We further hypothesized that this improvement in metacognitive accuracy would be greater for participants who reviewed non-interactive or interactive ChatGPT study guides compared to those who only re-read text passages in line with findings regarding recall practice by Dunlosky and colleagues (2005).

Finally, we hypothesized that participants' level of trust in AI would be related to both their test performance and metacognitive accuracy in either the non-interactive or interactive ChatGPT study guide conditions. This analysis was largely exploratory, as prior research has independently examined AI trust (e.g., Hoffman et al., 2023; Perrig et al., 2023) and the effects of AI on learning (e.g., Wang & Fan, 2025). It is conceivable that participants who are hesitant to trust AI may not be as willing to rely on the provided ChatGPT study guides to review the information from the text passages, and therefore would not see any change in their test performance or metacognitive accuracy as a result. We do not expect such a relationship for participants in the re-read condition, as they do not have any contact with AI-related materials throughout the experiment. However, it is also possible that the relationship between AI trust and test performance or metacognitive accuracy persists across all conditions, in which case the measure of trust may be related to some other underlying psychological construct that also predicts test performance and metacognitive accuracy. For example, Agler and colleagues (2021) found that Openness to Experience (one of the Big Five personality traits) was positively related to both comprehension performance and metacomprehension accuracy, while Böckle and colleagues (2021) and Oksanen and colleagues (2020) found that Openness to Experience was positively related to trust in AI systems.

Chapter 2: Methods

This study, including its hypotheses, methodology, and general analysis plan, was pre-registered on the Open Science Framework (https://osf.io/4zqs5). All materials, including text stimuli, test questions, and survey items are also available on the Open Science Framework.

Participants

Undergraduate students enrolled in psychology courses at the University of Maryland, College Park (N = 279) were randomly assigned to either re-read text passages (Re-Read), review non-interactive ChatGPT study guides (Non-Interactive), or review interactive ChatGPT study guides (Interactive). Six participants failed to complete the entirety of the experiment and were thus excluded from analysis. The final sample consisted of 273 participants, with 61.17% identifying as female, 38.09% identifying as male, 0.37% identifying as non-binary, and 0.37% identifying as other. In addition, 41.39% identified as White, 23.44% identified as Black or African American, 20.51% identified as Asian, 9.89% identified as Hispanic or Latino, 0.37% identified as American Indian or Alaskan Native, and 4.40% identified as other. Participants ranged from 18 to 23 years of age (M = 18.96, SD = 1.14).

Materials

Text Passages

Six expository text passages on various topics in the natural and social sciences were adapted from Griffin and colleagues (2019). Text passages were originally between 549–1086 words with varying levels of difficulty in terms of grade level, reading ease, and words per sentence. Text passages were re-written with the help of Perplexity on March 4th, 2025 to be "at a consistent

length (850–1000 words) and consistent difficulty level (college)". Text passages were re-generated by Perplexity until they matched the length and difficulty requirements as closely as possible, sometimes requiring re-prompting to remove imposed organizational structures or missing information, then manually revised to ensure that all critical content for answering test questions (in the manner described below) was present. The final versions of the text passages ranged from 873–937 words in length with similar readability levels, allowing for consistency in reading times across texts.

ChatGPT Study Guides

For each text passage, two study guides were created: a non-interactive, hierarchically organized outline-style study guide, and an interactive version. To create non-interactive study guides, revised text passages were pasted into ChatGPT-4o on March 4th, 2025 with the prompt, "generate a study guide based on the following text". ChatGPT was blind to the nature and content of test questions in an attempt to reduce potential bias in the study guides and increase external validity (as students typically do not know the exact content of tests when creating study materials). Minimal manual revisions were made; titles and bold or italic formatting within text were removed for consistency across passages, and small words such as "e.g." and "via" were added or removed for grammatical correctness. Study guides were also modified to ensure that they contained the same information as the text passages needed to answer all (five) of the inference-based test questions, half (four) of the memory-based test questions, and excluded the information in the text passages needed to answer the other half (four) of the memory-based test questions.

ChatGPT was then used to create interactive study guides, which were designed to encourage participants to engage with the material more than they might naturally if they were simply reviewing a non-interactive study guide. The non-interactive study guides were pasted into ChatGPT-4o on

March 23rd, 2025 with the prompt, “Edit the below study guide to remove key words and replace them with 4 alternative choices for each word (while keeping the rest exactly the same)”. In the experiment, interactive study guides were formatted such that each interactive component consisted of a blank space highlighted in blue on the study guide in place of one to three words. Selecting a dropdown arrow on the interactive component revealed 4 alternatives which could be used to fill in that particular blank. Once an alternative was selected, it automatically populated in the study guide in place of the blank space while remaining highlighted.

Due to the state of Large Language Models at the time of stimuli creation, more extensive revisions were required for the interactive study guides such as excluding blanks that did not have a correct answer option, had answer options that were too similar, or had words that were redundant (e.g., in the title, heading, or prior line). For the sake of time and consistency across passages, interactive study guides were reduced to have only 20 interactive components each by removing multiple interactive components in a sentence or list, as well as removal by section to ensure an approximately even distribution across study guides. Of the four memory-based questions whose answers were explicitly included in study guides, the answers to half (two) of those questions were included in interactive keyword components, while answers to the other half (two) of those questions were only included in their standard non-interactive form. The answers to inference-based questions were not always contained in the interactive components, as they often required multiple pieces of information and were not explicitly stated in texts or study guides, and in total ranged from 0–3 relevant interactive keyword components per question.

Test Questions

There were a total of 13 four-alternative multiple-choice test questions corresponding to each text passage. Eight questions corresponding to each passage were memory-based, regarding explicit details from the text passages, and five questions corresponding to each passage were comprehension-based, requiring connections to be made between concepts in the text passages that were not explicitly stated (these items are also commonly referred to as inference-based questions in the metacomprehension literature). All five comprehension-based questions and five of the eight memory-based questions were taken verbatim from Griffin and colleagues (2019). The three additional memory-based questions were created to conform to the structure of the existing test questions with the help of ChatGPT-4o on March 25th, 2025. As previously mentioned, half (four) of the eight memory-based questions could be answered directly from the ChatGPT study guides, and half (two) of those four memory-based questions could be answered directly from interactive components of ChatGPT study guides.

Procedure

Design & Randomization

A 3x6 mixed design was used. Participants were randomly assigned to one of three between-subjects study conditions (Re-Read: n = 92; Non-Interactive: n = 91; Interactive: n = 90). The repeated measure consisted of the six text passages presented to each participant in a random order, with the order of passages remaining consistent within participants throughout the experiment for the study and test phases. The 13 test questions for each text passage were presented in a random order for each participant. Answer choices for each test question were randomly ordered for each

participant, with the exception of questions whose answer choices followed a numerical order, had a final option of "all of the above", or had a final option combining two prior options such as "both A and B". For interactive ChatGPT study guides, multiple-choice options for interactive components were also randomly ordered for each participant following the same criteria as test questions.

Reading Phase

Participants were given five minutes to read each text passage. Texts were presented one at a time on a computer screen, and participants could not advance to the next passage until the five minutes were up (see Figure 1 for an example). The topic of the text passage was written in bold at the top of the page. A count-down timer was visible in the top right corner of the screen during all timed portions of the experiment. Immediately after reading each text passage, participants were asked to provide two forms of metacomprehension ratings: first, a judgment of understanding which asked, "How well do you understand the text about [topic]?" on a scale of 1–7, with 1 being "Not well at all" and 7 being "Very well"; and second, a prediction of performance which asked, "Imagine you were given a 100-question multiple-choice test on the text about [topic]. What percent of the questions do you think you would get right?" on a scale of 0–100, in increments of 1.

Study Phase

After reading and providing metacomprehension judgments for all six text passages, participants repeated the general procedure for their assigned study condition. In the Re-Read condition, participants were simply instructed to re-read each text passage. In the Non-Interactive and Interactive conditions, participants were instructed to "review each ChatGPT study guide and text passage", which were presented simultaneously, with text passages on the left side of the screen and

study guides on the right (see Figures 2 and 3 for examples of the Non-Interactive and Interactive conditions, respectively). In the Interactive condition, participants were additionally told, "There will be blank spaces highlighted in blue for you to fill in with the correct information. . . select the option you feel best fits into the study guide based on the text passage"; participants were not forced to respond to any or all of the blank spaces in interactive ChatGPT study guides. Participants were given 5 minutes to study each text passage in the manner assigned to them, and each text was again followed by two metacomprehension judgments as in the reading phase.

Test Phase

For each text passage, participants were presented with the entire set of 13 test questions (including both memory-based and comprehension-based items) on the same page and given an unlimited amount of time to answer them. The topic of the text which corresponded to each set of questions was denoted at the top of the page. Participants were required to provide a response to all test questions, and could not go back to change their answers once they moved on to the next set. Prior to the test, participants were not given any details regarding the content or nature of the test questions, aside from them being multiple-choice.

Self-Report Measures

In the last portion of the experiment, participants were given several self-report measures to complete. First, they were asked to rate how knowledgeable they felt they were on each of the text passage topics prior to the experiment on a scale of 1–5, with 1 being "No knowledge" and 5 being "Expert knowledge". Participants then completed a two-part scale about their perceptions of the study condition they were assigned to (either re-reading text passages, reviewing non-interactive

ChatGPT study guides, or reviewing interactive ChatGPT study guides). The first part of the scale consisted of three items which asked participants about their level of enjoyment and engagement while partaking in their respective study conditions, and whether they would use a similar study strategy in their own work, on a 5-point scale from "Strongly Disagree" to "Strongly Agree". The second part consisted of four items which asked how helpful or detrimental the study condition was for participants' ability to understand the texts, answer test questions, remember specific details, and make connections between concepts on a 5-point scale from "Very Detrimental" to "Very Helpful".

Finally, participants were asked to complete two AI-related measures. The first measure asked about their use of AI for academic purposes, including whether they had used it (yes/no), how frequently they used it (never/rarely/monthly/weekly/daily), average hours of use per week (0–20), and preference for specific chatbot models. The second was a measure of trust in AI adapted from the TXAI scale by Hoffman and colleagues (2023) using Perrig and colleagues' (2023) recommendations for optimizing psychometric properties. This measure consisted of seven statements about the reliability, accuracy, and usefulness of AI, which participants rated on a 5-point scale from "I disagree strongly" to "I agree strongly".

Statistical Analyses

All primary analyses were conducted using Bayesian regression models with the package brms in the R Statistical Environment (Bürkner, 2017; R Core Team, 2025). Pairwise comparisons were performed using the hypothesis() function from the brms package. All comparisons were non-directional. The 95% credible intervals (CIs) provided by the hypothesis() function were used as inference criteria; CIs that did not contain zero were considered to indicate a credible difference

between the groups being compared. Bayes Factors were computed using the Savage-Dickey density ratio to quantify the evidence in support of the alternative relative to the null hypothesis. For example, a BF_{10} of 10 would indicate that the data are 10 times more likely to be observed under the alternative hypothesis, whereas a BF_{10} of $1/10$ would indicate that the data are 10 times more likely to be observed under the null hypothesis.

Chapter 3: Results

Model Fitting

Bayesian regression models were run using 8 chains and 2,000 non-burn-in iterations per chain, resulting in a total of 16,000 posterior samples for each model. There were no divergences in any of the models. All R-hats were close to 1. Trace plots indicated that chains mixed well for all models. The minimum bulk effective sample size for a given model was 1,841, while the minimum tail effective sample size for a given model was 3,464, both well above the general recommendation of at least 100 per chain (or a minimum of 800 for the 8 chains used).

Test Performance

Our first set of hypotheses sought to investigate whether our study condition manipulation had any effect on participants' test performance. Bayesian mixed effects regression models were run predicting overall test performance, comprehension test performance, and memory test performance for each text passage from study condition (three levels: re-read text passages; review non-Interactive ChatGPT study guides; review interactive ChatGPT study guides). An aggregated binomial regression approach was used to appropriately model the discrete, bounded nature of test performance. Participants' self-rated prior knowledge for each text passage topic was standardized and included in the model as a fixed effect to control for its potential influence on performance (means and standard deviations of prior knowledge ratings for each study condition and text passage topic are presented in Table 1). Random intercepts were included for each participant, as well as for each text passage topic to account for potential differences in difficulty. The results of each of these models is presented below.

Overall Test Performance

Our first model examined overall test performance, which was defined as the number of questions answered correctly out of the 13 test questions for each text passage (collapsing across both memory-based and comprehension-based items). Means and standard deviations of overall test performance for each study condition and text passage topic are presented in Table 2. Model estimates are presented in Figure 4a; paired comparisons are visualized in Figure 4b. Prior knowledge had a credibly positive relationship with overall test performance (Estimate = 0.178, 95% CI [0.141, 0.216], $BF_{10} > 1.60 \times 10^4$). There were no credible differences in overall test performance between any of the study conditions (Re-Read vs Non-Interactive: Estimate = 0.128, 95% CI [-0.054, 0.307], $BF_{10} = 1/5.62$; Re-Read vs Interactive: Estimate = 0.021, 95% CI [-0.161, 0.197], $BF_{10} = 1/10.70$; Non-Interactive vs Interactive: Estimate = -0.108, 95% CI [-0.289, 0.075], $BF_{10} = 1/5.79$). These results suggest that using AI-generated study guides did not enhance participants' test performance above and beyond re-reading text passages, and engaging with interactive AI-generated study guides did not enhance participants' test performance above and beyond reviewing non-interactive AI-generated study guides.

Comprehension Test Performance

Comprehension test performance was defined as the number of questions answered correctly out of the five comprehension-based items (those requiring connections to be made between concepts in the text passages that were not explicitly stated). Means and standard deviations of comprehension test performance for each study condition and text passage topic are presented in Table 3. Model estimates are presented in Figure 5a; paired comparisons are visualized in Figure 5b.

Prior knowledge had a credibly positive relationship with comprehension test performance (Estimate = 0.190, 95% CI [0.129, 0.252], $BF_{10} > 1.60 \times 10^4$). There were no credible differences in comprehension test performance between any of the study conditions (Re-Read vs Non-Interactive: Estimate = 0.111, 95% CI [-0.105, 0.330], $BF_{10} = 1/7.90$; Re-Read vs Interactive: Estimate = -0.073, 95% CI [-0.290, 0.139], $BF_{10} = 1/7.56$; Non-Interactive vs Interactive: Estimate = -0.185, 95% CI [-0.404, 0.032], $BF_{10} = 1/2.43$). These results suggest that using AI-generated study guides did not enhance participants' comprehension-specific test performance above and beyond re-reading text passages, and engaging with interactive AI-generated study guides did not enhance participants' comprehension-specific test performance above and beyond reviewing non-interactive AI-generated study guides.

Memory Test Performance

Memory test performance for each text passage was defined as the number of questions answered correctly out of the eight memory-based items (those regarding explicit details from the text passages). Means and standard deviations of memory test performance for each study condition and text passage topic are presented in Table 4. Model estimates are presented in Figure 6a; paired comparisons are visualized in Figure 6b. Prior knowledge had a credibly positive relationship with memory test performance (Estimate = 0.176, 95% CI [0.129, 0.223], $BF_{10} > 1.60 \times 10^4$). There were no credible differences in memory test performance between any of the study conditions (Re-Read vs Non-Interactive: Estimate = 0.139, 95% CI [-0.042, 0.318], $BF_{10} = 1/4.83$; Re-Read vs Interactive: Estimate = 0.074, 95% CI [-0.106, 0.252], $BF_{10} = 1/7.62$; Non-Interactive vs Interactive: Estimate = -0.065, 95% CI [-0.243, 0.115], $BF_{10} = 1/8.28$). These results suggest that using AI-generated study guides did not enhance participants' memory-specific test performance

above and beyond re-reading text passages, and engaging with interactive AI-generated study guides did not enhance participants' memory-specific test performance above and beyond reviewing non-interactive AI-generated study guides.

Study Guide Status. We further sub-divided memory-specific test performance into accuracy for questions whose answers were included or excluded from ChatGPT study guides, as well as questions whose answers were included or excluded from interactive components of ChatGPT study guides, to address our research questions regarding whether ChatGPT-generated study guides are only beneficial for remembering the exact information they contain. A Bayesian mixed effects regression model was run predicting memory test performance from the interaction between study condition and study guide status (two levels: questions present in ChatGPT study guides; questions absent from ChatGPT study guides). The structure of this model was otherwise the same as for the models of overall, comprehension, and memory test performance. Model estimates are presented in Figure 7a. We conducted two sets of post-hoc comparisons to explore the effect of study guide status.

First, as shown in Figure 7b, we compared memory test performance for the four memory-based questions whose answers were present in ChatGPT study guides to the four memory-based questions whose answers were absent from ChatGPT study guides. Performance was greater for questions present in study guides than questions absent from study guides in all study conditions (Re-Read: Estimate = 0.354, 95% CI [0.231, 0.479], $BF_{10} > 1.60 \times 10^4$; Non-Interactive: Estimate = 0.382, 95% CI [0.256, 0.508], $BF_{10} > 1.60 \times 10^4$; Interactive: Estimate = 0.549, 95% CI [0.425, 0.674], $BF_{10} > 1.60 \times 10^4$). These results suggest that the questions whose answers were present in ChatGPT study guides may have been less difficult than the questions whose answers were absent

from ChatGPT study guides, but this was regardless of whether participants re-read text passages or reviewed one of the ChatGPT study guides.

Second, as shown in Figure 7c, we compared memory test performance between study conditions separately for the four memory-based questions whose answers were present in ChatGPT study guides and the four memory-based questions whose answers were absent from ChatGPT study guides. For questions present in study guides, there were no differences in memory test performance (Re-Read vs Non-Interactive: Estimate = 0.126, 95% CI [-0.071, 0.325], $BF_{10} = 1/9.93$; Re-Read vs Interactive: Estimate = -0.027, 95% CI [-0.229, 0.174], $BF_{10} = 1/13.83$; Non-Interactive vs Interactive: Estimate = -0.1531, 95% CI [-0.358, 0.050], $BF_{10} = 1/4.64$). For questions absent from study guides, there were also no differences (Re-Read vs Non-Interactive: Estimate = 0.154, 95% CI [-0.042, 0.352], $BF_{10} = 1/4.66$; Re-Read vs Interactive: Estimate = 0.167, 95% CI [-0.033, 0.366], $BF_{10} = 1/2.69$; Non-Interactive vs Interactive: Estimate = 0.014, 95% CI [-0.193, 0.218], $BF_{10} = 1/9.85$). These results suggest that participants who reviewed ChatGPT study guides did not solely remember the information contained in those study guides, or at least not above and beyond participants who re-read text passages.

Interactive Component Status. A Bayesian mixed effects regression model was run predicting memory test performance from the interaction between study condition and interactive component status (two levels: questions present in interactive components of ChatGPT study guides; questions absent from interactive components of ChatGPT study guides). The structure of this model was otherwise the same as for the models of overall, comprehension, and memory test performance. Model estimates are presented in Figure 8a. We again conducted two sets of post-hoc comparisons to explore the effect of interactive component status.

First, as shown in Figure 8b, we compared memory test performance for the two memory-based questions whose answers were present in interactive components of ChatGPT study guides to the two memory-based questions whose answers were absent from interactive components of ChatGPT study guides (but still present in the non-interactive content of ChatGPT study guides). For participants who re-read text passages, performance was credibly worse for questions present in interactive components compared to questions absent from interactive components (Estimate = -0.196, 95% CI [-0.373, -0.0186], $BF_{10} = 1/1.51$). However, this effect was not particularly convincing according to the Bayes Factor, which indicated that the data were approximately equally likely to be observed under the null as under the alternative hypothesis. For participants who reviewed non-interactive ChatGPT study guides, there was no difference in performance between questions present in interactive components and questions absent from interactive components (Estimate = -0.181, 95% CI [-0.361, 0.001], $BF_{10} = 1/2.23$). For participants who reviewed interactive ChatGPT study guides, performance was credibly better for questions present in interactive components compared to questions absent from interactive components (Estimate = 0.265, 95% CI [0.083, 0.442], $BF_{10} = 6.25$). These results suggest that participants who reviewed interactive ChatGPT study guides received a boost in performance for questions whose answers were present in interactive components.

Second, as shown in Figure 8c, we compared memory test performance between study conditions separately for the two memory-based questions whose answers were present in interactive components of ChatGPT study guides and the two memory-based questions whose answers were absent from interactive components of ChatGPT study guides (but still present in the non-interactive content of ChatGPT study guides). For questions whose answers were present in interactive com-

ponents, there was no difference in memory test performance between participants who re-read text passages and participants who reviewed non-interactive ChatGPT study guides (Estimate = 0.116, 95% CI [-0.138, 0.370], $BF_{10} = 1/6.96$), but performance was credibly better for participants who reviewed interactive ChatGPT study guides compared to both participants who re-read text passages and participants who reviewed non-interactive ChatGPT study guides (Re-Read vs Interactive: Estimate = -0.258, 95% CI [-0.507, -0.002], $BF_{10} = 1/1.11$; Non-Interactive vs Interactive: Estimate = -0.374, 95% CI [-0.624, -0.117], $BF_{10} = 7.14$). However, the difference in performance between participants who reviewed interactive ChatGPT study guides and participants who re-read text passages was not particularly convincing according to the Bayes Factor, which indicated that the data were approximately equally likely to be observed under the null as under the alternative hypothesis. For questions whose answers were present in the non-interactive content of ChatGPT study guides but absent from interactive components of ChatGPT study guides, there were no differences in memory test performance between study conditions (Re-Read vs Non-Interactive: Estimate = 0.131, 95% CI [-0.126, 0.385], $BF_{10} = 1/9.50$; Re-Read vs Interactive: Estimate = 0.203, 95% CI [-0.047, 0.455], $BF_{10} = 1/3.61$; Non-Interactive vs Interactive: Estimate = 0.072, 95% CI [-0.182, 0.330], $BF_{10} = 1/9.87$). These results suggest that, at least to some extent, participants who reviewed interactive ChatGPT study guides performed better than participants who re-read text passages or reviewed non-interactive ChatGPT study guides on questions whose answers were present in those interactive components, but did not perform worse on questions whose answers were not present in those interactive components.

Interactive Study Guide Performance

We initially asked whether differences in memory test performance for questions whose answers were present in interactive components of ChatGPT study guides was dependent on the accuracy with which those specific interactive components were answered during the study phase. However, it is not possible to compare this effect between study conditions because participants who re-read text passages or reviewed non-interactive ChatGPT study guides did not have the opportunity to complete the question-related interactive components in the study phase. Therefore, we instead conducted an exploratory analysis of the relationship between interactive component completion in the study phase and test performance only for participants assigned to review interactive ChatGPT study guides. Means and standard deviations of the number of interactive components completed, the number of interactive components completed correctly, and the number of interactive components completed incorrectly by participants in the Interactive study condition for each text passage topic are presented in Table 5.

Three separate Bayesian mixed effects regression models were run predicting overall test performance (out of 13 questions), comprehension-specific test performance (out of eight questions) and memory-specific test performance (out of five questions) from both the number of interactive components correctly completed in the study phase and the number of interactive components incorrectly completed in the study phase. As participants assigned to review interactive ChatGPT study guides were not forced to respond to all interactive components, controlling for the number of incorrectly completed interactive components in addition to the number of correctly completed interactive components allowed us to account for the total number of interactive components

completed, which varied by participant and text passage. The structures of these models were otherwise the same as for the models of overall, comprehension, and memory test performance.

The relationship between the number of interactive components correctly completed in the study phase and test performance was credibly positive across all three models (Overall test performance: Estimate = 0.051, 95% CI [0.031, 0.071], $BF_{10} > 1.60 \times 10^4$; Comprehension test performance: Estimate = 0.076, 95% CI [0.046, 0.107], $BF_{10} > 1.60 \times 10^4$; Memory test performance: Estimate = 0.051, 95% CI [0.028, 0.072], $BF_{10} = 50$). Results were mixed for the relationship between the number of interactive components incorrectly completed in the study phase and test performance. For overall test performance, this relationship was credibly negative (Estimate = -0.030, 95% CI [-0.056, -0.003], $BF_{10} = 1/7.24$), but the Bayes Factor indicated that the data were over seven times more likely to be observed under the null hypothesis of no relationship. For comprehension test performance, this relationship was not different from zero (Estimate = -0.018, 95% CI [-0.060, 0.026], $BF_{10} = 1/32.51$). Finally, for memory test performance, this relationship was credibly negative (Estimate = -0.054, 95% CI [-0.083, -0.024], $BF_{10} = 8.33$). These results suggest that participants who completed more interactive components correctly in the study phase performed better on all types of test questions, but participants who completed more interactive components incorrectly did not always perform worse.

Metacomprehension Accuracy

Our next set of hypotheses sought to investigate whether metacomprehension accuracy improved between judgments made during the reading phase (time 1) and judgments made during the study phase (time 2). For each time point, two types of metacomprehension judgments were elicited: judgments of understanding, in which participants rated how well they understood each

text passage on a scale of 1 ("Not well at all") to 7 ("Very well"), and predictions of performance, in which participants rated how many questions they thought they would get right on a test out of 100. Means and standard deviations of judgments of understanding and predictions of performance for each time point, study condition, and text passage topic are presented in Tables 6 and 7, respectively. Two different measures of metacomprehension accuracy were used: absolute accuracy, computed as the predicted proportion of questions correct subtracted by the actual proportion of questions correct (including both memory-based and comprehension-based items) for each text passage, and relative accuracy, computed using intraindividual Kendall's tau-a correlations between metacomprehension judgments and overall test performance across text passages. Relative accuracy was computed and analyzed separately for both judgments of understanding and predictions of performance at each time point. Absolute accuracy was only computed and analyzed for predictions of performance at each time point because they were assumed to follow a continuous numeric scale, unlike judgments of understanding, which were measured on a Likert scale and therefore not assumed to have equal distances between scale points. Means and standard deviations of absolute accuracy scores for each timepoint, study condition, and text passage topic are presented in Table 8. Means and standard deviations of relative metacomprehension accuracy scores as measured by Kendall's tau-a for each timepoint, type of judgment, and study condition are presented in Table 9.

Absolute Accuracy

A Bayesian mixed effects linear regression model was run predicting absolute metacomprehension accuracy from study condition, judgment time (two levels: time 1, during the reading phase; time 2, during the study phase), and their interaction.[1] Other effects in this model included the fixed effect of standardized prior knowledge ratings, random intercepts for each participant, and

random intercepts for each text passage topic. Model estimates are presented in Figure 9a; paired comparisons are visualized in Figure 9b. Prior knowledge had a credibly positive relationship with absolute accuracy (Estimate = 0.028, 95% CI [0.020, 0.036], $BF_{10} > 1.60 \times 10^4$). There was a credible difference between absolute accuracy at time 2 and time 1 for the Re-Read condition (Estimate = 0.061, 95% CI [0.041, 0.080], $BF_{10} > 1.60 \times 10^4$) and the Non-Interactive condition (Estimate = 0.072, 95% CI [0.053, 0.092], $BF_{10} > 1.60 \times 10^4$), but not for the Interactive condition (Estimate = 0.001, 95% CI [-0.019, 0.021], $BF_{10} = 1/98.75$). Absolute accuracy is computed such that a score of 0 indicates perfect absolute accuracy (an exact match between predicted and actual performance), a score below 0 indicates underconfidence (lower predicted than actual performance), and a score above 0 indicates overconfidence (higher predicted than actual performance). Based on model estimates, absolute accuracy scores for the Re-Read condition increased from 0.05 at time 1 to 0.11 at time 2 and absolute accuracy scores for the Non-Interactive condition increased from 0.08 at time 1 to 0.15 at time 2, both moving away from zero. Therefore, these results suggest that absolute accuracy became poorer and overconfidence increased following the study phase for participants who re-read text passages or reviewed non-interactive ChatGPT study guides. In contrast, absolute accuracy remained the same for participants who reviewed interactive ChatGPT study guides, with a mean estimate of 0.06 at both time 1 and time 2.

Since our initial hypothesis that absolute metacomprehension accuracy would improve between judgments made during initial reading and judgments made during the study phase across study conditions was not supported, it was not possible to then compare that improvement between conditions. Instead, to evaluate the effect of the study condition on absolute metacomprehension accuracy, we conducted a Bayesian mixed effects linear regression predicting absolute accuracy

of judgments made during the study phase (time 2) from study condition, controlling for absolute accuracy at time 1. Other effects in this model included the fixed effect of standardized prior knowledge ratings, random intercepts for each participant, and random intercepts for each text passage topic. Model estimates are presented in Figure 10a; paired comparisons are visualized in Figure 10b. Absolute accuracy at time 1 had a credibly positive relationship with absolute accuracy at time 2 (Estimate = 0.791, 95% CI [0.760, 0.822], $BF_{10} > 1.60 \times 10^4$), but prior knowledge had no relationship with absolute accuracy at time 2 when controlling for the influence of judgments made at time 1 (Estimate = 0.000, 95% CI [-0.007, 0.007], $BF_{10} = 1/261.09$). There was no credible difference in absolute accuracy between the Re-Read condition and the Non-Interactive condition (Estimate = -0.017, 95% CI [-0.041, 0.008], $BF_{10} = 1/44.62$). There was a credible difference in absolute accuracy between the Re-Read condition and the Interactive condition (Estimate = 0.058, 95% CI [0.034, 0.082], $BF_{10} > 1.60 \times 10^4$), as well as between the Non-Interactive condition and the Interactive condition (Estimate = 0.075, 95% CI [0.051, 0.099], $BF_{10} > 1.60 \times 10^4$). Based on model estimates, absolute accuracy was 0.12 at time 2 for the Re-Read condition, 0.14 for the Non-Interactive condition, and 0.06 for the Interactive condition. Therefore, these results suggest that absolute accuracy for judgments made during the study phase was better for participants who reviewed interactive ChatGPT study guides compared to participants who either re-read text passages or reviewed non-interactive ChatGPT study guides.

Relative Accuracy

Two separate Bayesian mixed effects linear regression models were run predicting relative metacomprehension accuracy for Judgments of Understanding and predictions of performance from study condition, judgment time, and their interaction with random intercepts for each participant.[2,3]

Model estimates are presented in Figure 11a; paired comparisons are visualized in Figure 11b. For participants in the Re-Read condition, there was a credible decrease in relative accuracy between time 1 and time 2 for both judgments of understanding (Estimate = -0.094, 95% CI [-0.151, -0.034], $BF_{10} = 1/2.94$) and predictions of performance (Estimate = -0.077, 95% CI [-0.139, -0.017], $BF_{10} = 1/2.22$). However, these differences were not particularly convincing for either type of judgment according to the Bayes Factors, which would be considered non-informative by conventional guidelines (Lee & Wagenmakers, 2014). For participants in the Non-Interactive condition, there was no difference between relative accuracy at time 2 and time 1 for either judgments of understanding (Estimate = -0.025, 95% CI [-0.084, 0.034], $BF_{10} = 1/33.92$) or predictions of performance (Estimate = -0.018, 95% CI [-0.078, 0.042], $BF_{10} = 1/40.21$). For participants in the Interactive condition, there was also no difference between relative accuracy at time 2 and time 1 for either judgments of understanding (Estimate = -0.002, 95% CI [-0.062, 0.056], $BF_{10} = 1/33.03$) or predictions of performance (Estimate = 0.012, 95% CI [-0.048, 0.073], $BF_{10} = 1/30.43$). These results indicate that relative accuracy did not improve between the reading phase and the study phase in any condition, and may have worsened to some extent for participants who re-read text passages.

As was the case with absolute accuracy, since our initial hypothesis that relative accuracy would improve between judgments made during initial reading and judgments made during the study phase across study conditions was not supported, it was not possible to then compare that improvement between conditions. Instead, to evaluate the effect of the study condition on relative metacomprehension accuracy, we ran two separate Bayesian linear regression models predicting relative accuracy of judgments of understanding and predictions of performance made during the

study phase (time 2) from study condition, controlling for relative accuracy at time 1. Model estimates are presented in Figure 12a, while post-hoc comparisons are visualized in Figure 12b. There was no credible difference in relative accuracy between the Re-Read condition and the Non-Interactive condition for either judgments of understanding (Estimate = -0.043, 95% CI [-0.114, 0.029], $BF_{10} = 1/19.79$) or predictions of performance (Estimate = -0.033, 95% CI [-0.107, 0.043], $BF_{10} = 1/26.42$). Relative accuracy was credibly worse in the Re-Read condition than in the Interactive condition for both judgments of understanding (Estimate = -0.116, 95% CI [-0.187, -0.044], $BF_{10} = 5.56$) and predictions of performance (Estimate = -0.114, 95% CI [-0.188, -0.038], $BF_{10} = 2.04$). Relative accuracy was also credibly worse in the Non-Interactive condition than in the Interactive condition for both judgments of understanding (Estimate = -0.073, 95% CI [-0.145, -0.002], $BF_{10} = 1/3.64$) and predictions of performance (Estimate = -0.081, 95% CI [-0.157, -0.006], $BF_{10} = 1/2.93$). However, the difference between the Re-Read and Interactive conditions for predictions of performance was not particularly convincing according to the Bayes Factor, which would be considered non-informative by conventional guidelines (Lee & Wagenmakers, 2014). Similarly, the difference between the Non-Interactive and Interactive conditions was not particularly convincing, with Bayes Factors indicating that the data were 3.64 times more likely under the null hypothesis of no difference for judgments of understanding and 2.93 times more likely under the null hypothesis of no difference for predictions of performance. These results suggest that relative accuracy during the study phase was somewhat better for participants who reviewed interactive ChatGPT study guides compared to participants who re-read text passages or reviewed non-interactive ChatGPT study guides.

Trust in Artificial Intelligence

Our final set of hypotheses sought to evaluate whether trust in AI was related to test performance and metacomprehension accuracy for participants who reviewed interactive or non-interactive ChatGPT study guides. Responses to each of the items on the TXAI scale (Hoffman et al., 2023; Perrig et al., 2023) were numerically re-coded from 1 ("I disagree strongly") to 5 ("I agree strongly"). Numeric responses were then summed across the seven items for each participant, resulting in a composite AI trust score which could range from 5 (indicating a complete lack of trust in AI) to 35 (indicating a strong sense of trust in AI). The average AI trust score across participants was 20.05 with a standard deviation of 4.66. A Bayesian linear regression was run to predict AI trust score from study condition. This model was then compared with an intercept-only equivalent to evaluate whether trust differed by study condition. The intercept-only (null) model was favored over the model including study condition ($BF_{10} = 1/259.07$), indicating that AI trust scores did not vary by condition.

Four separate Bayesian mixed-effects regression models were run predicting overall test performance, absolute metacomprehension accuracy, and relative metacomprehension accuracy of judgments of understanding and predictions of performance (at time 2) from study condition, AI trust score, and their interaction. Other fixed and random effects followed the same structure as the primary analyses for each of the above outcomes, respectively. Post-hoc tests evaluating the slope of AI trust for each study condition and outcome measure are presented in Table 10. There were only two cases in which the slope for AI trust was credibly different from zero (and in fact was negative): when predicting overall test performance for the Interactive condition (Estimate

= -0.024, 95% CI [-0.047, -0.000], $BF_{10} = 1/12.02$) and when predicting relative accuracy of judgments of understanding for the Non-Interactive condition (Estimate = -0.024, 95% CI [-0.044, -0.004], $BF_{10} = 1/9.45$). However, these results were not particularly convincing according to the Bayes Factors, which indicated that the data were 12 times more likely in the case of overall test performance and 9 times more likely in the case of relative accuracy to be observed under the null hypothesis of no relationship with AI trust. All other slopes for AI trust were not credibly different from zero in either direction, with Bayes Factors strongly favoring the null. Therefore, these results suggest that level of trust in AI had no relationship with overall test performance, absolute accuracy, or relative accuracy for participants in any study condition.

Usage of Artificial Intelligence

Though not part of the primary hypotheses we sought out to test, we also explored participants' self-reported usage of AI for academic purposes to contribute to the growing literature on the popularity of AI among college students. Survey responses by study condition are presented in Table 11. Three Bayesian regression models were run to predict participants' responses for each of the AI usage self-report items (whether or not they had ever used it, the frequency with which they use it, and the average hours per week they use it) from study condition. These models were then compared with intercept-only equivalents to evaluate whether responses varied by condition. The intercept-only (null) model was favored for all items: whether participants had ever used AI for academic purposes ($BF_{10} = 1/12.11$), the frequency with which participants used AI for academic purposes ($BF_{10} = 1/13.44$), and the average hours per week participants used AI for academic purposes ($BF_{10} = 1/1.30 \times 10^3$), indicating that there was no evidence for a difference in self-reported AI usage by study condition. Therefore, we highlight notable results collapsing

across study conditions below.

Of the 273 participants included in the final sample, 255 (93.4%) reported having used AI chatbots such as ChatGPT for academic purposes. Participants most frequently reported using AI for academic purposes on a weekly basis (39%), followed by using AI for academic purposes rarely (23%), with an average usage of 2.41 hours per week (SD = 2.92). As for which AI chatbot models students most frequently use, ChatGPT took first place with 239 out of 273 participants (87.5%) reporting using it for academic purposes, followed by Gemini with 53 out of 273 participants (19.4%). It is important to note that, as with all self-reported measures, there may be some inconsistencies in participant responses. Specifically, three participants who reported not having used AI for academic purposes subsequently reported using AI for academic purposes "rarely" as opposed to "never", and two of those participants reported using AI for academic purposes for an average of 1.8 and 4 hours per week, respectively, as well as reporting specific models used (including ChatGPT, Perplexity, and Gemini). Nonetheless, these results suggest that usage of AI for academic purposes is rampant among students in our sample.

Study Condition Perceptions

Finally, we explored participants' attitudes towards the study conditions they were assigned to (re-reading text passages, reviewing non-interactive ChatGPT study guides, or reviewing interactive ChatGPT study guides) with a self-report scale. Because responses were provided on Likert scales (from "Strongly disagree" to "Strongly agree" and "Very detrimental" to "Very helpful"), Bayesian ordinal regression models were used to predict response distributions for each item from study condition. These models were then compared with intercept-only equivalents to evaluate whether responses varied by condition, followed by post-hoc comparisons between conditions

when appropriate.

In the first part of the survey, participants were asked to rate their level of agreement with a series of statements regarding their enjoyment of their study condition, their engagement with their study condition, and their willingness to use their study condition as a strategy in their own academic work. Participants' responses to the first part of the survey by study condition are presented in Table 12. For enjoyment, the model including study condition was favored over the intercept-only model ($BF_{10} = 1.02x10^{17}$), indicating that there was evidence for a difference between study conditions. Post-hoc comparisons revealed that participants reported credibly lower enjoyment for re-reading text passages than for reviewing non-interactive ChatGPT study guides (Estimate = -2.604, 95% CI [-3.218, -2.020], $BF_{10} > 1.60 \times 10^4$) or interactive ChatGPT study guides (Estimate = -1.678, 95% CI [-2.209, -1.145], $BF_{10} > 1.60 \times 10^4$). Furthermore, participants reported credibly higher enjoyment for reviewing non-interactive ChatGPT study guides than for reviewing interactive ChatGPT study guides (Estimate = 0.926, 95% CI [0.395, 1.473], $BF_{10} = 100$). For engagement, the model including study condition was favored over the intercept-only model ($BF_{10} = 2.64 \times 10^{17}$), indicating that there was evidence for a difference between study conditions. Post-hoc comparisons further revealed that participants reported credibly lower engagement with re-read text passages than with reviewing non-interactive ChatGPT study guides (Estimate = -2.252, 95% CI [-2.842, -1.690], $BF_{10} > 1.60 \times 10^4$) or interactive ChatGPT study guides (Estimate = -2.300, 95% CI [-2.863, -1.743], $BF_{10} > 1.60 \times 10^4$). There was no credible difference between participants who reviewed non-interactive compared to interactive ChatGPT study guides (Estimate = -0.048, 95% CI [-0.567, 0.467], $BF_{10} = 1/3.70$). Finally, for willingness to use the study strategy, the model including study condition was not favored over the intercept-only model ($BF_{10} = 1/1.85$),

indicating that there was no evidence for a difference between study conditions. These results suggest that participants enjoyed and were engaged with reviewing ChatGPT study guides more than re-reading text passages, but were not more likely to select them as a study strategy in their own academic work.

In the second part of the survey, participants were asked to rate how helpful or detrimental their study condition was for their ability to understand the text passages, accurately answer the test questions, remember details from the text passages, and make connections between concepts within the text passages. Participants' responses to the second part of the survey by study condition are presented in Table 13. The models including study condition were not favored over intercept-only models for any of the items (understand text passages: $BF_{10} = 1.93$; accurately answer questions: $BF_{10} = 1/14.29$; remember details: $BF_{10} = 1/8.33$; make connections between concepts: $BF_{10} = 1/4.35$), indicating that there was no evidence for differences between conditions. These results suggest that participants perceived both types of ChatGPT study guides to be equally effective as re-reading text passages for learning purposes, which aligns with the findings for actual test performance.

Chapter 4: Discussion

The present study investigated whether AI-generated study materials could be used to facilitate memory, comprehension, and metacomprehension of text materials. While reviewing ChatGPT study guides did not provide a benefit to test performance above and beyond re-reading text passages, participants reported more enjoyment and engagement with the study guides than with re-reading. Furthermore, interactive ChatGPT study guides reduced overconfidence in metacomprehension judgments, resulting in better metacomprehension accuracy than non-interactive ChatGPT study guides or re-reading text passages. We also found that participants reported enjoying and feeling engaged with reviewing ChatGPT study guides to a greater extent than simply re-reading text passages. Though ChatGPT study guides did not provide improve test performance over re-reading, they also did not worsen test performance. Therefore, the results of the present study suggest that participants who enjoy ChatGPT study guides more than the "traditional" study method of re-reading can do so with an equal level of effectiveness for test performance, and potentially additional benefits for metacomprehension.

In contrast to prior literature on outlines, we failed to find an adjunct display effect for AI-generated study guides. This finding is not particularly surprising with regard to comprehension test performance; the effect of instructor-provided outlines is generally weaker for comprehension-based evaluations than for memory-based evaluations because they do not prime integration of information as strongly as the act of generating an outline (Ponce et al., 2023). With regard to memory test performance, however, there are several possible explanations for this finding. Some studies have presented outlines to students as a preface for the to-be-learned material (e.g., Krug

et al., 1989) or at the same time as their first and only opportunity to study the material (e.g., Kiewra et al., 1999). In the present study, the adjunct displays (non-interactive and interactive ChatGPT study guides) were only presented the second time participants were exposed to the material in comparison with a third group who was also exposed to the material twice. Therefore, it is possible that the study guides simply provided the same amount of benefit, if any, as re-reading text passages. However, work by Marefat and Ghahari (2009) and Moradi and colleagues (2020) has shown an adjunct display effect when outlines are provided to students following initial reading in comparison with students who are provided the same texts for a second time. Importantly, though, both of the aforementioned studies employed the exact same text materials, which differed from those used in the present study. Differences in text passage length, difficulty, and questions used may account for the discrepancy in the findings. Another key difference is that previous studies on the adjunct display effect have employed instructor- or expert-generated outlines, as opposed to the AI-generated study guides employed here. The current capabilities of AI simply may not be sufficient to create study materials which match the level of an expert or instructor. In addition, instructor- or expert-generated outlines likely contain all of the information required to answer all test questions used. Our AI-generated study guides, however, were created "blind" to all test questions and were further modified to ensure that only half of the memory-based questions could be answered solely from the study guides. This was done to reflect the fact that students rarely know the exact contents of an exam when generating study materials. It is therefore possible that if AI-generated study guides contained all test-relevant information, they may show a greater benefit to memory-based test performance than re-reading alone. This explanation is not strongly supported by the present findings, though, as there was no difference in memory test performance between participants who

reviewed either type of ChatGPT study guide and participants who re-read text passages regardless of whether or not test question answers were included in the study guide.

Though interactive ChatGPT study guides did not provide a general improvement in test performance over re-reading text passages or reviewing non-interactive ChatGPT study guides, we did find some evidence in support of a testing effect (Roediger & Karpicke, 2006). That is, participants who reviewed interactive ChatGPT study guides performed better on memory-based questions whose answers were included in interactive components. This is consistent with Hinze and Wiley's (2011) findings on fill-in-the-blank recall practice, which also only benefited test performance for items matching the fill-in-the-blanks exactly. It is possible that if there were more interactive components matching memory-based test questions, there would have been a greater benefit in performance overall for participants who reviewed ChatGPT study guides. In fact, interactive ChatGPT study guides originally contained far more interactive components overall, including those which aligned with answers to memory-based test questions, but these interactive components had to be manually reduced in the present study to ensure that participants would have enough time to complete most of the study guides. Furthermore, we found that participants who completed more interactive components of ChatGPT study guides correctly displayed better overall, memory, and comprehension test performance. While completing more interactive components correctly may have evidenced better learning of text passages in the first place, which would also correspond to better test performance, it is also possible that participants who completed interactive components correctly did so by referring back to the original texts to find information, which would provide a benefit for test performance later. On average, participants were not able to complete all 20 interactive components of ChatGPT study guides within the 5 minutes allotted, which

likely disproportionately affected participants who had a slower reading speed or response time. Therefore, test performance for participants who reviewed interactive ChatGPT study guides may have further improved if participants had unlimited time to complete those interactive components, as they would (to some extent) if studying on their own.

As for metacomprehension, we failed to find a rereading effect in the present study. Contrary to our hypotheses, participants' metacomprehension accuracy was worse for judgments made after re-reading text passages (or reviewing non-interactive ChatGPT study guides) compared to judgments made after initial reading. Several methodological and analytical factors may contribute to this finding. First, participants in the present study were shown each text passage in its entirety for a set amount of time, in contrast to studies in which participants either read text passages one sentence at a time or in their entirety but can advance when finished (e.g., Dunlosky, 2005; Rawson & Dunlosky, 2002; Rawson et al., 2000). Therefore, it is possible for participants in the present study to have read some or all of a text passage multiple times during "initial reading", depending on individual factors such as reading speed and motivation. The ability to re-read elements of text passages may have allowed participants to construct a sufficient situation model during initial reading, providing no benefit when reading again. Additionally, participants in the present study were not aware that they would be shown text passages a second time (either on their own or in conjunction with a ChatGPT study guide) in advance. This may have encouraged participants to read more carefully initially, similarly allowing them to construct a better understanding of text passages during initial reading, than if they knew that text passages would be presented for a second time prior to testing. Second, studies examining the rereading effect typically do so between-subjects, where one group of participants read a text passage once, then provide metacomprehension judgments,

and another group of participants read a text passage twice, then provide metacomprehension judgments (e.g., Dunlosky, 2005; Griffin et al., 2008; Rawson & Dunlosky, 2002; Rawson et al., 2000). Instead, participants in the present study each made the same metacomprehension judgments at two different timepoints, allowing for a within-subjects evaluation of the effect of re-reading. It is possible that the act of making metacomprehension judgments following initial reading of text passages created an anchor based on which participants adjusted their second round of metacomprehension judgments. In one of very few studies which has examined the rereading effect within-subjects, Margolin and Snyder (2018) had the same participants read a series of text passages only one time and a separate series of text passages twice, and elicited repeated metacomprehension judgments (following initial reading and following re-reading) for text passages read twice. Their analyses revealed that the magnitude of metacomprehension judgments (i.e., participants' confidence in their understanding of the material) significantly increased between initial reading and re-reading, as was the case with predictions of performance in the present study. However, Margolin and Snyder (2018) did not compute a measure of relative or absolute metacomprehension accuracy to compare these two timepoints within text passages, so it is unclear whether the increase in judgment magnitude made participants more or less metacognitively accurate. In the present study, increases in confidence for predictions of performance moved participants' judgments further away from their actual test performance, making metacognitive accuracy worse after re-reading. This general pattern of results supports Margolin and Snyder's (2018) idea that reading a text passage twice makes students think they understand it better, even when this is not necessarily the case.

Even when controlling for metacomprehension accuracy at time 1 (during initial reading), participants who reviewed interactive ChatGPT study guides had better metacomprehension ac-

curacy at time 2 (during the study phase) compared to participants who re-read text passages or reviewed non-interactive ChatGPT study guides. This provides some evidence that retrieval practice (in this case, the act of filling in interactive components with the correct information) can support metacomprehension in addition to memory-based test performance (e.g., Dunlosky et al., 2005). However, effects were weaker for relative as opposed to absolute accuracy, with some evidence in support of no difference between interactive and non-interactive ChatGPT study guides for relative accuracy. Similarly, while there were sharp decreases in absolute metacomprehension accuracy between initial reading and the study phase for participants who re-read text passages or reviewed non-interactive ChatGPT study guides, these effects were not as dramatic for relative metacomprehension accuracy. In fact, for participants who reviewed non-interactive ChatGPT study guides, there was no change in relative accuracy between initial reading and the study phase for either judgments of understanding or predictions of performance. Therefore, it seems that although reviewing non-interactive ChatGPT study guides increased overconfidence in metacomprehension judgments, these increases were consistent across text passages, resulting in no difference for relative accuracy. For participants who re-read text passages, there were credible but weak decreases in relative accuracy between initial reading and re-reading. Differences in metacomprehension accuracy between study conditions at time 2 (during the study phase) were also weaker for relative than absolute accuracy. One possible reason for this pattern of results is that relative and absolute accuracy provide different perspectives on metacomprehension abilities, and likely serve different purposes as far as guiding students' study decisions. Good relative accuracy can help students determine which test materials they should prioritize when studying, while good absolute accuracy can help students determine whether or not they need to study specific test materials. However, the

limitations of relative accuracy as a measure of metacomprehension may also contribute to these findings. Judgments of understanding are typically given on a Likert scale (such as "Not well at all" to "Very well" for the present study), they are generally not considered to be continuous in nature or to have equal distances between scale points (such as the distance between "Somewhat well" and "Very well"). Therefore, researchers use measures of relative accuracy such as Kendall's tau or Goodman-Kruskal's gamma, which are non-parametric correlations based on the rank-order of both metacomprehension judgments and test performance across text passages within individuals. These measures ignore the magnitude of differences between rankings, which limits the amount of information they can convey. In addition, because relative accuracy is computed as a single score for each participant, one loses the ability to employ a mixed-effects modeling approach, which can account for text passage difficulty, self-rated prior knowledge, and participant abilities.

Future work should confirm the results obtained here to ensure their reliability across samples. To this extent, an exact replication of the present study is currently in progress. More research is needed to continue to assess the effectiveness of AI-generated study materials for text materials as the capabilities of AI advance. In particular, researchers should explore the implementation of AI in real educational contexts that rely on primarily text-based materials such as textbook chapters or journal articles, moving beyond the literature's focus on skills- and project-based areas. Furthermore, as the present study only provided students with already-generated AI study materials, researchers should explore whether interacting with AI chatbots directly differently impacts the learning and studying of text materials. Future work should also study the rereading effect for metacomprehension accuracy within-subjects to further disentangle potential theoretical reasons why it may not present in the same way as in between-subjects designs. Lastly, future

research should examine whether the improvements in metacomprehension accuracy gained by reviewing interactive ChatGPT study guides are meaningful in helping students better regulate their own studying and learning.

Footnotes

1. Since absolute accuracy scores were continuous but bounded between -1 and 1, we also conducted analyses using Beta regression models by transforming the data to a (0, 1) interval. The results from both modelling techniques converged, and thus only the results for the linear regression models on raw absolute accuracy scores are reported here for simplicity's sake. Models and post-hoc comparisons for Beta regression models can be found on OSF.

2. Since our measure of relative accuracy, Kendall's tau-a, is continuous but bounded between -1 and 1, we also conducted analyses using Beta regression models by transforming the data to a (0, 1) interval and linear regression models by transforming the data to Fisher's z through Pearson's r. The results from all three modelling techniques converged, and thus only the results for the linear regression models on raw Kendall's tau-a values are presented for simplicity's sake. Models and post-hoc comparisons for Beta regression models and transformed linear regression models can be found on OSF.

3. Because relative accuracy was computed across text passages for each participant, fixed effects for prior knowledge and random effects for text passage topic could not be included in this model.

Tables

Table 1

Prior Knowledge Ratings by Study Condition and Text Passage Topic

Topic	Re-Read		Non-Interactive		Interactive	
	M	*SD*	*M*	*SD*	*M*	*SD*
Volcanoes	2.32	0.86	2.44	0.87	2.31	0.87
Evolution	3.05	0.80	3.02	0.88	3.11	0.89
Ice Ages	2.27	0.90	2.23	0.90	2.22	0.83
Antibiotics	2.25	1.03	2.33	0.97	2.24	1.00
Intelligence Tests	3.30	0.95	3.13	1.04	3.27	0.98
Monetary Policy	2.16	1.19	2.24	1.12	2.14	0.98

Note. Prior knowledge was rated on a 1–5 scale for each text passage.

Table 2

Overall Test Performance by Study Condition and Text Passage Topic

Topic	Re-Read		Non-Interactive		Interactive	
	M	*SD*	*M*	*SD*	*M*	*SD*
Volcanoes	5.22	2.45	5.86	2.50	5.53	2.67
Evolution	8.33	2.38	7.53	2.90	8.17	2.98
Ice Ages	7.24	2.41	6.98	2.38	7.41	2.52
Antibiotics	5.63	2.15	5.10	2.03	5.00	2.23
Intelligence Tests	8.47	2.12	8.01	1.97	8.77	2.09
Monetary Policy	7.50	2.81	6.71	2.85	7.20	2.69

Note. Overall test performance was measured as the number of test questions answered correctly out of a possible 13 for each text passage.

Table 3

Comprehension Test Performance by Study Condition and Text Passage Topic

Topic	Re-Read		Non-Interactive		Interactive	
	M	*SD*	*M*	*SD*	*M*	*SD*
Volcanoes	2.09	1.22	2.58	1.33	2.38	1.39
Evolution	3.32	1.26	2.91	1.31	3.24	1.30
Ice Ages	2.99	1.19	2.81	1.22	3.21	1.24
Antibiotics	1.71	1.21	1.37	1.12	1.50	1.12
Intelligence Tests	3.57	1.14	3.49	1.00	3.83	1.19
Monetary Policy	2.63	1.40	2.41	1.39	2.59	1.38

Note. Comprehension test performance was measured as the number of comprehension-based test questions answered correctly out of a possible 5 for each text passage.

Table 4

Memory Test Performance by Study Condition and Text Passage Topic

Topic	Re-Read		Non-Interactive		Interactive	
	M	*SD*	*M*	*SD*	*M*	*SD*
Volcanoes	3.13	1.58	3.27	1.59	3.16	1.73
Evolution	5.01	1.65	4.62	1.99	4.92	1.98
Ice Ages	4.25	1.69	4.16	1.66	4.20	1.66
Antibiotics	3.92	1.40	3.73	1.31	3.50	1.55
Intelligence Tests	4.90	1.33	4.52	1.29	4.93	1.23
Monetary Policy	4.87	1.76	4.31	1.90	4.61	1.72

Note. Memory test performance was measured as the number of memory-based test questions answered correctly out of a possible 8 for each text passage.

Table 5

Number of Interactive Components Completed, Correct, and Incorrect by Text Passage Topic

Topic	Number Completed		Number Correct		Number Incorrect	
	M	*SD*	*M*	*SD*	*M*	*SD*
Volcanoes	15.14	5.20	11.33	4.81	3.81	3.31
Evolution	18.96	2.80	16.84	3.58	2.11	2.70
Ice Ages	16.62	3.99	14.33	3.88	2.29	2.46
Antibiotics	14.84	4.61	10.89	3.95	3.96	4.10
Intelligence Tests	18.49	3.27	17.27	3.62	1.22	1.78
Monetary Policy	16.71	4.13	13.62	3.95	3.09	3.23

Note. Specific to the 90 participants in the Interactive study condition. Number completed, correct, and incorrect are out of a possible 20 interactive components per study guide.

Table 6

Judgments of Understanding by Study Condition and Text Passage Topic

Topic	Re-Read				Non-Interactive				Interactive			
	Time 1		Time 2		Time 1		Time 2		Time 1		Time 2	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Volcanoes	4.32	1.37	5.13	1.34	4.08	1.20	4.92	1.24	4.32	1.18	4.19	1.43
Evolution	5.09	1.33	5.26	1.37	4.96	1.39	5.19	1.23	5.18	1.30	5.34	1.20
Ice Ages	4.33	1.49	4.89	1.38	4.35	1.25	4.92	1.30	4.51	1.26	4.70	1.25
Antibiotics	3.90	1.25	4.55	1.26	4.03	1.37	4.66	1.27	3.86	1.31	3.88	1.39
Intelligence Tests	5.63	1.31	5.90	1.09	5.35	1.33	5.77	1.14	5.80	1.11	5.73	1.22
Monetary Policy	4.08	1.45	4.64	1.40	4.20	1.19	4.93	1.28	3.99	1.40	4.19	1.37

Note. Judgments of Understanding were made on a 1–5 scale for each text passage at each time point.

Table 7

Predictions of Performance by Study Condition and Text Passage Topic

Topic	Re-Read				Non-Interactive				Interactive			
	Time 1		Time 2		Time 1		Time 2		Time 1		Time 2	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Volcanoes	57.24	21.12	66.33	20.58	54.97	22.42	65.23	21.47	55.58	19.62	53.77	23.05
Evolution	65.59	20.80	68.46	20.13	64.57	23.38	68.80	20.75	66.13	21.20	68.31	21.41
Ice Ages	57.58	22.71	64.38	21.04	57.22	21.87	64.42	22.14	57.97	19.68	59.81	21.18
Antibiotics	51.01	21.40	58.10	21.70	52.53	21.54	61.10	21.94	50.78	21.07	48.38	22.19
Intelligence Tests	71.99	20.03	77.29	17.76	70.35	22.38	75.84	18.10	75.33	16.61	74.26	19.96
Monetary Policy	54.92	22.30	60.17	21.60	56.82	21.72	64.53	21.60	54.31	22.54	56.18	22.79

Note. Predictions of Performance were made on a 0–100 scale for each text passage at each time point.

Table 8

Absolute Metacomprehension Accuracy by Study Condition and Text Passage Topic

Topic	Re-Read				Non-Interactive				Interactive			
	Time 1		Time 2		Time 1		Time 2		Time 1		Time 2	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Volcanoes	0.17	0.23	0.26	0.25	0.10	0.25	0.20	0.23	0.13	0.26	0.11	0.26
Evolution	0.02	0.23	0.04	0.24	0.07	0.24	0.11	0.24	0.03	0.24	0.05	0.25
Ice Ages	0.02	0.26	0.09	0.26	0.04	0.23	0.11	0.24	0.01	0.23	0.03	0.23
Antibiotics	0.08	0.25	0.15	0.25	0.13	0.21	0.22	0.22	0.12	0.23	0.10	0.26
Intelligence Tests	0.07	0.20	0.12	0.18	0.09	0.22	0.14	0.20	0.08	0.18	0.07	0.20
Monetary Policy	-0.03	0.24	0.02	0.24	0.05	0.25	0.13	0.26	0.01	0.26	0.01	0.28

Note. Absolute metacomprehension accuracy was computed as the predicted proportion of test questions correct subtracted by the actual proportion of test questions correct, ranging from -1.0 to +1.0.

Table 9

Relative Metacomprehension Accuracy by Study Condition and Judgment Type

Judgment Type	Re-Read				Non-Interactive				Interactive			
	Time 1		Time 2		Time 1		Time 2		Time 1		Time 2	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Judgments of Understanding	0.26	0.29	0.17	0.31	0.21	0.30	0.18	0.24	0.31	0.26	0.31	0.30
Predictions of Performance	0.28	0.31	0.21	0.33	0.22	0.31	0.21	0.28	0.34	0.31	0.35	0.31

Note. Relative metacomprehension accuracy was computed using intraindividual Kendall's tau-a correlations.

Table 10

Post-Hoc Analyses of AI Trust Predicting Overall Test Performance, Absolute Accuracy, and Relative Accuracy by Study Condition

Outcome	Re-Read				Non-Interactive				Interactive			
	Est	*2.5%*	*97.5%*	BF_{10}	*Est*	*2.5%*	*97.5%*	BF_{10}	*Est*	*2.5%*	*97.5%*	BF_{10}
Overall Test Performance	-0.001	-0.028	0.026	1 / 99.18	-0.020	-0.046	0.007	1 / 33.32	-0.024	-0.047	-0.000	1 / 12.02
Absolute Accuracy	-0.001	-0.005	0.003	1 / 631.46	0.001	-0.002	0.005	1 / 604.81	-0.002	-0.005	0.002	1 / 329.40
Relative Accuracy (JoU)	0.011	-0.010	0.031	1 / 78.55	-0.024	-0.044	-0.004	1 / 9.45	-0.006	-0.024	0.012	1 / 85.50
Relative Accuracy (PoP)	0.005	-0.006	0.017	1 / 166.16	-0.008	-0.019	0.003	1 / 99.15	-0.004	-0.014	0.007	1 / 149.52

Note. Est is the Bayesian posterior distribution mean for the slope of Trust in AI. 2.5% and 97.5% are the lower and upper bounds, respectively, of the 95% credible interval. For relative accuracy, JoU is Judgments of Understanding and PoP is Predictions of Performance.

Table 11

Table of Counts for Artificial Intelligence Usage by Study Condition

Condition	*n*	Usage		Frequency				
		Yes	No	Daily	Weekly	Monthly	Rarely	Never
Re-Read	92	84	8	14	38	14	19	7
Non-Interactive	91	87	4	13	35	17	23	3
Interactive	90	84	6	19	33	12	21	5
Total	273	255	18	46	106	43	63	15

Note. Usage: ”Have you ever used Artificial Intelligence for academic purposes?” Frequency: ”How frequently do you use Artificial Intelligence for academic purposes?”

Table 12

Table of Counts for Perceived Enjoyment, Engagement, and Strategy Usage by Study Condition

Item	Condition	*n*	Strongly Disagree	Somewhat Disagree	Neither Agree nor Disagree	Somewhat Agree	Strongly Agree
Item 1	Re-Read	92	17	40	15	20	0
	Non-Interactive	91	1	9	12	45	24
	Interactive	90	3	13	24	40	10
Item 2	Re-Read	92	17	44	15	13	3
	Non-Interactive	91	2	14	10	44	21
	Interactive	90	3	8	11	47	21
Item 3	Re-Read	92	8	20	8	40	16
	Non-Interactive	91	7	10	13	30	31
	Interactive	90	7	14	15	32	22

Note. Item 1: ”I enjoyed [study condition]”; Item 2: ”I was engaged in [study condition]”; Item 3: ”I would choose to use [study condition] in my own academic work”.

Table 13

Table of Counts for Perceived Helpfulness by Study Condition

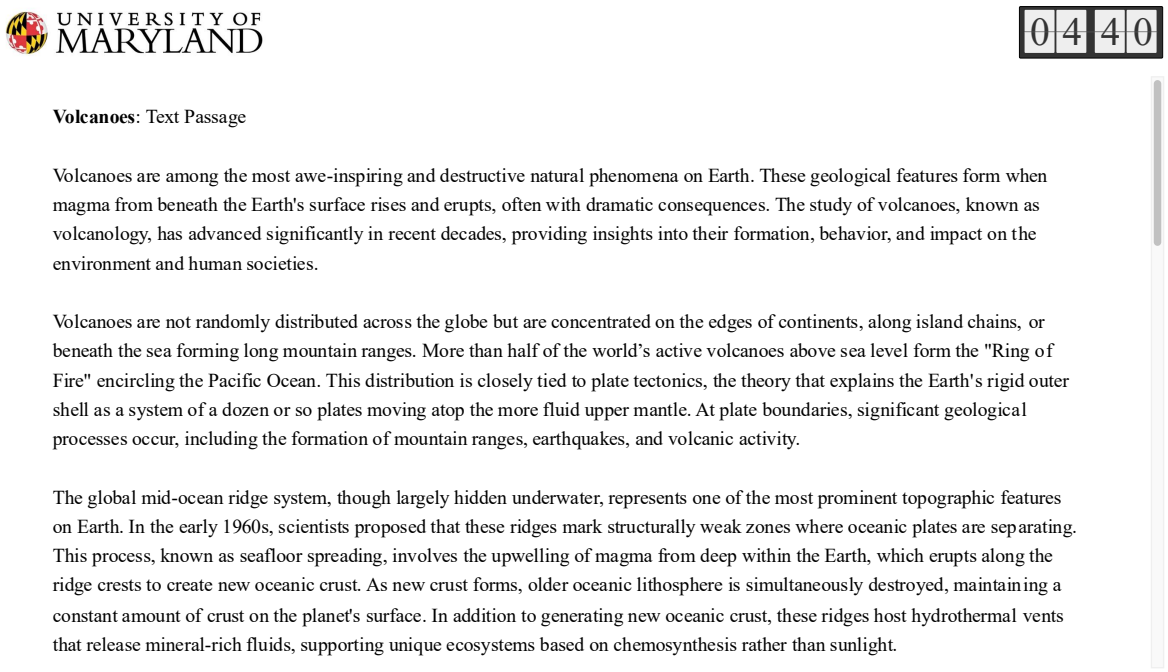
Item	Condition	*n*	Very Detrimental	Somewhat Detrimental	Neither Helpful nor Detrimental	Somewhat Helpful	Very Helpful
Item 1	Re-Read	92	1	11	17	51	12
	Non-Interactive	91	0	2	15	53	21
	Interactive	90	3	5	10	59	13
Item 2	Re-Read	92	3	9	23	49	8
	Non-Interactive	91	2	11	26	45	7
	Interactive	90	3	7	27	46	7
Item 3	Re-Read	92	2	9	20	45	16
	Non-Interactive	91	3	8	15	41	24
	Interactive	90	3	7	18	41	21
Item 4	Re-Read	92	2	7	26	48	9
	Non-Interactive	91	2	7	24	39	19
	Interactive	90	5	6	31	36	12

Note. Item 1: ”Your ability to understand the text passages”; Item 2: ”Your ability to accurately answer the test questions”; Item 3: ”Your ability to remember specific details from the text passages”; Item 4: ”Your ability to make connections between concepts within the text passages”.

Figures

Figure 1

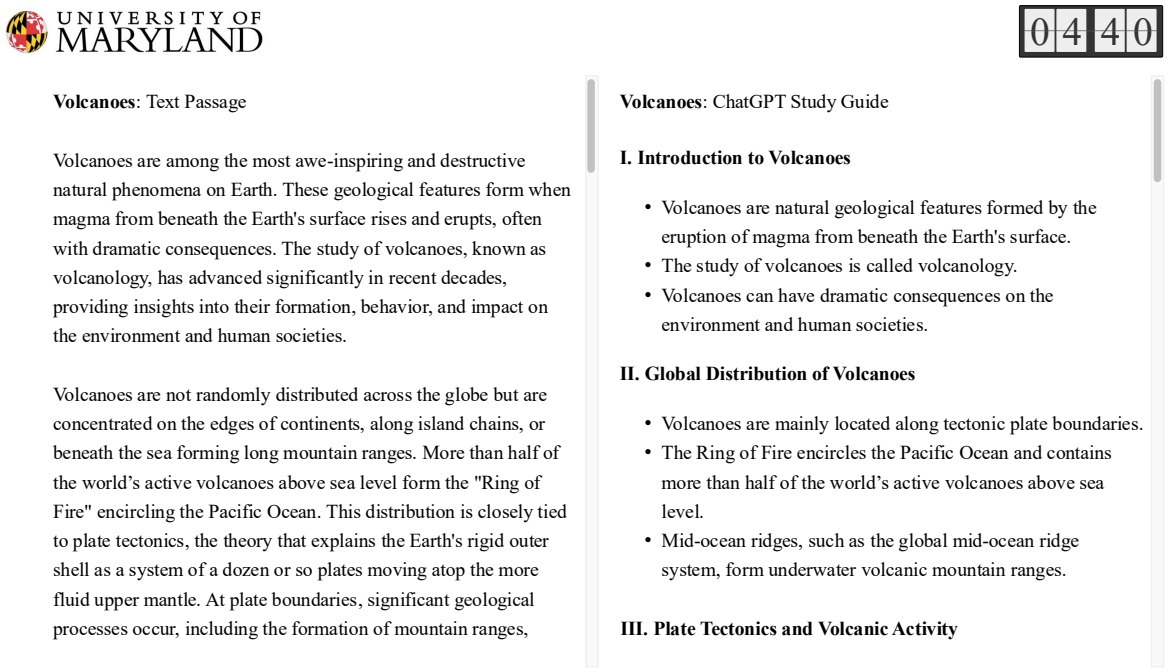
Example of Text Passage Presented During Initial Reading



Note. Excerpt of text passage on Volcanoes. Timer in top-right corner shows 4 minutes, 40 seconds remaining out of 5 minutes. Participants could use the scroll bar to view the remainder of the text passage.

Figure 2

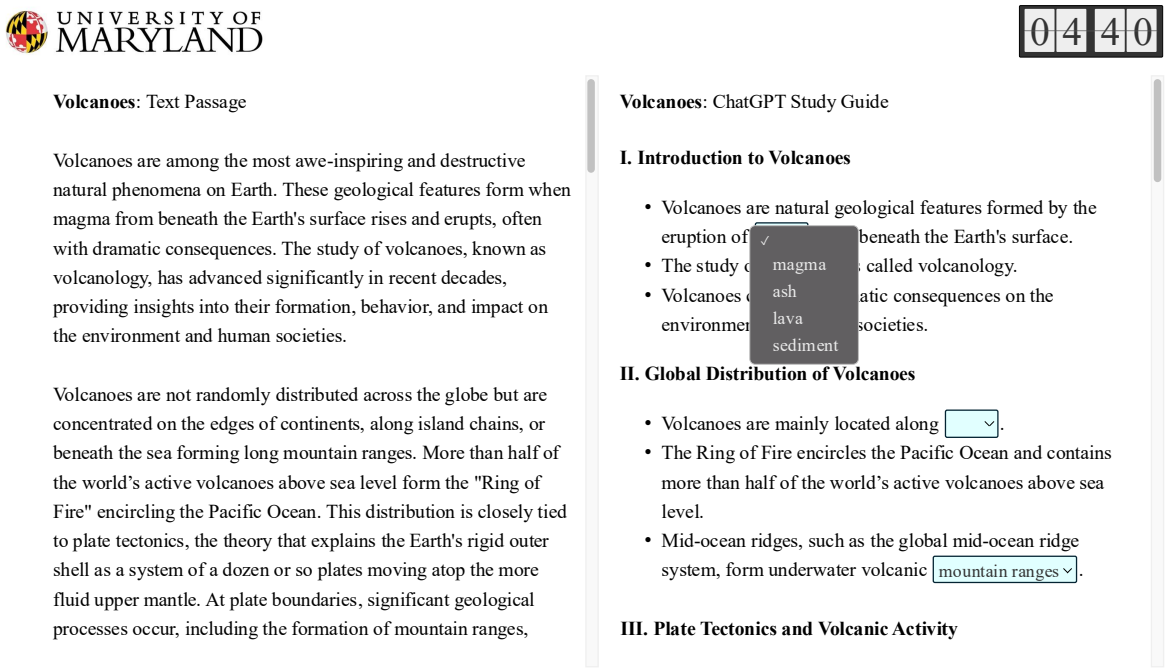
Example of Non-Interactive ChatGPT Study Guide Presented During Study Phase



Note. Excerpts of text passage (left) and non-interactive ChatGPT study guide (right) on Volcanoes. Timer in top-right corner shows 4 minutes, 40 seconds remaining out of 5 minutes. Participants could use the scroll bars to view the remainders of the text passage and study guide independently.

Figure 3

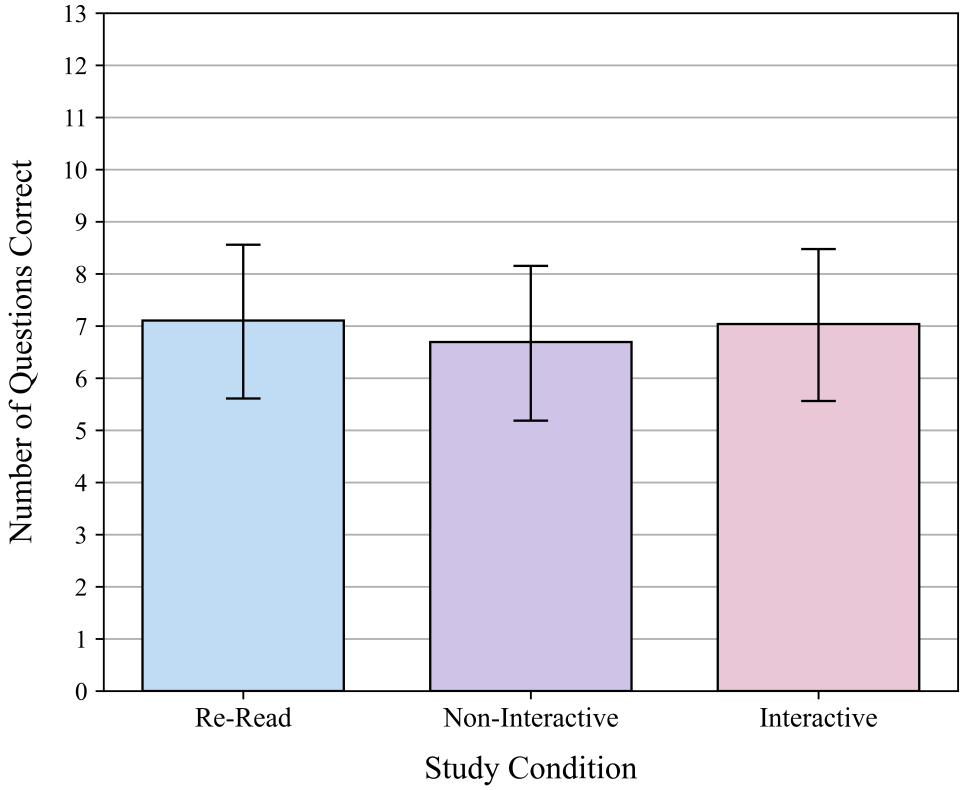
Example of Interactive ChatGPT Study Guide Presented During Study Phase



Note. Excerpts of text passage (left) and interactive ChatGPT study guide (right). Interactive components highlighted in light blue at various stages: first interactive component selected to view drop-down answer options; second interactive component blank; third interactive component with "mountain ranges" selected as answer, embedded in text. Timer in top-right corner shows 4 minutes, 40 seconds remaining out of 5 minutes. Participants could use the scroll bars to view the remainders of the text passage and study guide independently.

Figure 4a

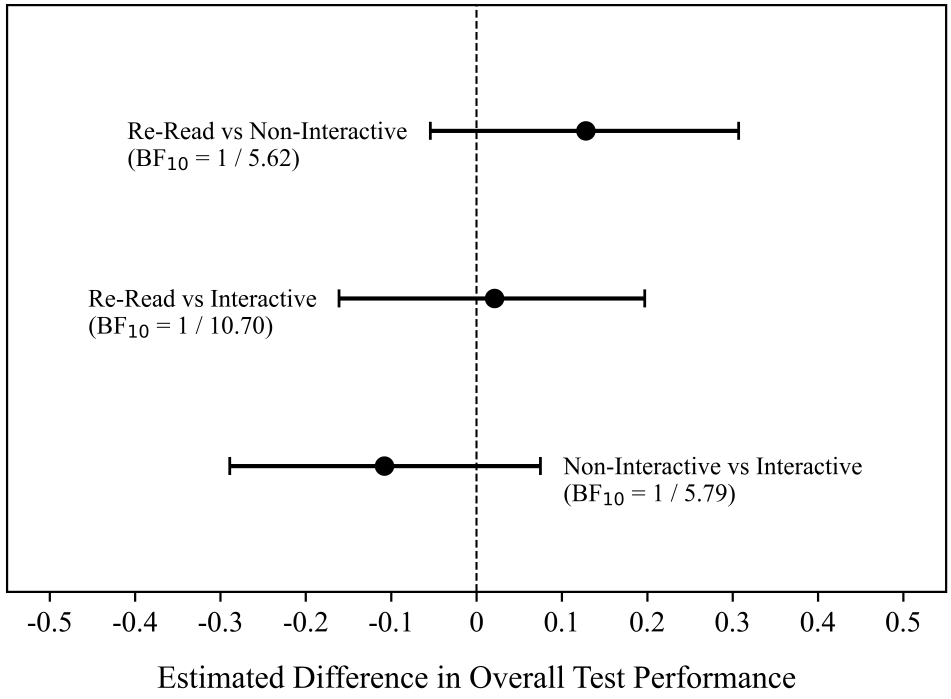
Model Estimates of Overall Test Performance by Study Condition



Note. Bars represent Bayesian posterior distribution means for the number of test questions (including both memory-based and comprehension-based items) answered correctly. Error bars represent 95% credible intervals for Bayesian posterior distributions.

Figure 4b

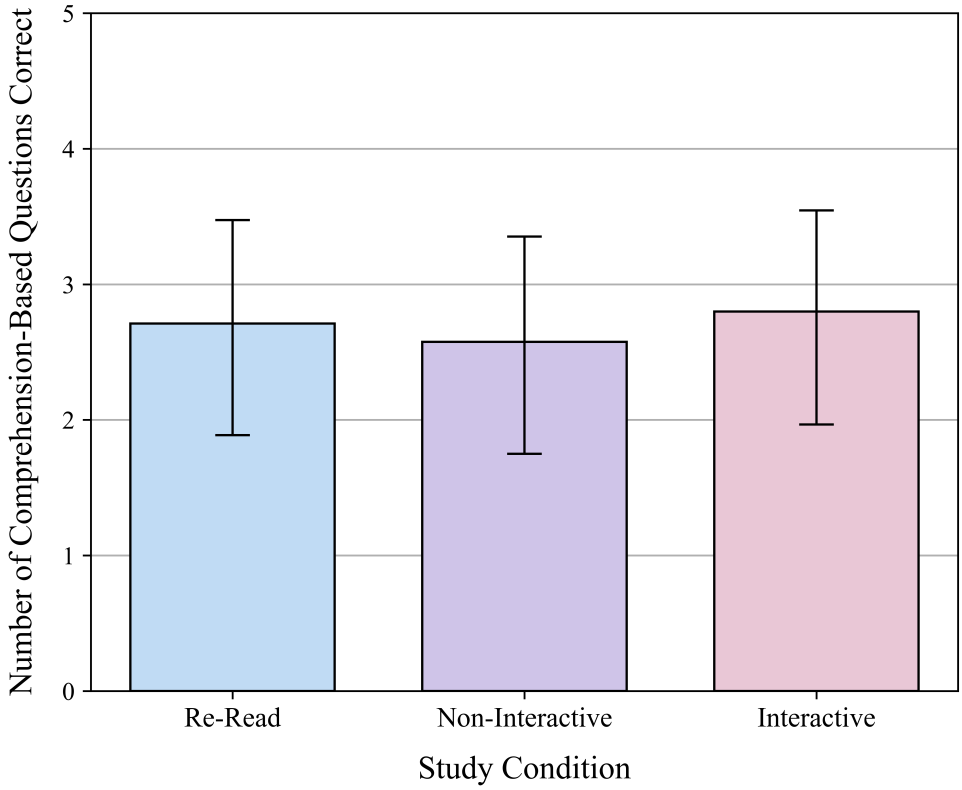
Paired Comparisons of Overall Test Performance Between Study Conditions



Note. Points represent differences in Bayesian posterior distribution means between study conditions for the number of test questions (including both memory-based and comprehension-based items) answered correctly. Error bars represent 95% credible intervals for differences in Bayesian posterior distribution means between study conditions. Bayes Factors computed via the Savage-Dickey density ratio.

Figure 5a

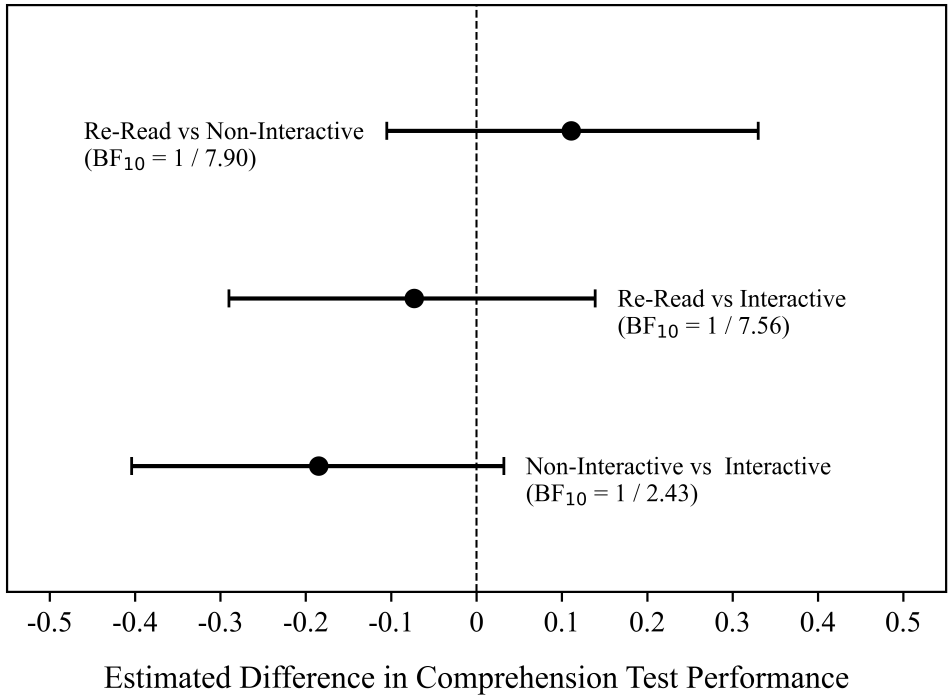
Model Estimates of Comprehension Test Performance by Study Condition



Note. Bars represent Bayesian posterior distribution means for the number of comprehension-based test questions answered correctly. Error bars represent 95% credible intervals for Bayesian posterior distributions.

Figure 5b

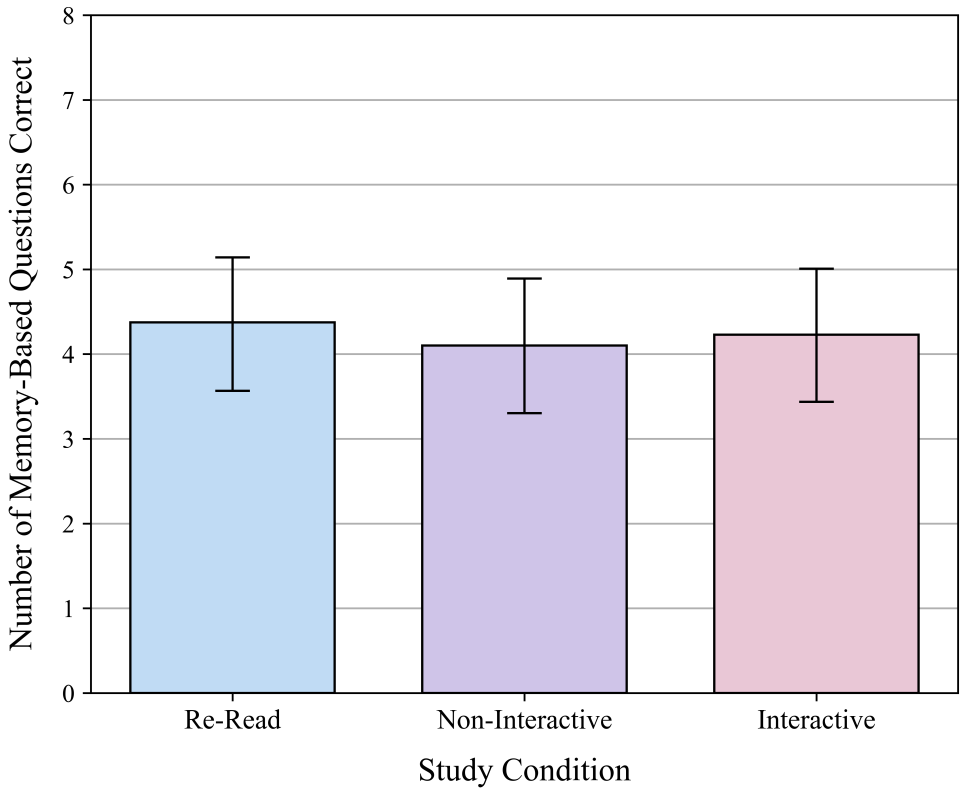
Paired Comparisons of Comprehension Test Performance Between Study Conditions



Note. Points represent differences in Bayesian posterior distribution means between study conditions for the number of comprehension-based test questions answered correctly. Error bars represent 95% credible intervals for differences in Bayesian posterior distribution means between study conditions. Bayes Factors computed via the Savage-Dickey density ratio.

Figure 6a

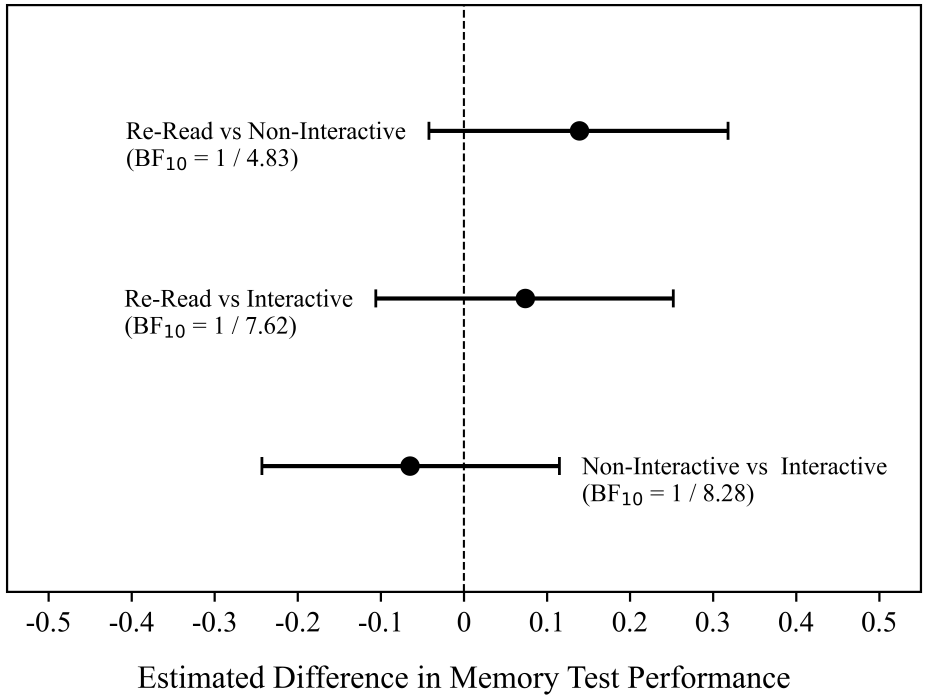
Model Estimates of Memory Test Performance by Study Condition



Note. Bars represent Bayesian posterior distribution means for the number of memory-based test questions answered correctly. Error bars represent 95% credible intervals for Bayesian posterior distributions.

Figure 6b

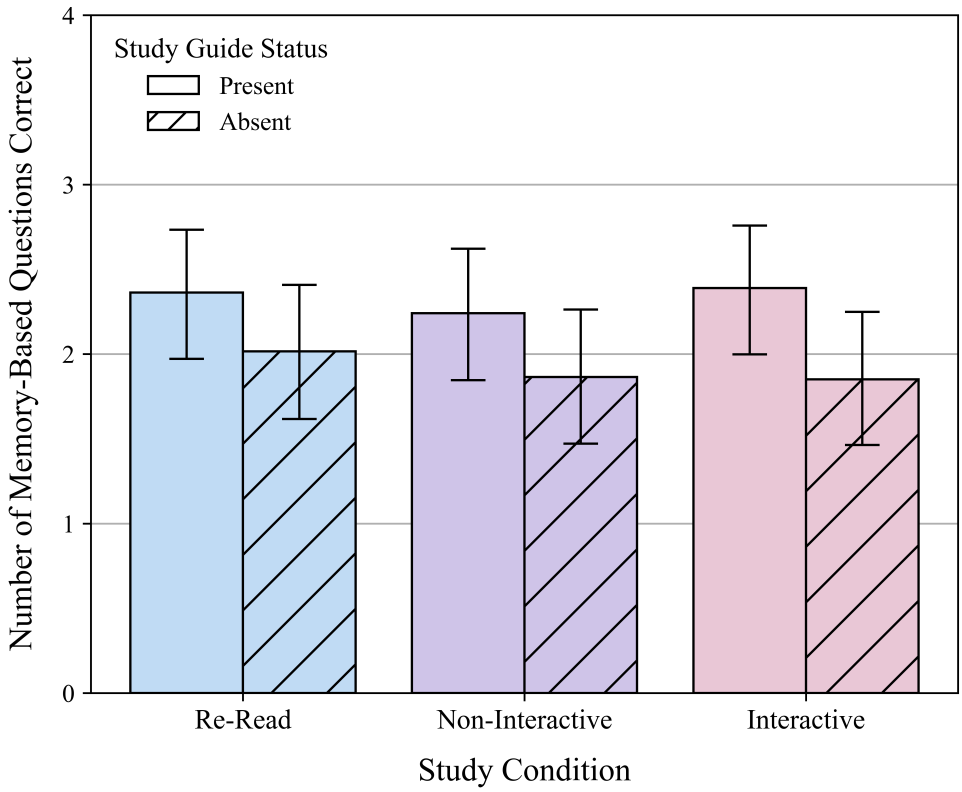
Paired Comparisons of Memory Test Performance Between Study Conditions



Note. Points represent differences in Bayesian posterior distribution means between study conditions for the number of memory-based test questions answered correctly. Error bars represent 95% credible intervals for differences in Bayesian posterior distribution means between study conditions. Bayes Factors computed via the Savage-Dickey density ratio.

Figure 7a

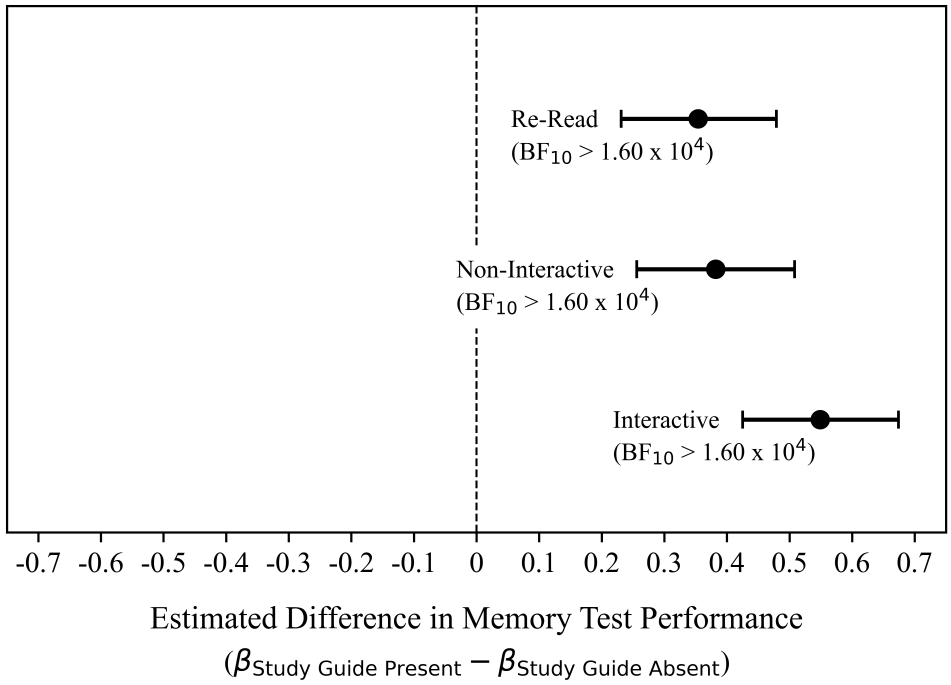
Model Estimates of Memory Test Performance by Study Condition and Study Guide Status



Note. Solid bars represent Bayesian posterior distribution means for the number of questions answered correctly of the four memory-based questions whose answers were included in ChatGPT study guides ("Present"). Striped bars represent Bayesian posterior distribution means for the number of questions answered correctly of the four memory-based questions whose answers were not included in ChatGPT study guides ("Absent"). Error bars represent 95% credible intervals for Bayesian posterior distributions.

Figure 7b

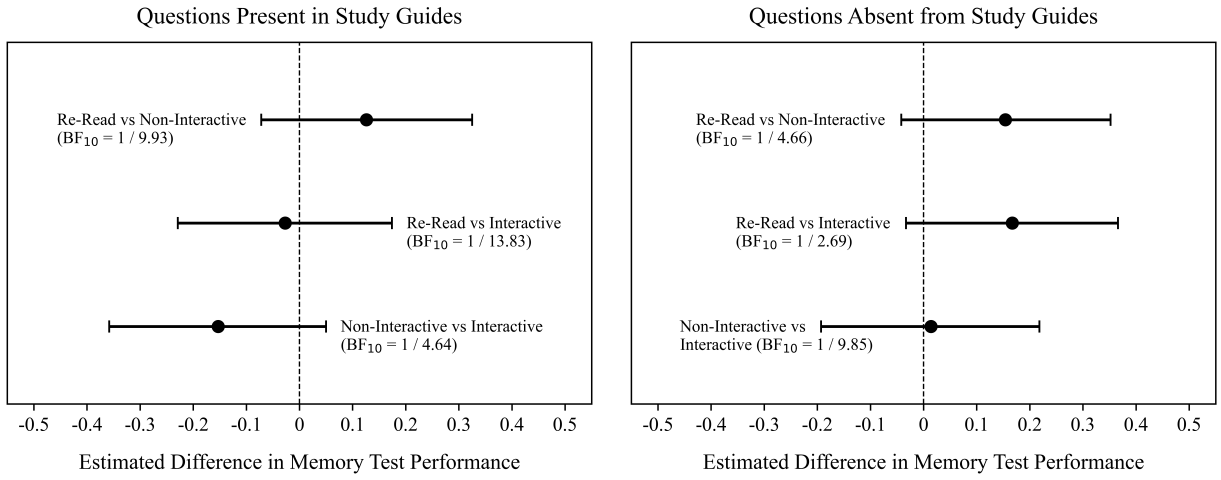
Paired Comparisons of Memory Test Performance Between Questions Present and Absent from Study Guides by Study Condition



Note. Points represent differences in Bayesian posterior distribution means between accuracy for memory-based questions whose answers were included in ChatGPT study guides and accuracy for memory-based questions whose answers were not included in ChatGPT study guides within each study condition. Error bars represent 95% credible intervals for differences in Bayesian posterior distribution means between question types. Bayes Factors were computed via the Savage-Dickey density ratio.

Figure 7c

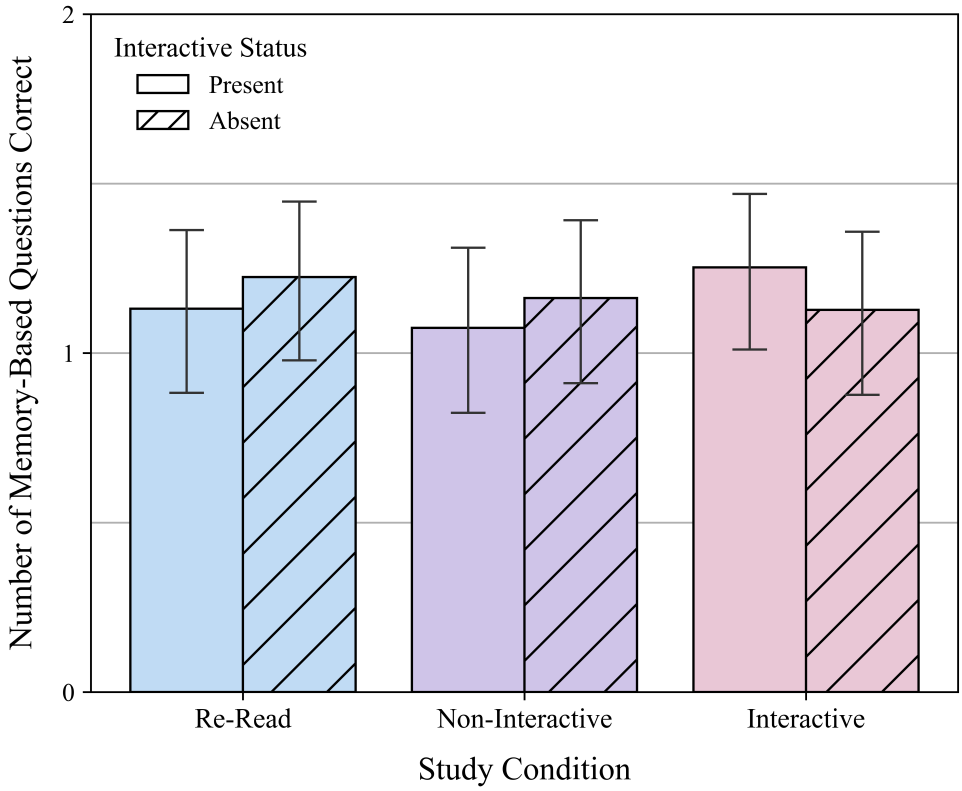
Paired Comparisons of Memory Test Performance Between Questions Present and Absent from Study Guides by Study Condition



Note. Points represent differences in Bayesian posterior distribution means between accuracy for memory-based questions whose answers were included in ChatGPT study guides and accuracy for memory-based questions whose answers were not included in ChatGPT study guides within each study condition. Error bars represent 95% credible intervals for differences in Bayesian posterior distribution means between question types. Bayes Factors were computed via the Savage-Dickey density ratio.

Figure 8a

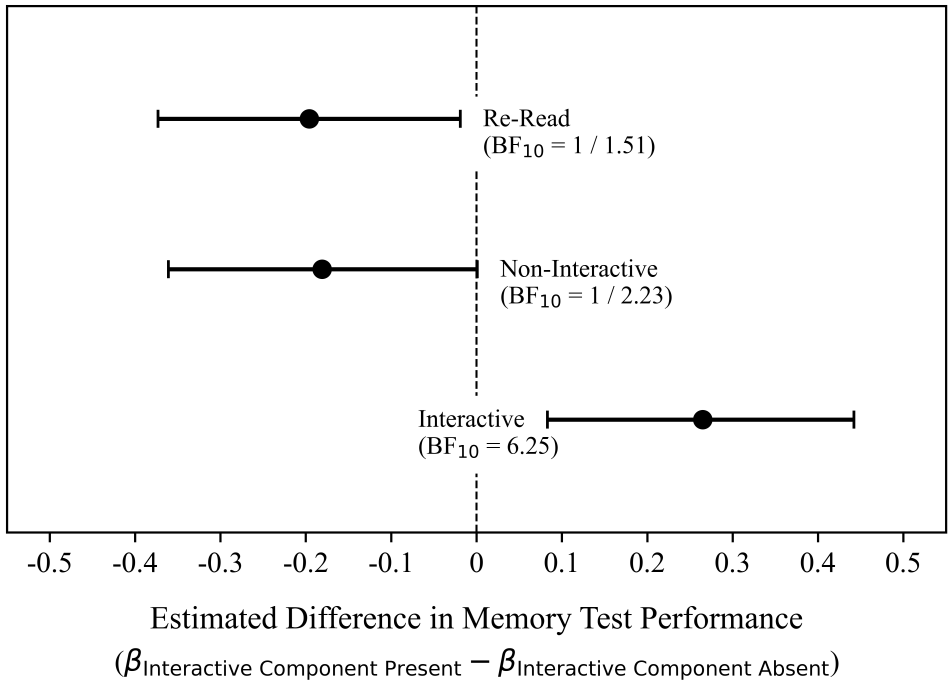
Model Estimates of Memory Test Performance by Study Condition and Interactive Component Status



Note. Solid bars represent Bayesian posterior distribution means for the number of questions answered correctly of the two memory-based test questions whose answers were included in interactive components of ChatGPT study guides ("Present"). Striped bars represent Bayesian posterior distribution means for the number of questions answered correctly of the two memory-based test questions whose answers were included in ChatGPT study guides, but not included in interactive components ("Absent"). Error bars represent 95% credible intervals for Bayesian posterior distributions.

Figure 8b

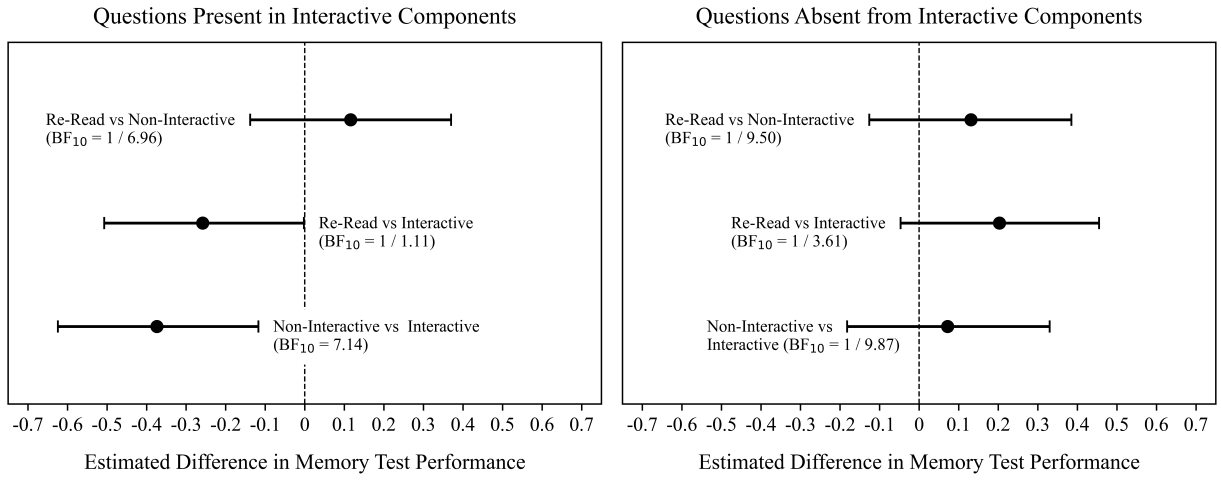
Paired Comparisons of Memory Test Performance Between Questions Present and Absent from Interactive Components by Study Condition



Note. Points represent differences in Bayesian posterior distribution means between accuracy for memory-based questions whose answers were included in interactive components of ChatGPT study guides and accuracy for memory-based questions whose answers were included in ChatGPT study guides, but not included in interactive components, within each study condition. Error bars represent 95% credible intervals for differences in Bayesian posterior distribution means between question types. Bayes Factors were computed via the Savage-Dickey density ratio.

Figure 8c

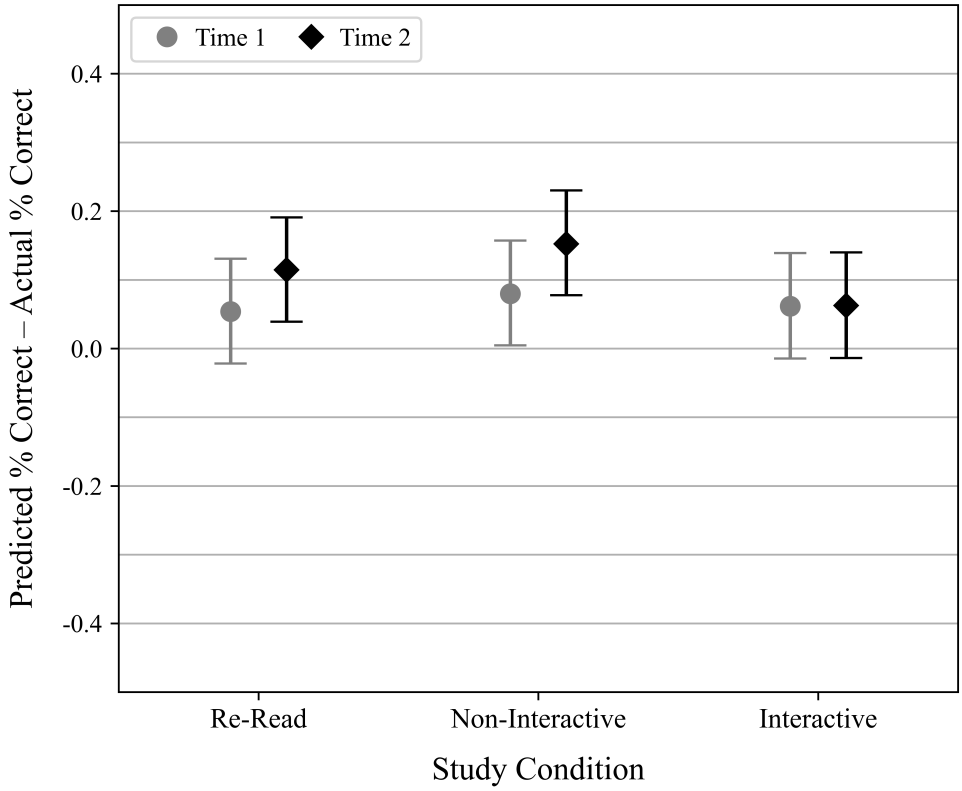
Paired Comparisons of Memory Test Performance Between Study Conditions by Interactive Component Status



Note. Points represent differences in Bayesian posterior distribution means between study conditions for the number of questions answered correctly of either the two memory-based test questions whose answers were included in interactive components of ChatGPT study guides (left plot) or the two memory-based test questions whose answers were included in ChatGPT study guides, but not included in interactive components (right plot). Error bars represent 95% credible intervals for differences in Bayesian posterior distribution means between study conditions. Bayes Factors computed via the Savage-Dickey density ratio.

Figure 9a

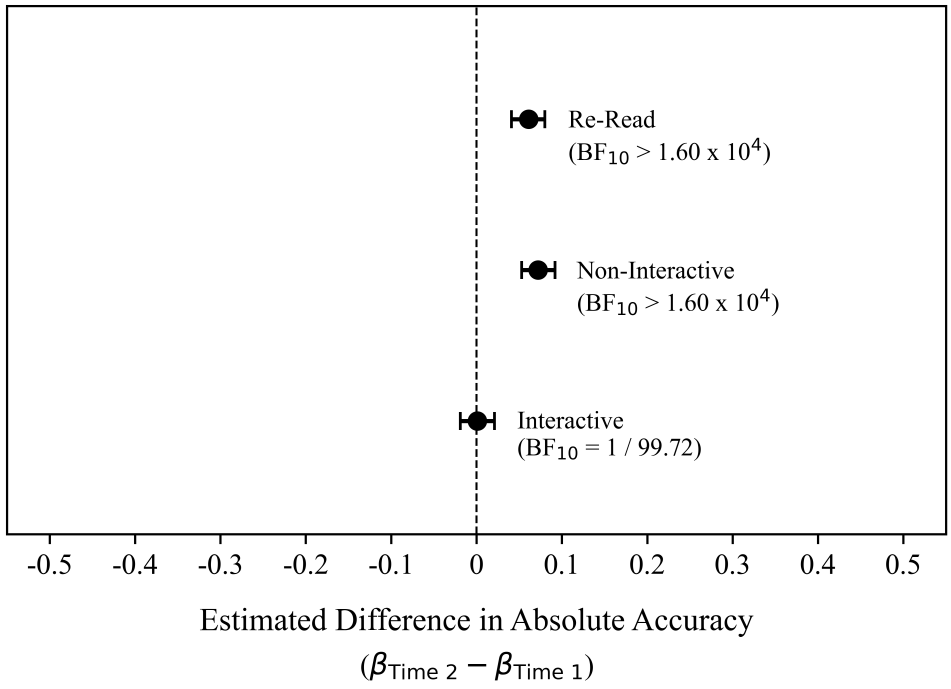
Model Estimates of Absolute Metacomprehension Accuracy by Study Condition and Time of Judgment



Note. Round gray points represent Bayesian posterior distribution means for absolute metacomprehension accuracy of predictions of performance made following the initial reading phase ("Time 1"). Diamond-shaped black points represent Bayesian posterior distribution means for absolute metacomprehension accuracy of predictions of performance made following the study phase ("Time 2"). Error bars represent 95% credible intervals for Bayesian posterior distributions.

Figure 9b

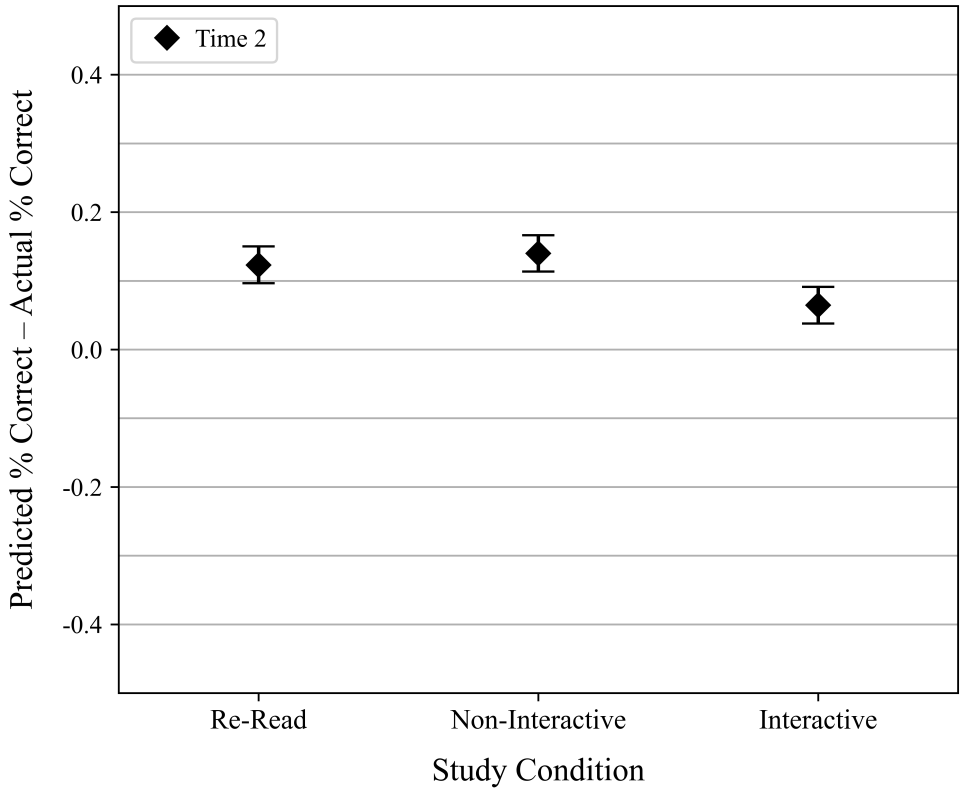
Paired Comparisons of Absolute Metacomprehension Accuracy Between Time 1 and Time 2 Judgments by Study Condition



Note. Points represent differences in Bayesian posterior distribution means between absolute metacomprehension accuracy of predictions of performance made following the study phase ("Time 2") and absolute metacomprehension accuracy of predictions of performance made following the initial reading phase ("Time 1") within each study condition. Error bars represent 95% credible intervals for differences in Bayesian posterior distribution means between time points. Bayes Factors were computed via the Savage-Dickey density ratio.

Figure 10a

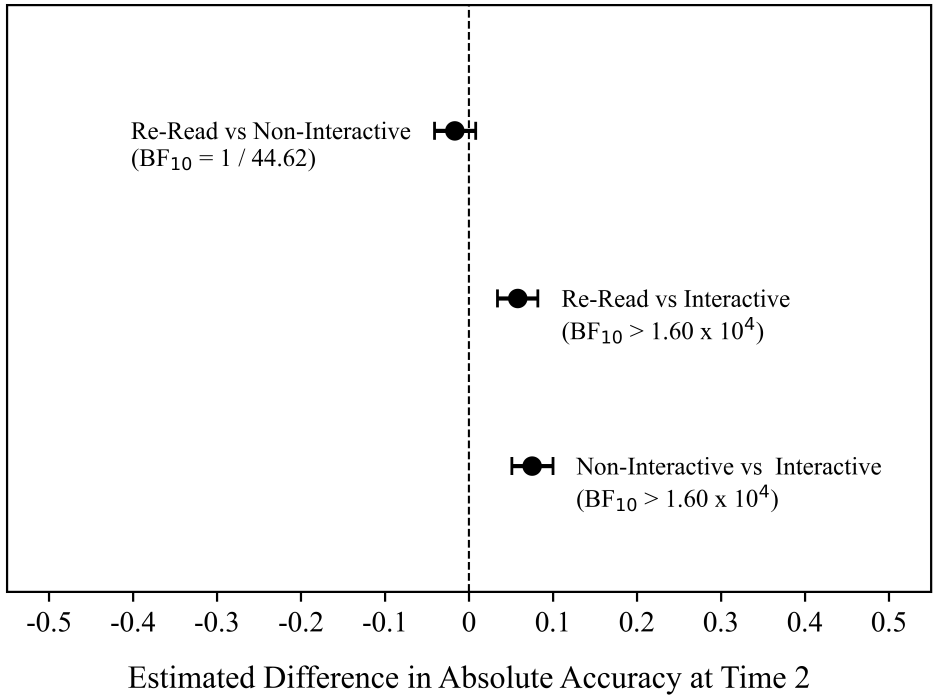
Model Estimates of Absolute Metacomprehension Accuracy at Time 2 by Study Condition



Note. Diamond-shaped black points represent Bayesian posterior distribution means for absolute metacomprehension accuracy of predictions of performance made following the study phase ("Time 2"). Error bars represent 95% credible intervals for Bayesian posterior distributions.

Figure 10b

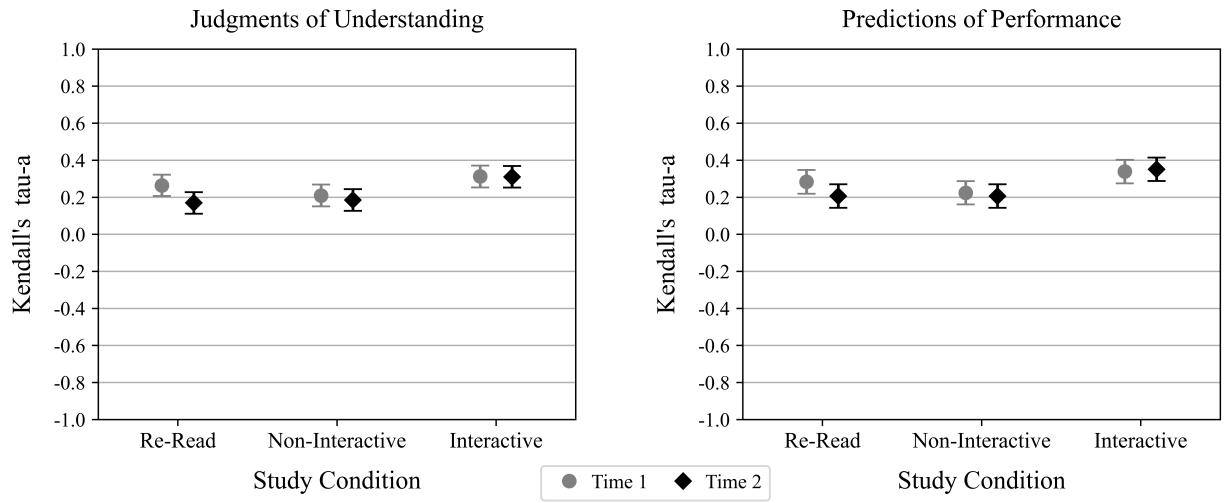
Paired Comparisons of Absolute Metacomprehension Accuracy at Time 2 between Study Conditions



Note. Points represent differences in Bayesian posterior distribution means between study conditions for absolute metacomprehension accuracy of predictions of performance made following the study phase ("Time 2"). Error bars represent 95% credible intervals for differences in Bayesian posterior distribution means between study conditions. Bayes Factors computed via the Savage-Dickey density ratio.

Figure 11a

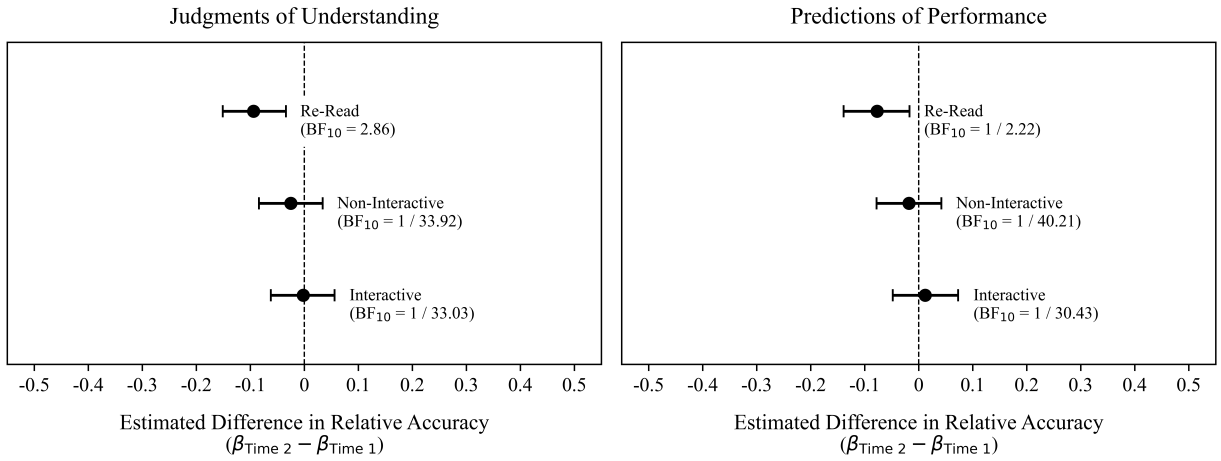
Model Estimates of Relative Metacomprehension Accuracy by Judgment Type, Judgment Time, and Study Condition



Note. Round grey points represent Bayesian posterior distribution means for relative metacomprehension accuracy of judgments of understanding (left plot) and predictions of performance (right plot) made following the initial reading phase ("Time 1"). Diamond-shaped black points represent Bayesian posterior distribution means for relative metacomprehension accuracy of judgments of understanding (left plot) and predictions of performance (right plot) made following the study phase ("Time 2"). Error bars represent 95% credible intervals for Bayesian posterior distributions.

Figure 11b

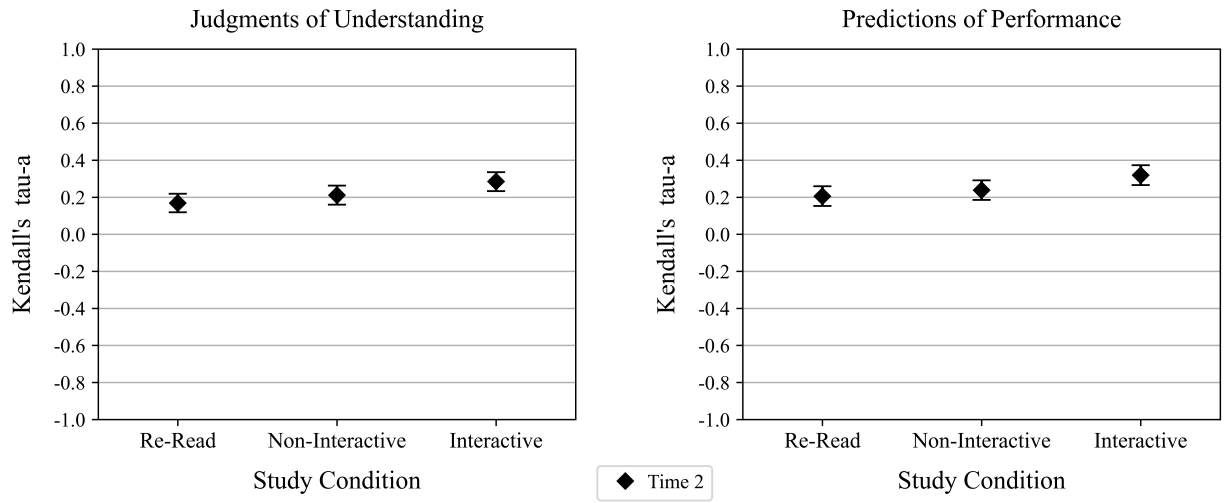
Paired Comparisons of Relative Metacomprehension Accuracy Between Time 1 and Time 2 Judgments by Study Condition and Judgment Type



Note. Left plot: points represent differences in Bayesian posterior distribution means between relative metacomprehension accuracy of judgments of understanding made following the study phase ("Time 2") and relative metacomprehension accuracy of judgments of understanding made following the initial reading phase ("Time 1") within each study condition. Right plot: points represent differences in Bayesian posterior distribution means between relative metacomprehension accuracy of predictions of performance made following the study phase ("Time 2") and relative metacomprehension accuracy of predictions of performance made following the initial reading phase ("Time 1") within each study condition. Error bars represent 95% credible intervals for differences in Bayesian posterior distribution means between time points. Bayes Factors were computed via the Savage-Dickey density ratio.

Figure 12a

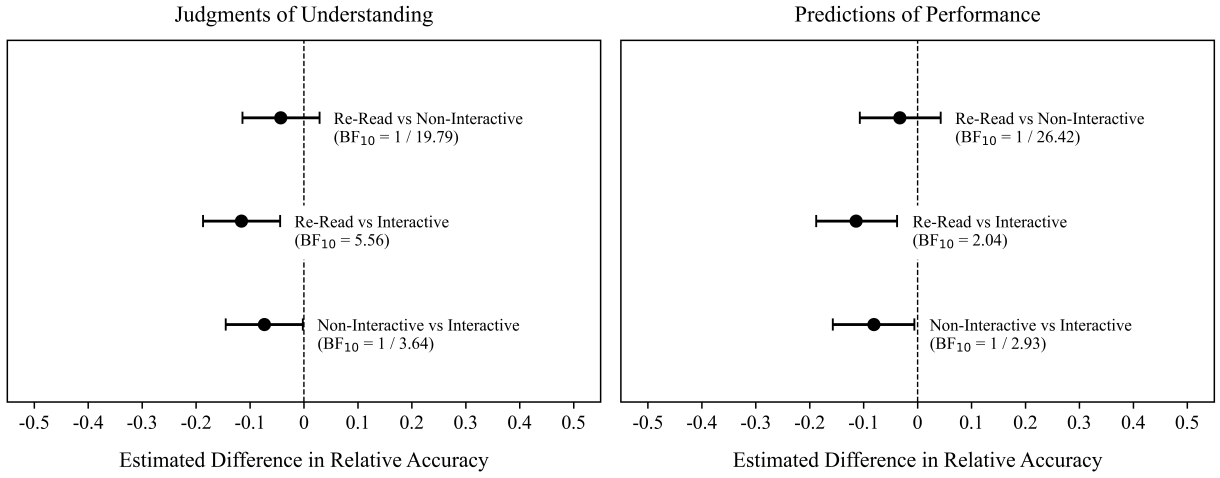
Model Estimates of Relative Metacomprehension Accuracy at Time 2 by Judgment Type and Study Condition



Note. Diamond-shaped black points represent Bayesian posterior distribution means for relative metacomprehension accuracy of judgments of understanding (left plot) and predictions of performance (right plot) made following the study phase ("Time 2"). Error bars represent 95% credible intervals for Bayesian posterior distributions.

Figure 12b

Paired Comparisons of Relative Metacomprehension Accuracy at Time 2 Between Study Conditions by Judgment Type



Note. Points represent differences in Bayesian posterior distribution means between study conditions for relative metacomprehension accuracy of judgments of understanding (left plot) and predictions of performance (right plot) made following the study phase ("Time 2"). Error bars represent 95% credible intervals for differences in Bayesian posterior distribution means between study conditions. Bayes Factors computed via the Savage-Dickey density ratio.

References

Agler, L.-M. L., Noguchi, K., & Alfsen, L. K. (2021). Personality traits as predictors of reading comprehension and metacomprehension accuracy. *Current Psychology*, *40*(10), 5054–5063. https://doi.org/10.1007/s12144-019-00439-y

Amoah, A., Asiama, R. K., & Kwablah, E. (2025). ChatGPT early usage among students: A global evidence of determinants. *Development and Sustainability in Economics and Finance*, *7*, 100065. https://doi.org/10.1016/j.dsef.2025.100065

Anderson, M. C., & Thiede, K. W. (2008). Why do delayed summaries improve metacomprehension accuracy? *Acta psychologica*, *128*(1), 110–118. https://doi.org/10.1016/j.actpsy.2007.10.006

Böckle, M., Yeboah-Antwi, K., & Kouris, I. (2021). Can you trust the black box? the effect of personality traits on trust in AI-enabled user interfaces. *International Conference on Human-Computer Interaction*, 3–20. https://doi.org/10.1007/978-3-030-77772-2_1

Bürkner, P.-C. (2017). Brms: An R package for bayesian multilevel models using stan. *Journal of Statistical Software*, *80*(1), 1–28. https://doi.org/10.18637/jss.v080.i01

Chiang, E. S., Therriault, D. J., & Franks, B. A. (2010). Individual differences in relative metacomprehension accuracy: Variation within and across task manipulations. *Metacognition and Learning*, *5*(2), 121–135. https://doi.org/10.1007/s11409-009-9052-6

Dirkx, K. J., Kester, L., & Kirschner, P. A. (2014). The testing effect for learning principles and procedures from texts. *The Journal of Educational Research*, *107*(5), 357–364. https://doi.org/10.1080/00220671.2013.823370

Dunlosky, J. (2005). Why does rereading improve metacomprehension accuracy? Evaluating the levels-of-disruption hypothesis for the rereading effect. *Discourse Processes*, *40*(1), 37–55. https://doi.org/10.1207/s15326950dp4001_2

Dunlosky, J., Rawson, K. A., & Middleton, E. L. (2005). What constrains the accuracy of meta-comprehension judgments? Testing the transfer-appropriate-monitoring and accessibility hypotheses [Special Issue on Metamemory]. *Journal of Memory and Language*, *52*(4), 551–565. https://doi.org/https://doi.org/10.1016/j.jml.2005.01.011

Griffin, T. D., Wiley, J., & Thiede, K. W. (2008). Individual differences, rereading, and self-explanation: Concurrent processing and cue validity as constraints on metacomprehension accuracy. *Memory & Cognition*, *36*(1), 93–103. https://doi.org/10.3758/MC.36.1.93

Griffin, T. D., Wiley, J., & Thiede, K. W. (2019). The effects of comprehension-test expectancies on metacomprehension accuracy. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *45*(6), 1066. https://doi.org/10.1037/xlm0000634

Hinze, S. R., & Wiley, J. (2011). Testing the limits of testing effects using completion tests. *Memory*, *19*(3), 290–304. https://doi.org/10.1080/09658211.2011.560121

Hoffman, R. R., Mueller, S. T., Klein, G., & Litman, J. (2023). Measures for explainable AI: Explanation goodness, user satisfaction, mental models, curiosity, trust, and human-AI performance. *Frontiers in Computer Science*, *5*. https://doi.org/10.3389/fcomp.2023.1096257

Ibrahim, H., Liu, F., Asim, R., Battu, B., Benabderrahmane, S., Alhafni, B., Adnan, W., Alhanai, T., AlShebli, B., Baghdadi, R., et al. (2023). Perception, performance, and detectability of

conversational artificial intelligence across 32 university courses. *Scientific reports*, *13*(1), 12187. https://doi.org/10.1038/s41598-023-38964-3

Johnson, D. M., Doss, W., & Estepp, C. M. (2024). Using chatGPT with novice arduino programmers: Effects on performance, interest, self-efficacy, and programming ability. *Journal of Research in Technical Careers*, *8*(1), 1. https://doi.org/10.9741/2578-2118.1152

Kiewra, K. A., Kauffman, D. F., Robinson, D. H., Dubois, N. F., & Staley, R. K. (1999). Supplementing floundering text with adjunct displays. *Instructional Science*, *27*(5), 373–401. https://doi.org/10.1007/BF00892032

Kintsch, W. (1988). The role of knowledge in discourse comprehension: A construction-integration model. *Psychological review*, *95*(2), 163. https://doi.org/10.1037/0033-295X.95.2.163

Kintsch, W. (1994). Text comprehension, memory, and learning. *American psychologist*, *49*(4), 294. https://doi.org/10.1037/0003-066X.49.4.294

Kosar, T., Ostojić, D., Liu, Y. D., & Mernik, M. (2024). Computer science education in chatGPT era: Experiences from an experiment in a programming course for novice programmers. *Mathematics*, *12*(5). https://doi.org/10.3390/math12050629

Krug, D., George, B., Hannon, S. A., & Glover, J. A. (1989). The effect of outlines and headings on readers' recall of text. *Contemporary Educational Psychology*, *14*(2), 111–123. https://doi.org/10.1016/0361-476X(89)90029-5

Lee, C., Myung, J., Han, J., Jin, J., & Oh, A. (2023). Learning from teaching assistants to program with subgoals: Exploring the potential for AI teaching assistants. *arXiv preprint arXiv:2309.10419*. https://doi.org/10.48550/arXiv.2309.10419

Lee, M., & Wagenmakers, E. (2014). *Bayesian cognitive modeling: A practical course*. Cambridge university press.

Liu, R., Zenke, C., Liu, C., Holmes, A., Thornton, P., & Malan, D. J. (2024). Teaching CS50 with AI: Leveraging generative artificial intelligence in computer science education. *Proceedings of the 55th ACM Technical Symposium on Computer Science Education V. 1*, 750–756. https://doi.org/10.1145/3626252.3630938

Maki, R. H. (1998). Test predictions over text material. In D. Hacker, J. Dunlosky, & A. Graesser (Eds.), *Metacognition in educational theory and practice* (pp. 117–144). Routledge.

Maki, R. H., & Berry, S. L. (1984). Metacomprehension of text material. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*(4), 663. https://doi.org/10.1037/0278-7393.10.4.663

Marefat, H., & Ghahari, S. (2009). (incorporating) adjunct displays: A step toward facilitation of reading comprehension. *Porta Linguarum*, *11*, 179–188. https://doi.org/10.30827/Digibug.31846

Margolin, S. J., & Snyder, N. (2018). It may not be that difficult the second time around: The effects of rereading on the comprehension and metacomprehension of negated text. *Journal of Research in Reading*, *41*(2), 392–402. https://doi.org/10.1111/1467-9817.12114DigitalObjectIdentifier(DOI)

McDaniel, M. A., Anderson, J. L., Derbish, M. H., & Morrisette, N. (2007). Testing the testing effect in the classroom. *European journal of cognitive psychology*, *19*(4-5), 494–513. https://doi.org/10.1080/09541440701326154

Moradi, S., Ghahari, S., & Abbas Nejad, M. (2020). Learner- vs. expert-constructed outlines: Testing the associations with L2 text comprehension and multiple intelligences. *Studies in Second Language Learning and Teaching*, *10*(2), 359–384. https://doi.org/10.14746/ssllt.2020.10.2.7

Oksanen, A., Savela, N., Latikka, R., & Koivula, A. (2020). Trust toward robots and artificial intelligence: An experimental approach to human–technology interactions online. *Frontiers in Psychology*, *Volume 11 - 2020*. https://doi.org/10.3389/fpsyg.2020.568256

Perrig, S. A. C., Scharowski, N., & Brühlmann, F. (2023). Trust issues with trust scales: Examining the psychometric quality of trust measures in the context of AI. *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3544549.3585808

Ponce, H. R., Mayer, R. E., & Mendez, E. E. (2023). Effects of learner-generated outlining and instructor-provided outlining on learning from text: A meta-analysis. *Educational Research Review*, *39*, 100538. https://doi.org/10.1016/j.edurev.2023.100538

R Core Team. (2025). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

Rawson, K. A., & Dunlosky, J. (2002). Are performance predictions for text based on ease of processing? *Journal of experimental psychology: Learning, memory, and cognition*, *28*(1), 69. https://doi.org/10.1037/0278-7393.28.1.69

Rawson, K. A., Dunlosky, J., & Thiede, K. W. (2000). The rereading effect: Metacomprehension accuracy improves across reading trials. *Memory & Cognition*, *28*(6), 1004–1010. https://doi.org/10.3758/BF03209348

Rawson, K. A., & Kintsch, W. (2005). Rereading effects depend on time of test. *Journal of educational psychology*, *97*(1), 70. https://doi.org/10.1037/0022-0663.97.1.70

Roediger, H. L., & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological science*, *17*(3), 249–255.

Thiede, K. W., Anderson, M., & Therriault, D. (2003). Accuracy of metacognitive monitoring affects learning of texts. *Journal of educational psychology*, *95*(1), 66.

Thiede, K. W., & Anderson, M. C. (2003). Summarizing can improve metacomprehension accuracy. *Contemporary Educational Psychology*, *28*(2), 129–160. https://doi.org/10.1016/S0361-476X(02)00011-5

Thiede, K. W., Dunlosky, J., Griffin, T. D., & Wiley, J. (2005). Understanding the delayed-keyword effect on metacomprehension accuracy. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(6), 1267. https://doi.org/10.1037/0278-7393.31.6.1267

Thiede, K. W., Wiley, J., & Griffin, T. D. (2011). Test expectancy affects metacomprehension accuracy. *British Journal of Educational Psychology*, *81*(2), 264–273. https://doi.org/10.1348/135910710X510494

Urban, M., Děchtěrenko, F., Lukavský, J., Hrabalová, V., Svacha, F., Brom, C., & Urban, K. (2024). ChatGPT improves creative problem-solving performance in university students: An experimental study. *Computers and Education*, *215*, 105031. https://doi.org/10.1016/j.compedu.2024.105031

van Gog, T., & Sweller, J. (2015). Not new, but nearly forgotten: The testing effect decreases or even disappears as the complexity of learning materials increases. *Educational Psychology Review*, *27*(2), 247–264. https://doi.org/10.1007/s10648-015-9310-x

Wang, J., & Fan, W. (2025). The effect of chatGPT on students' learning performance, learning perception, and higher-order thinking: Insights from a meta-analysis. *Humanities and Social Sciences Communications*, *12*(1), 1–21. https://doi.org/10.1057/s41599-025-04787-y

Wong, A. Y., & Moss, J. (2022). Metacomprehension and regressions during reading and rereading. *Metacognition and Learning*, *17*(1), 53–71. https://doi.org/10.1007/s11409-021-09272-w

Yu, S.-C., Huang, Y.-M., & Wu, T.-T. (2024). Tool, threat, tutor, talk, and trend: College students' attitudes toward chatGPT. *Behavioral Sciences*, *14*(9). https://doi.org/10.3390/bs14090755