AMERICAN PSYCHOLOGICAL ASSOCIATION
SARMAC
Journal of Applied Research in Memory and Cognition


ISSN: 2211-3681
https://doi.org/10.1037/mac0000254

EMPIRICAL ARTICLE

# Using ChatGPT-Generated Prequestions to Improve Memory and Text Comprehension

Steven C. Pan[1], Judith Schweppe[2, 3], Andy Z. J. Teo[1], Alyssa Indrajaya[1], and Niklas Wenzel[4]
[1] Department of Psychology, Faculty of Arts and Social Sciences, National University of Singapore, Singapore
[2] Faculty of Social and Educational Sciences, University of Passau, Germany
[3] Faculty of Education, University of Erfurt, Germany
[4] School of Health Professions Education, Maastricht University, Netherlands


Prequestioning is a learning method that involves guessing the answers to questions about to-be-learned information (i.e., prequestions), followed by the opportunity to discover the correct answers. Its application during self-regulated learning requires access to practice questions, which can often be quite limited. Across four experiments, we investigated a potential solution: using artificial intelligence (AI)—specifically, the large language model-based chatbot, ChatGPT—to generate prequestions. Experiment 1 found that AI-generated prequestioning enhanced the learning of an educational text relative to simply reading. Experiment 2 found that AI-generated questions and human-generated questions were equally effective at promoting learning. Experiments 3–4 revealed that prequestioning with questions generated using different AI prompts aided learning relative to studying an AI-generated outline. Across experiments, prequestioning benefited subsequent test performance on questions that were identical and nonidentical to those previously encountered. Together, these results highlight the potential of generative AI to facilitate prequestioning, yielding substantial learning benefits.

***General Audience Summary***

This study investigated a learning method known as prequestioning, which involves making guesses in response to practice questions about information that has not yet been learned, followed by an opportunity to learn the correct answers. For example, one might guess the answers to questions about atomic structure before viewing a lesson on that topic. Doing so requires practice questions, which are called prequestions when used for prequestioning, but such questions are not always readily available or easy to create. Consequently, it can be quite difficult to use prequestioning in many circumstances, such as when a student is studying on their own. To potentially solve this problem, we explored the use of artificial intelligence (AI) to generate prequestions for learning purposes. In each of four experiments, study participants either used or did not use AI-generated prequestions before reading a text about brakes and then took a memory and comprehension test. In Experiment 1, participants who used AI-generated prequestions scored higher on the test than those who did not. In Experiment 2, participants who used AI-generated prequestions performed just as well as those who used prequestions generated by human beings. In Experiments 3 and 4, participants who used AI-generated prequestions of various types scored better than those who studied an AI-generated outline. Together, these results suggest that using AI to generate prequestions is a useful way to implement prequestioning in a widely accessible and customizable manner. It is of similar effectiveness as prequestioning using questions made by human beings and of greater effectiveness than some approaches that do not involve prequestioning. We conclude that learners should be able to use AI to easily make and use prequestions in a variety of ways, with a potential result being improved learning of educational texts and other materials.

*Keywords:* prequestioning, question generation, generative artificial intelligence, large language model, ChatGPT

*Supplemental materials:* https://doi.org/10.1037/mac0000254.supp

Sean Kang served as action editor.
Steven C. Pan https://orcid.org/0000-0001-9080-5651
Judith Schweppe https://orcid.org/0000-0002-8661-3072
Andy Z. J. Teo https://orcid.org/0009-0009-3842-1031
Alyssa Indrajaya https://orcid.org/0000-0003-1719-2777
Niklas Wenzel https://orcid.org/0009-0008-0222-467X

*continued*

A growing body of research suggests that prequestioning, a learning method that involves guessing the answers to questions about information that one has yet to learn (i.e., prequestions), followed by an opportunity to learn the correct answers, can benefit the learning of text passages (e.g., Richland et al., 2009), video lectures (e.g., Carpenter & Toftness, 2017), and other materials (for reviews, see Pan & Carpenter, 2023; St Hilaire et al., 2024; see also Metcalfe, 2017). Studies of prequestioning typically involve research participants using prequestioning, no prequestioning, or an alternative approach (e.g., reading learning objectives) to initially learn or become familiar with a set of materials. Next, they receive further instruction on those materials (e.g., reading a text passage), and, after a retention interval of a few minutes or more, take a criterial test. On that test, a prequestioning effect—that is, better performance following prequestioning than without—is often observed. Various theories explain the prequestioning effect, with several major accounts suggesting that prequestioning activates cognitive mechanisms (e.g., improved attention) or prior knowledge (e.g., mental models) that enable learners to better profit from subsequent learning opportunities (Kornell & Vaughn, 2016; Mera et al., 2021; see also Motz et al., 2024; Pan & Sana, 2021; Sana & Carpenter, 2023).

Although the pedagogical benefits of prequestioning have become increasingly apparent to researchers, the method remains obscure and rarely used in classrooms, during self-regulated learning, and in other settings (Pan et al., 2020). The paucity of suitable practice questions appears to be a contributing factor. Survey data indicate that undergraduate instructors offer practice questions prior to instruction as infrequently as 37%, compared with 76% afterward, which is too late for prequestioning to occur (Pan et al., 2020). Despite some instructors having access to question banks, a lack of provided practice questions is reflected in a search of the RateMyProfessors.com website in May 2025, which reveals at least 300 complaints about insufficient practice or review questions, as well as insufficient practice exercises. Moreover, across domains ranging from aeronautics to medicine, common textbooks lack practice questions for key topics or entirely (e.g., I. Anderson & Sheikh, 2018; Domingo et al., 2022; Thobroni et al., 2022).

Although students might obtain practice questions from online repositories or other sources, such questions are not necessarily plentiful, well-written, accurate, or suitable for prequestioning (for related findings, see Pan et al., 2023; Zung et al., 2022). For prequestions to be effective, they must align with upcoming material in a way that facilitates learning through answer discovery (St. Hilaire & Carpenter, 2020). For example, Hausman and Rhodes (2018) found that "conceptual" prequestions—requiring inferences from a forthcoming text—were less effective than "factual" prequestions with verbatim answers, possibly because the answers were difficult to locate. At the same time, overly narrow prequestions may also be suboptimal. St. Hilaire et al. (2019) showed that "isolative" prequestions, targeting a single detail, were less effective than "integrative" ones that spanned multiple ideas. Overall, these findings suggest that the effectiveness of prequestions depends on the scope of the question, how easily the question can be understood and remembered, and the likelihood that the correct answers can be located in subsequent materials. Notably, such distinctions do not apply equally to retrieval practice, where the memorability of the questions themselves is less important, there is no need to engage in answer discovery, and "conceptual"-type questions are among the most beneficial for learning (see Butler, 2010; Pan & Rickard, 2018).

Compounding these challenges, practice questions are commonly written with the assumption that learners have already studied the material, which is incompatible with prequestioning (to foreshadow, we encountered this challenge in developing the prequestions for the present study). Moreover, given their unfamiliarity with the material, it can be prohibitively difficult for students to evaluate the quality or appropriateness of a given set of practice questions that might be used for prequestioning, much less generate them independently. Given these complexities and the early stage of research

Materials, data, and analysis code are publicly available on the Open Science Framework and can be accessed at https://osf.io/x3enc/. The authors have no conflicts of interest to declare. Portions of this research were presented by Steven C. Pan in a guest lecture at the University of Passau in August 2024.

This research was supported by a Research Experience Programme Grant (Undergraduate Research Opportunities Programme-Research Experience) from the Undergraduate Research Opportunities Programme awarded to Andy Z. J. Teo under the supervision of Steven C. Pan and a Faculty of Arts and Social Sciences (FASS) grant awarded to Steven C. Pan, both funded by the National University of Singapore, as well as an International Visiting Academic's program grant from the Faculty of Social and Educational Sciences at the University of Passau, facilitated by Judith Schweppe on behalf of Steven C. Pan. The authors thank Aruna Kandasamy, Nathaneal Teo, and Valerie Tan for their scoring assistance.

Conceptualization: Steven C. Pan, Judith Schweppe, and Andy Z. J. Teo; data curation: Andy Z. J. Teo and Alyssa Indrajaya; formal analysis: Steven C. Pan, Andy Z. J. Teo, and Alyssa Indrajaya; funding acquisition: Steven C. Pan; investigation: Andy Z. J. Teo and Alyssa Indrajaya; methodology: Steven C. Pan, Judith Schweppe, Andy Z. J. Teo, and Niklas Wenzel; project administration: Steven C. Pan and Andy Z. J. Teo; Resources: Steven C. Pan, Andy Z. J. Teo, and Niklas Wenzel; software: Niklas Wenzel; supervision: Steven C. Pan; validation: Alyssa Indrajaya; visualization: Steven C. Pan; writing–original draft: Steven C. Pan; writing–review and editing: Steven C. Pan, Judith Schweppe, Andy Z. J. Teo, Alyssa Indrajaya, and Niklas Wenzel.

Steven C. Pan played a lead role in conceptualization, funding acquisition, methodology, project administration, supervision, visualization, and writing–original draft, a supporting role in resources, and an equal role in formal analysis and writing–review and editing. Judith Schweppe played a lead role in conceptualization and a supporting role in methodology and writing–review and editing. Andy Z. J. Teo played a lead role in data curation, formal analysis, and investigation and a supporting role in conceptualization, funding acquisition, methodology, project administration, resources, and writing–review and editing. Alyssa Indrajaya played a supporting role in data curation, formal analysis, investigation, validation, and writing–review and editing. Niklas Wenzel played a supporting role in methodology, resources, and writing–review and editing.

The data are available at https://osf.io/x3enc/.

The experimental materials are available at https://osf.io/jeahn/.

The preregistered design is available at https://aspredicted.org/N5N_PS1, https://aspredicted.org/2XJ_KD6, https://aspredicted.org/4bzz-jr93.pdf.

Correspondence concerning this article should be addressed to Steven C. Pan, Department of Psychology, Faculty of Arts and Social Sciences, National University of Singapore, 9 Arts Link, Singapore City 117572, Singapore. Email: scp@nus.edu.sg

in this area, students may struggle to obtain or create effective prequestions without expert guidance or expertise.

## Are AI-Generated Questions the Solution?

Recent developments in generative artificial intelligence (AI), especially in natural language processing, suggest a possible solution: use AI to produce practice questions. Initial attempts at programming AI to produce questions relied on rule-based techniques before progressing to data-driven machine learning approaches, with somewhat mixed results (W. Chan et al., 2023). New advancements in large language models (LLMs), however, have enabled the creation of AI-generated practice questions with substantial efficiency, ease, customizability, and abundance (W. Chan et al., 2023; Indran et al., 2024; Singh et al., 2023). Even popular learning websites (e.g., Quizlet) have recently introduced AI-based question generation features. These advancements suggest that current LLMs can be successfully applied to generate prequestions.

OpenAI's Chat Generative Pre-trained Transformer (ChatGPT), an advanced LLM, is an AI chatbot that can analyze text, compose coherent text, comprehend context, and engage in conversations with human beings. Built upon artificial neural networks, it is fine tuned for language understanding and generation, heightening its potential for educational applications such as question generation, answering questions, and providing feedback (Indran et al., 2024; Kasneci et al., 2023; see also Dai et al., 2023; Imundo et al., 2024). Since its public launch in November 2022, ChatGPT has become one of the most popular AI tools in the world, with over 600 million regular users (Apostolopoulos et al., 2023). Although its underlying architecture is not public, its capabilities have improved across successive versions, including GPT-3.5 (2022), GPT-4 (2023), and GPT-4o (2024).

Initial research on the question-generating capabilities of ChatGPT has focused on the quality of its outputs (Kıyak & Emekli, 2024). For instance, W. Chan et al. (2023) found that without carefully fine-tuned prompts (i.e., typed instructions or commands given to the chatbot), GPT-3.5 generated "convoluted" questions consisting of several questions combined into a single sentence. Singh et al.'s (2023) analysis of GPT-3.5-generated questions using Bloom's taxonomy of cognitive complexity (L. W. Anderson, 2009) found that ChatGPT tended to produce lower level questions at the "remember," "understand," and "apply" levels, suggesting difficulties with higher order reasoning skills (i.e., at the evaluate and create levels). These results suggest that although promising, the question-generating capacities of ChatGPT can still be refined. Accordingly, some researchers advise users to engage ChatGPT in an iterative process of question generation, employing increasingly refined prompts until a desirable result is achieved (e.g., Indran et al., 2024; Lee et al., 2024). As previously noted, however, in the case of prequestioning, learners may struggle to evaluate prequestion quality.

Although several researchers have suggested using ChatGPT to generate prequestions (e.g., Arango-Ibanez et al., 2024), the approach has yet to be explored. Indeed, the pedagogical implications of using AI-generated questions more generally remain largely underexamined. With respect to prequestioning, the effectiveness of AI-generated prequestions in common learning contexts, such as for mastering the content of an educational text passage, is unclear. Related issues, such as the capacity of ChatGPT to produce integrative and other types of prequestions, are also yet to be addressed.

## The Present Study

We investigated the potential of AI-generated prequestioning for enhancing memory and comprehension of educational texts. In each of four experiments, participants engaged in prequestioning, no prequestioning, or an alternative activity prior to reading a text passage and then took a criterial test that assessed directly prequestioned and not directly prequestioned content. Collectively, these experiments answered five research questions pertaining to plausible uses of AI-generated prequestioning in pedagogical contexts, including effects on learning, implementation factors, and relative to alternative approaches. The questions (and the experiments that addressed them) were as follows:

1. What is the effectiveness of AI-generated practice questions in producing a prequestioning effect for directly tested information (Experiments 1–4)?
2. Do the benefits of AI-generated prequestioning transfer to criterial test questions that differ from the practice questions in some way, such as by targeting different content, being generated by human beings instead of AI (Experiments 1–4), or being generated using different prompts (Experiments 3–4)?
3. How do AI-generated and human-generated prequestioning compare (Experiment 2)?
4. Do prompts impact the efficacy of AI-generated prequestioning (Experiments 3–4)?
5. How does AI-generated prequestioning compare with a time-matched alternative learning activity such as studying an outline (Experiments 3–4)?

Together, these experiments explored several common circumstances that an instructor or student attempting to implement AI-generated prequestioning might encounter, at least in the following ways. For instance, reflecting how most users currently access ChatGPT, all experiments used the latest free version available at the time (GPT-3.5 or GPT-4o) instead of paid versions. In the initial experiments, maximum reading time was not restricted, as occurs during self-regulated learning, whereas later experiments controlled for reading time. Further, in line with W. Chan et al.'s (2023), Indran et al.'s (2024), and Lee et al.'s (2024) findings and advice to instructors, we used iteratively refined prompts to generate prequestions and, following the findings of St. Hilaire et al. (2019), focused on generating integrative prequestions to maximize learning benefits.

# Experiment 1

Experiment 1 investigated whether engaging in AI-generated prequestioning before reading an educational text improves the learning of directly tested and untested content, relative to reading the text alone.

## Method

This article reports, for all experiments, how we determined our sample size, all data exclusions, all manipulations, and all measures.

### *Design*

The experiment used a 2 × 2 design wherein each participant was randomly assigned to an AI-prequestion group or a read-only control group. All participants completed a criterial test featuring two question types: *tested* questions, which were identical to those used during prequestioning in the AI-prequestion group, and *untested* questions, which were not identical to those used during prequestioning.

### *Participants*

The target sample size of 90 participants, split evenly between groups, was determined via an a priori power analysis conducted in G*Power (Faul et al., 2007). That analysis involved a two-tailed, independent-samples *t* test with the assumptions of a medium effect size (Cohen's $d = 0.6$, the prequestion effect magnitude by St. Hilaire et al., 2019), $\alpha = .05$, and 80% power, reflecting our interest in detecting a prequestioning effect for tested or untested content. One hundred nine participants were recruited via the research platform Prolific Academic, each receiving $4.00. All participants had to be from an English-speaking country (i.e., Australia, Canada, New Zealand, the United Kingdom, or the United States), between 21 and 45 years old, fluent in English, have a ≥95% approval rate on the platform, and to not have prior knowledge of the relevant subject matter (i.e., types and functions of vehicle brakes). The entire study was conducted with ethics approval obtained at the first author's affiliated university. All participants gave informed consent before participating and were treated in accordance with the Declaration of Helsinki.

Data from 19 participants were excluded prior to data analysis due to evidence of off-task browser activity, as detected via the TaskMaster add-on for Qualtrics (Permut et al., 2019). Participants were characterized as off-task if more than 20% of total time or 100 s consecutively were spent away from the experiment during the training phase or if any time was spent away from the experiment during the criterial test. The final sample of 90 participants (AI-prequestion group, $n = 45$; read-only control group, $n = 45$) had a mean age of 33.1 years and was 50% female. Forty percent of these participants were from the United Kingdom, 27% were from the United States, and 33% were from the remaining eligible countries; 10% of participants were Asian, 14% were Black, 8% were mixed, 61% were White, and 8% were from other ethnic groups or declined to provide ethnicity information.

### *Materials*

The materials consisted of an expository text passage, two AI-generated questions, and two human-generated questions. The text passage, "Brakes," was a 675-word, eleven-paragraph text adapted from the *World Book Encyclopedia*. Describing various types of brakes, the passage has been used in prior studies, including to demonstrate a prequestioning effect (St. Hilaire et al., 2019). Analysis using publicly available tools showed that the passage had a Flesch–Kincaid grade level of 7.44 and a Gunning Fog Index of 9.00 (indicating the grade level and years of education needed to understand the text; https://www.online-utility.org/). The passage was not accompanied by diagrams or pictures (cf. Mayer & Gallini, 1990).

The two AI-generated questions were created using an iteratively tested and adapted prompt in the latest version of ChatGPT that was freely available at the time: GPT-3.5 (https://chat.openai.com; see Appendix A for the final prompt). As advised by Indran et al. (2024), the prompt was applied to the "Brakes" passage and refined until it consistently produced desirable outputs, including questions that were of the integrative type described by St. Hilaire et al. (2019). To judge the suitability of the outputs, we applied a standard based on St. Hilaire et al., wherein an integrative question was defined as a question that targeted multiple pieces of information from the passage and not a single, isolated detail. Moreover, the question had to refer to actual passage content and be written in a sufficiently clear and understandable manner. In practice, we found that GPT-3.5, in response to the final prompt, produced questions that met our criteria with little apparent difficulty.

Despite the relative ease of producing suitable questions, it was necessary to instruct GPT-3.5 to avoid referring to the passage directly, as participants had not yet read it (GPT-3.5 assumed that the questions would be used after the passage had been read, which would be suitable for retrieval practice but is antithetical to prequestioning) and to refrain from generating multiple questions within a single question (just as observed by W. Chan et al., 2023). Further, given GPT-3.5's tendency to create long and very detailed questions, instructions to avoid inserting multiple clauses and a 15-word limit had to be instituted. The two resulting questions, created in October 2023 (see Table 1), required a relatively thorough understanding of the "Brakes" passage (referring to hydraulic brakes vs. mechanical brakes; components of drum brakes). These questions served as prequestions for the AI-prequestion group and as tested questions on the criterial test for both groups. Although such questions were new to the read-only group, they were consistently labeled as tested questions for comparison purposes.

The two human-generated questions (see Table 1) were originally used by St. Hilaire et al. (2019). These questions targeted mostly nonoverlapping content than the AI-generated questions but also required deep understanding of the passage to answer correctly (referring to train brakes vs. bicycle brakes; types of mounts in hydraulic brakes). A comparison of the idea units represented in the AI-generated versus human-generated questions is presented in the Supplemental Online Materials. Both questions were used as untested questions on the criterial test for both groups. The choice of questions generated by human beings and not AI was intentional; doing so served to explore the effects of AI-generated prequestioning versus no prequestioning when learning was assessed using never-before-seen questions not developed by the same or similar AI sources. Moreover, it reflected a plausible scenario wherein students practice with AI-generated prequestions prior to a human (i.e., instructor)-generated high-stakes exam. Readability and cognitive level metrics for all questions used in this study are presented in this article after the Results section for Experiment 4.

In addition, as an exploratory measure, we asked participants to guess the source of each question used in the entire study (human or AI generated). The results, which are included in the Supplemental Online Materials, revealed that adult learners

**Table 1**
*Artificial Intelligence-Generated and Human-Generated Questions*

| Type | Expt. | Question |
|---|---|---|
| Human generated | 1–4 | There are two key types of mounts in hydraulic brakes. What are the two types? Next, name two ways in which they differ from one another.[a] |
| | 1–4 | Train brakes and bicycle brakes rely on different systems to function. In what way are the two systems fundamentally different, and in what way are they similar?[a] |
| | 2–4 | Are air brakes more similar to mechanical brakes or hydraulic brakes? Why? |
| | 2–4 | What is the primary difference between mechanical brakes and hydraulic brakes? |
| AI-/elaborate generated | 1 | What distinguishes hydraulic brakes from mechanical brakes in automobiles? |
| | 1–3 | Can you describe the components and operation of drum brakes in detail?[a] |
| | 2, 3 | Explain the principle behind power brakes and their advantage for drivers[a] |
| | 2, 3 | How do mechanical brakes, like the caliper brake, function on a bicycle? |
| | 2, 3 | What is the role of compressed air in air brakes for buses, trucks, and trains? |
| AI-/elaborate generated (comparison) | 4 | How do hydraulic brakes differ from mechanical brakes in how they generate braking force? |
| | 4 | What is the primary similarity between disk brakes and bicycle caliper brakes in their operation? |
| | 4 | In what way do air brakes differ from hydraulic brakes in terms of the medium used to activate the brake shoes? |
| | 4 | How do power brakes enhance braking compared to regular hydraulic brakes? |
| AI-/basic generated | 3 | What are the three major kinds of brakes? |
| | 3 | What is the purpose of the main brake pipe in train brake systems? |
| | 3 | Explain the function of hydraulic brakes and outline the components involved in their operation. |
| | 3 | Describe the mechanism of electric brakes and identify where they are commonly used. |
| AI-/basic generated (comparison) | 4 | How do mechanical brakes on a bicycle differ from mechanical brakes on an automobile in terms of their components and function? |
| | 4 | In what ways do hydraulic brakes and air brakes differ in the mechanisms they use to apply pressure to the brake shoes? |
| | 4 | Compare the structure and function of drum brakes and disk brakes as used in automobiles. How does each type achieve the same result? |
| | 4 | What are the differences between power brakes and regular hydraulic brakes in terms of the additional force provided to the brake system? |

*Note.* AI-generated questions in Experiments 1–3 and 4 were produced by GPT-3.5 and GPT-4o, respectively. AI-generated questions in Experiments 1 and 2 were labeled as AI-elaborate questions in Experiments 3 and 4 to distinguish them from AI-basic questions. Four human-generated questions were adapted from St. Hilaire et al. (2019, p. 1213); all other questions in the table are original to the present study. Expt. = experiment; AI = artificial intelligence.
[a] Questions used as untested different-source questions in Experiment 2.

performed no better than chance at detecting AI-generated questions but were more successful at doing so for human-generated questions.

### Procedure

Experiment 1 entailed three phases: training, distractor task, and criterial test. An overview of the procedure is depicted in Figure 1 (top panel). Group assignment determined the activities that occurred during the training phase. In the read-only group, participants were instructed to read the "Brakes" passage very carefully and spent a minimum of 5 min doing so. In the AI-prequestion group, participants first attempted two prequestions, one question at a time. They were instructed to answer each question to the best of their ability and to remember both questions as an aid to finding the correct answers in subsequent reading materials. Each question had to be viewed for at least 20 s before a response, which was mandatory, could be typed into a provided textbox. Then, as a brief attention check, participants were prompted to type at least one recognizable detail from at least one prequestion (similar to that of St. Hilaire et al., 2019). Afterward, participants spent a minimum of 5 min reading the "Brakes" passage. They were instructed to do so until they had discovered the answers to the prequestions and were theoretically able to correctly answer such questions.[1]

The distractor task involved participants typing examples of five popular categories (e.g., movies) for 1 min per category. The criterial test followed. During the test, four questions were presented, one at a time, for a minimum of 20 s each and with no maximum time limit. The four criterial test questions (two AI generated, two human generated) were shuffled such that no two questions of the same type were presented consecutively. Participants were required to type an answer for each question into a provided textbox to advance. Once participants had finished the criterial test, they were debriefed and dismissed.

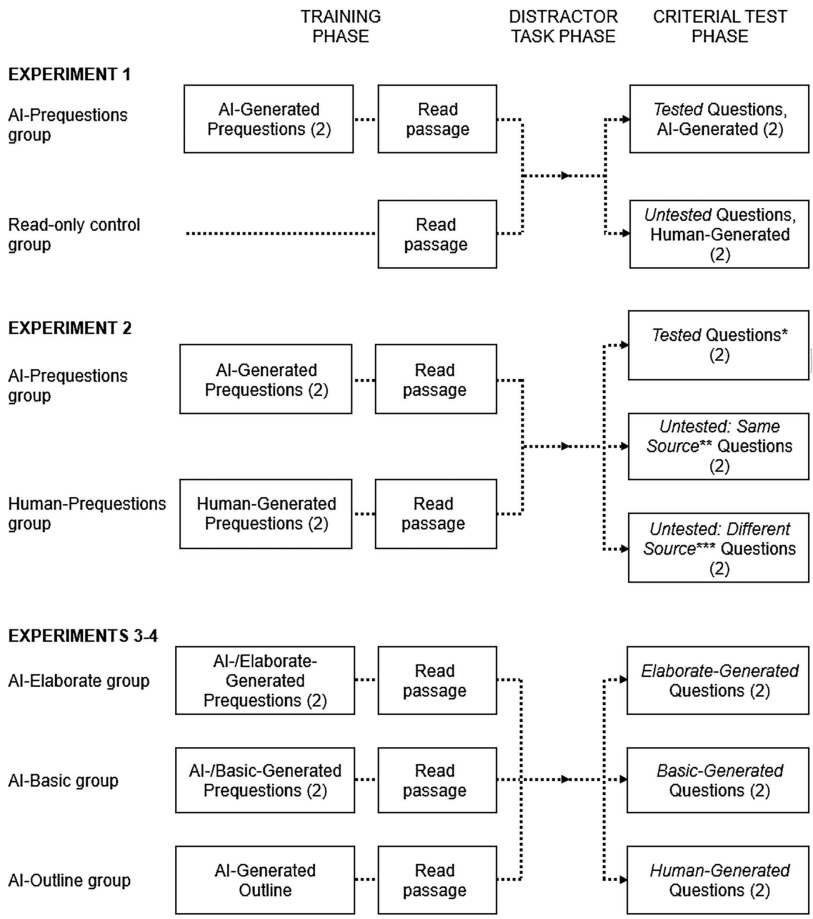
### Scoring

A scoring rubric was developed wherein the correct answer to each question consisted of 2–3 idea units. To attain a perfect score to a given question, a participant's answer had to be entirely accurate with all idea units expressed and described correctly (each idea unit earned 1 point). All scoring for Experiments 1–3 was performed by a single rater who was blind to group assignment; a different rater scored Experiment 4. To assess scoring reliability, 53% of all responses for Experiment 1 were scored by an additional blind

[1] One to two prequestions before a text has precedent (e.g., Carpenter et al., 2018), and participants often struggle to recall more questions (St. Hilaire & Carpenter, 2020). Pilot testing indicated that attention checks, recalling questions, and detailed instructions were necessary to screen out inattentive participants. Although participants were not explicitly warned of an upcoming test, the post-prequestioning instructions could have been interpreted as implying one.

**Figure 1**

*Overview of the Experimental Design and Procedure*




*Note.* Numbers in parentheses indicate the total number of questions of a given type. For Experiment 2, * = identical to the prequestions in the respective groups; ** = different questions from the same source as the prequestions; *** = different questions from a different source as the prequestions.



rater, yielding a reasonable level of interrater agreement (intraclass correlation coefficient = 0.84).

Formal scoring was not performed on attention check data from the AI-prequestion group. Instead, responses were simply checked to verify whether at least one recognizable detail from at least one prequestion was recalled. In this and subsequent experiments, all participants that were not off task according to TaskMaster met that threshold, and hence, attention check data are not discussed further.

## Results

Data and analysis code for all experiments are available on the Open Science Framework and are accessible at https://osf.io/x3enc/ (Pan et al., 2025).

### *Training*

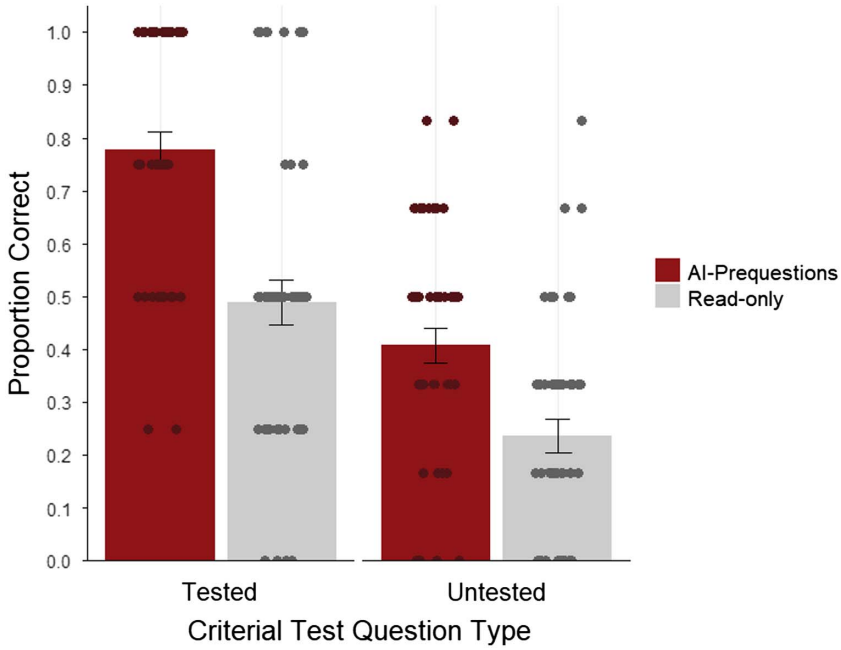
Performance on the prequestions was very low (accuracy of $M$ = 0.056 proportion correct, $SD$ = 0.19), as expected given minimal prior knowledge (performance on prequestions is typically quite low, e.g., Kornell et al., 2009; Richland et al., 2009). Participants in the AI-prequestion group spent, on average, 1.10 min ($SD$ = 1.06) per question. The AI-prequestion group tended to spend longer reading the text passage ($M$ = 7.22 min, $SD$ = 2.47) than the read-only group ($M$ = 5.64 min, $SD$ = 1.08).

### *Criterial Test*

We performed a 2 (training group: AI-prequestion vs. read-only) × 2 (question type: tested vs. untested) mixed-design analysis of variance (ANOVA) on participant-level mean criterial test scores. That analysis revealed a significant main effect of training group, $F(1, 88) = 32.95$, $p < .001$, $\eta_p^2 = .27$; a significant main effect of question type, $F(1, 88) = 101.26$, $p < .0001$, $\eta_p^2 = .54$; and no significant interaction, $F(1, 88) = 3.67$, $p = .059$, $\eta_p^2 = .040$. Inspection of criterial test data in Figure 2 sheds light on the ANOVA results. In the figure there are clear indications of better performance in the AI-prequestion group than in the read-only group, with a substantial prequestioning effect for both tested

**Figure 2**

*Experiment 1 Criterial Test Results*




*Note*. Error bars represent standard error of the mean. See the online article for the color version of this figure.

and untested questions and, moreover, better performance for tested (i.e., AI-generated) than untested (i.e., human-generated) questions.

To further characterize the observed prequestioning effects, we performed pairwise comparisons separately for tested and untested questions. To supplement these comparisons, Bayes factors (using the BayesFactor package in R; Morey & Rouder, 2023) are also reported. A Bayes factor ($BF_{10}$) represents the likelihood ratio of the alternative hypothesis to the null hypothesis. A $BF_{10} > 1$ favors the alternative hypothesis, $BF_{10}$ of 1 suggests equal likelihood, and a $BF_{10} < 1$ favors the null hypothesis (Rouder et al., 2009; Wagenmakers, 2007). For cases wherein the null hypothesis is more likely, the reciprocal $BF_{01}$ is reported for ease of interpretation.

For tested questions, there was a significant difference between the AI-prequestion and read-only groups, $t(88) = 5.29$, $p < .001$, $d = 1.12$, $BF_{10} > 15{,}000$, indicating a robust prequestioning effect. For untested questions, there was also a significant group difference, $t(88) = 3.69$, $p = .0004$, $d = 0.78$, $BF_{10} = 67.35$, again indicating a robust prequestioning effect.

## Experiment 2

Experiment 1 demonstrated that prequestioning with AI-generated questions prior to reading a text passage yields substantial learning benefits relative to only reading. Building on that finding, Experiment 2 examined whether prequestioning with AI-generated questions might be as effective as prequestioning with questions created by human beings and whether the advantage of tested over untested questions observed in Experiment 1 was partially due to the untested questions also coming from a different source (i.e., human beings instead of AI).

### Method

The design, hypotheses, and analysis plan for Experiment 2 were preregistered at https://aspredicted.org/qp76-8jf3.pdf. We hypothesized that criterial test performance would be best for questions targeting directly tested, as opposed to untested, content. As an exploratory question, we also examined potential differences between untested questions from the same source and from a different source.

#### *Design*

Experiment 2 used a 2 × 3 design wherein each participant was randomly assigned to an AI-prequestion group or a human-prequestion group. All participants completed a criterial test that included three question types: (a) tested questions, which were identical to those used during prequestioning in each group; (b) untested same-source questions, which were generated by the same source as the prequestions (ChatGPT for the AI-prequestion group and human beings for the human-prequestion group); and (c) untested different-source questions, which were generated by a different source (human beings for the AI-prequestion group and ChatGPT for the human-prequestion group). The criterial test therefore addressed memory for tested materials and transfer to untested materials along with the impact of question source on transfer performance.

### *Participants*

The target sample size of 128 participants, split evenly between two groups, was determined via an a priori power analysis conducted in G*Power involving a two-tailed, independent-samples *t* test with the assumptions of a medium effect size ($d = 0.5$, smaller than in the preceding experiment), $\alpha = .05$, and 80% power (although we did not hypothesize an advantage of one type of prequestioning over another, we were still interested in being able to detect a reasonably sized difference between groups if it occurred). One hundred sixty-eight participants were recruited via Prolific Academic, each meeting the same qualifying requirements as in Experiment 1 and receiving $6.00. Prior to formal analysis, data from 40 participants were excluded for evidence of off-task browser activity (again using the same criteria as in the prior experiment), yielding a final sample of 128 participants (AI-prequestion group, $n = 64$; human-prequestion group, $n = 64$) which had a mean age of 31.2 years and was 48% female. Thirty-seven percent of these participants were from the United Kingdom, 28% were from the United States, and 35% were from the remaining eligible countries; 11% of participants were Asian, 11% were Black, 8% were mixed, 62% were White, and 9% were from other ethnic groups or declined to provide ethnicity information.

### *Materials, Procedure, and Scoring*

Four AI-generated questions were created in January 2024 using GPT-3.5 and the same prompt as in Experiment 1 (see Table 1). One of the generated questions was identical to an AI-generated prequestion used in Experiment 1 (we observed that ChatGPT commonly generates the same or similar questions in response to the same prompt and source materials). The questions were divided into pairs, with one pair presented to participants in the AI-prequestion group during the training phase and pair presentation counterbalanced over participants. All four questions were presented to the AI-prequestion group on the criterial test (as tested same- and untested same-source questions, respectively), whereas two questions were presented to the human-prequestion group on the criterial test (as untested different-source questions).

The four human-generated questions, sourced from St. Hilaire et al. (2019), included the same questions used in Experiment 1, as well as two additional questions (see Table 1). These questions were divided into pairs, with the presentation of pairs during the training phase counterbalanced in the human-prequestion group. All four questions were presented to the human-prequestion group on the criterial test (as tested same- and untested same-source questions, respectively), whereas two questions were presented to the AI-prequestion group on the criterial test (as untested different-source questions).

The questions presented as untested different-source questions were always the same two questions for the AI-prequestion group and the same two questions for the human-prequestion group. These questions, which are noted with asterisks in Table 1, had minimal overlap in idea units with any of tested or untested same-source questions for any participant. For instance, for the human-prequestion group, the tested and untested same-source questions referred to hydraulic brakes, train versus bicycle brakes, air brakes, and mechanical versus hydraulic brakes, whereas the untested different-source questions referred to components of drum brakes and power brakes, respectively (see Online Supplemental Materials for further discussion).

As overviewed in Figure 1 (middle panel), the procedure was similar to that for Experiment 1, except that participants were assigned to the AI-prequestion or human-prequestion groups, with all participants attempting two prequestions, and the criterial test featured six rather than four questions (shuffled in a manner akin to that for Experiment 1). Scoring occurred in the same fashion as in the prior experiment.

## Results

### *Training*

Performance in the AI-prequestion and human-prequestion groups was very low (accuracy of $M = 0.016$, $SD = 0.088$; $M = 0.063$, $SD = 0.19$, respectively), as expected. Both groups spent similar amounts of time answering the prequestions (AI-prequestion group, $M = 1.22$ min, $SD = 1.26$; human-prequestion group, $M = 1.13$ min, $SD = 0.66$). Reading time appeared to be longer on average in the AI-prequestion group ($M = 7.54$ min, $SD = 3.57$) than in the human-prequestion group ($M = 6.87$ min, $SD = 2.72$).

### *Criterial Test*

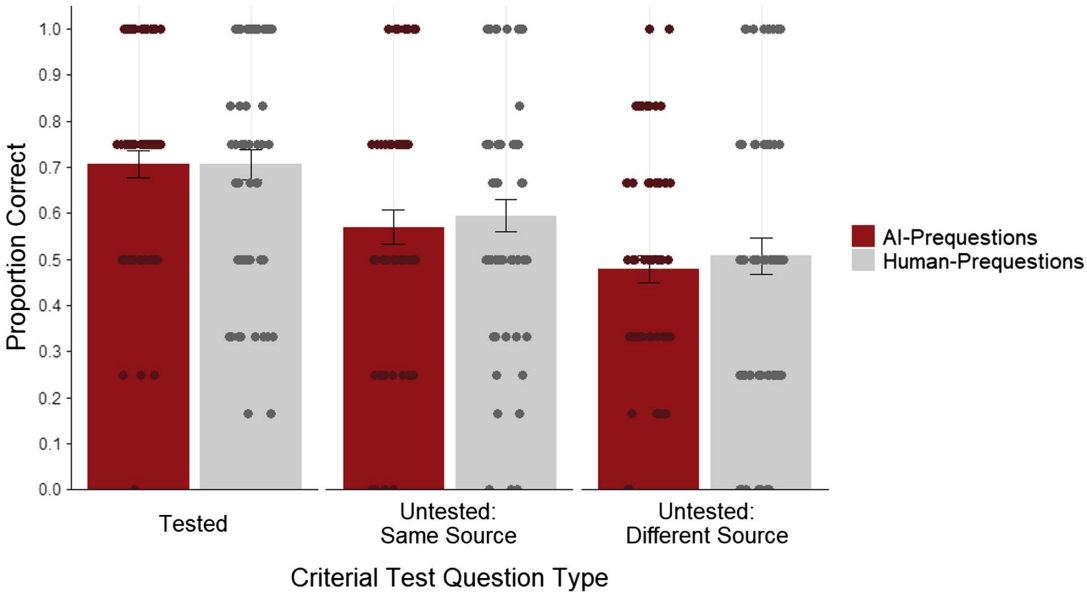
We performed a 2 (training group: AI-prequestion vs. human-prequestion) × 3 (question type: tested vs. untested same-source vs. untested different-source) mixed-design ANOVA on participant-level mean criterial test scores. There was no significant main effect of training group, $F(1, 126) = 0.31$, $p = .58$, $\eta_p^2 = .0024$; a significant main effect of question type, $F(2, 252) = 22.99$, $p < .0001$, $\eta_p^2 = .15$; and no significant interaction, $F(2, 252) = 0.13$, $p = .88$, $\eta_p^2 = .0011$. Inspection of Figure 3 suggests comparable levels of performance for the AI-prequestion and human-prequestion groups and an apparent stepwise decrease in accuracy across tested questions to the two types of untested questions. The latter pattern suggests that as the degree of transfer increases—both in terms of tested versus untested content and the source of the questions—performance decreases.

Although both groups appeared to perform similarly on all question types, we performed pairwise comparisons for each question type in accordance with our preregistered analysis plan. There were no significant differences between the AI-prequestion and human-prequestion groups for any question type, $t$s $\leq 0.58$, $p \geq 0.56$, $d$s $\leq 0.10$, $BF_{01}$s $\geq 4.53$, which corresponds to the ANOVA results.

To explore the main effect of question type, we performed pairwise comparisons on test results collapsed across training groups. There were significant differences between tested and untested same-source questions, $t(127) = 3.89$, $p < .001$, $d = 0.46$, $BF_{10} = 109$, as well as between untested same-source and untested different-source questions, $t(127) = 2.86$, $p = .005$, $d = 0.32$, $BF_{10} = 4.70$. These results reflect the aforementioned stepwise pattern and are consistent with our hypothesis. Specifically, the highest levels of performance occurred with questions identical to those encountered during training, followed by reduced performance when the questions differed, with the lowest performance involving questions generated by a different source than encountered previously. Bayesian evidence, however, was stronger for the

**Figure 3**

*Experiment 2 Criterial Test Results*




*Note*. Error bars represent standard error of the mean. See the online article for the color version of this figure.

former comparison than for the latter one, which is also in line with the observed effect sizes.

## Experiment 3

Building on the benefits of AI-generated prequestioning revealed in the previous experiments, Experiment 3 investigated whether similar benefits might occur with questions not generated using specially designed prompts, under strict controls for time on task, and versus a competitive non-prequestioning condition. Accordingly, this experiment featured a control group that studied an outline before reading, which can benefit text comprehension (Ponce et al., 2023), and compared it against two AI-generated prequestioning groups: one using questions from the same detailed prompt as before and another using questions from a simplified prompt that an untrained student might use. Recent surveys reveal that undergraduate students commonly use generative AI to provide summaries of to-be-learned materials (Balabdaoui et al., 2024), which can include outlines (Vieriu & Petrea, 2025), as well as a prelude to writing assignments (Malik et al., 2023). The outline was also AI generated.

### Method

The experiment design, hypotheses, and analysis plan were preregistered at https://aspredicted.org/hpjn-kgsp.pdf. We hypothesized that both prequestioning groups (i.e., using prequestions developed via a specially designed and a simplified prompt; hereinafter, AI-elaborate and AI-basic groups, respectively) would outperform the outline group (hereinafter, AI-outline group) and that performance would be best in AI-elaborate group. In addition, the design allowed for another exploratory comparison of transfer to different types of untested questions.

#### *Design*

Experiment 3 used a 3 × 3 design wherein each participant was randomly assigned to the AI-elaborate, AI-basic, or AI-outline group. For an overview of the different groups, see Figure 1 (bottom panel). During the training phase, the AI-elaborate group engaged in prequestioning with the same prequestions used in Experiment 2, whereas the AI-basic group engaged in prequestioning with prequestions that were not generated using a specially designed prompt. The AI-outline group did not engage in prequestioning and instead studied an outline.

All groups completed a criterial test that resembled Experiment 2 in featuring three types of questions, both AI and human generated, but in this case the types were (a) elaborate-generated questions, which were developed using the same prompt as used in the preceding experiments; (b) basic-generated questions, which were developed using a simplified prompt; and (c) human-generated questions, which were again drawn from St. Hilaire et al. (2019). As in the prior experiment, this approach addressed the learning of tested and untested materials as well as the impact of question source on transfer performance.

In summary, Experiment 3 featured three groups that differed in training phase characteristics (the AI-elaborate, AI-basic, and AI-outline groups), with all three groups completing a criterial test that featured three question types (elaborate generated, basic generated, and human generated).

#### *Participants*

An a priori power analysis was conducted in G*Power to determine the minimum sample size needed for detecting an advantage of either prequestioning group over the AI-outline group,

with an effect size of $d = 0.5$ or larger, in planned one-tailed pairwise comparisons with $\alpha = .05$ and 80% power (to reiterate, we hypothesized an advantage of prequestioning over outline study). A minimum of 51 participants per group was recommended, and thus, we targeted at least 153 participants, equally split across the three groups. One hundred seventy-four participants were recruited via Prolific Academic, each meeting the same qualifying requirements as in prior experiments and receiving $5.35. Prior to formal analysis, data from 21 participants were excluded for evidence of off-task browser activity, yielding a final sample of 153 participants (AI-elaborate group, $n = 51$; AI-basic group, $n = 51$; and AI-outline group, $n = 51$), which had a mean age of 32.6 years and was 58% female. Forty-five percent of these participants were from the United Kingdom, 33% were from the United States, and 22% were from the remaining eligible countries; 10% of participants were Asian, 11% were Black, 5% were mixed, 71% were White, and 3% were from other ethnic groups or declined to provide ethnicity information.

### *Materials, Procedure, and Scoring*

Materials drawn from Experiment 2 included the four questions that had been generated using a purposefully designed and iteratively refined prompt in GPT-3.5 (i.e., elaborate-generated questions) and four human-generated questions. In addition, four basic-generated questions were newly generated in February 2024 using GPT-3.5 and a simplified, one-line prompt that any learner with no specialized training might use (i.e., "Can you create a series of 2 questions from a passage?"). This approach was indeed much simpler than that employed to create elaborate-generated questions, which might not represent much time or effort savings over having to manually devise practice questions. It should be noted, however, that because no instructions were given to avoid referring directly to the passage as if it had already been read, one of the basic-generated questions produced by GPT-3.5 included such a reference and had to be manually revised. A counterbalancing protocol was implemented to ensure that each participant received two elaborate-generated, two basic-generated, and two human-generated questions on the criterial test. Unlike Experiments 1 and 2, reflecting the increased complexity of having three experimental groups, criterial test question assignments were not preselected and simply counterbalanced across participants. Participants in the AI-elaborate and AI-basic groups were exposed to the same two elaborate-generated or basic-generated questions on the criterial test as they had encountered during the training phase. An AI-generated outline of approximately 350 words in length was also created at the same time using a detailed prompt (see Appendix B).

The procedure for the AI-elaborate and AI-basic groups was akin to that of Experiment 2, except that 80 s was allotted for each of the two prequestions and passage reading time was fixed at 6 min. The AI-outline group underwent a similar procedure, with 160 s allotted for studying the outline and passage reading time fixed at 6 min. Hence, time on task was equivalent across all groups. The instructions for the prequestion groups were identical to those of the previous experiments, while the AI-outline group was instructed to carefully read an outline related to an upcoming text passage as an aid to understanding the text. Each group completed a six-question criterial test containing two elaborate-generated, two basic-generated, and two human-generated questions. All other procedures, including counterbalancing of prequestions, shuffling of criterial test questions, and scoring of participants' responses, resembled that of the prior experiment.

## Results

### *Training*

Participants in the AI-elaborate and AI-basic groups were rarely able to answer the prequestions correctly (accuracy of $M = 0.093$, $SD = 0.29$; $M = 0.039$, $SD = 0.20$, respectively), as expected.

### *Criterial Test*

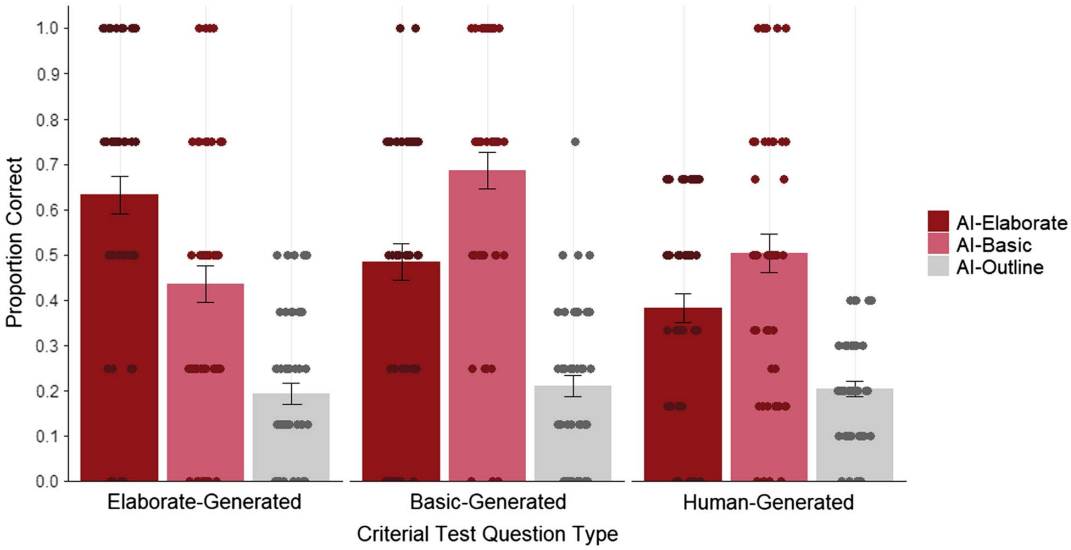
We performed a 3 (training group: AI-elaborate vs. AI-basic vs. AI-outline) × 3 (question type: elaborate-generated vs. basic-generated vs. human-generated) mixed-design ANOVA on participant-level mean criterial test scores. That analysis revealed a significant main effect of training group, $F(2, 150) = 56.96$, $p < .0001$, $\eta_p^2 = .43$; a significant main effect of question type, $F(2, 300) = 8.03$, $p < .001$, $\eta_p^2 = .051$; and a significant interaction, $F(4, 300) = 14.11$, $p < .0001$, $\eta_p^2 = .016$. These patterns are reflected in Figure 4, in which the advantages of the prequestioning groups over the AI-outline group and variations in performance across question types are evident. The interaction appears to reflect, at least in part, different performance between prequestioning groups across question types—including advantages for the AI-elaborate group on elaborate-generated questions and the AI-basic group on basic-generated questions. Moreover, although the AI-elaborate group appeared to exhibit a stepwise decrease in accuracy from tested questions to the two types of untested questions, consistent with Experiment 2, that pattern appeared to be minimal or absent in the other groups. It should be emphasized, however, that a decrease in similarity across question types in Experiment 2 is less clear-cut than that for Experiment 3 (in which the questions are identical to those used during training in the case of elaborate-generated and basic-generated questions for the AI-elaborate and AI-basic groups, but otherwise different for the other two types of questions).

To further characterize the observed prequestioning effects, we compared performance in each prequestioning group versus the AI-outline group (uniquely among the pairwise comparisons reported for this experiment, these tests were one-tailed given our preregistered hypothesis that prequestioning would be more effective than studying an outline). The AI-elaborate group outperformed the AI-outline group on all question types, $t$s $\geq 4.99$, $p < .0001$, $d$s $\geq 0.99$, $BF_{10} \geq 6{,}000$. The AI-basic group also outperformed the AI-outline group on all question types, $t$s $\geq 5.21$, $p < .0001$, $d$s $\geq 1.03$, $BF_{10} \geq 14{,}000$.

To explore the efficacy of the questions generated by different types of prompts, we examined criterial test performance on human-generated questions. Those questions were the only questions on the criterial test that were entirely new to both groups and, hence, should be equally "novel" in both cases. Performance on human-generated questions was significantly better in the AI-basic group than in the AI-elaborate group, $t(100) = 2.29$, $p = .024$, $d = 0.45$, $BF_{10} = 2.08$. Even though the Bayesian evidence was not strong, these patterns falsify the hypothesis that prequestions made using a detailed prompt are generally more

**Figure 4**

*Experiment 3 Criterial Test Results*




*Note*. Error bars represent standard error of the mean. See the online article for the color version of this figure.



effective, as the advantage is in favor of the AI-basic rather than the AI-elaborate prompt.

Finally, to explore the effect of question type during the criterial test, we performed pairwise comparisons on data collapsed across training group. Performance on elaborate-generated questions was not significantly different than performance on basic-generated questions, $t(152) = 1.47$, $p = .14$, $d = 0.12$, $BF_{01} = 3.90$. Relative to human-generated questions, however, performance on elaborated-generated questions was significantly higher, $t(152) = 2.15$, $p = .033$, $d = 0.17$, $BF_{10} = 0.84$, although Bayesian evidence weakly favors the null hypothesis. Performance on basic-generated questions was also significantly higher than performance on human-generated questions, $t(152) = 3.84$, $p < .001$, $d = 0.31$, $BF_{10} = 91.33$.

## Experiment 4

Experiment 4 aimed to conceptually replicate the findings of Experiment 3 while addressing two observations about its method. First, the human-generated questions all involved making comparisons, which might impact understanding and transfer (Rittle-Johnson & Star, 2011), whereas most of the AI-generated questions did not. Second, the complexity or quality of the outline might have detracted from its efficacy (Colliot & Jamet, 2018). Hence, this experiment only used questions that involved making comparisons and featured a simplified outline in the AI-Outline group. Additionally, whereas the prior experiments used GPT-3.5, this experiment involved the recently released GPT-4o.

### Method

The experiment design, hypotheses, and analysis plan were preregistered at https://aspredicted.org/4bzz-jr93.pdf. We hypothesized that there would be a prequestioning effect relative to the AI-outline condition for all criterial test question types.

#### *Design*

Experiment 4 used a 3 × 3 design identical to that used in Experiment 3.

#### *Participants*

The target sample size was identical to that of Experiment 3. One-hundred eighty-three participants were recruited via Prolific Academic using the same minimum requirements and compensated identically as in the prior experiment. Prior to formal analysis, data from 21 participants were excluded due to off-task browser activity, and data from five participants were excluded for noncompliance with instructions, resulting in a final sample of 157 participants (AI-elaborate group, $n = 53$; AI-basic group, $n = 51$; and AI-outline group, $n = 53$) which had a mean age of 32.9 years and was 60% female (4% nonbinary). Thirty-eight percent of these participants were from the United Kingdom, 37% were from the United States, and 25% were from the remaining eligible countries; 11% of participants were Asian, 6% were Black, 8% were mixed, 71% were White, and 4% were from other ethnic groups or declined to provide ethnicity information.

#### *Materials, Procedure, and Scoring*

Experiment 4 differed from Experiment 3 in all materials except for the human-generated questions. New sets of four questions each were generated in August 2024 using an AI-elaborate and an AI-basic prompt (see Table 1) and the latest free version of ChatGPT, GPT-4o, which was released in May 2024. These prompts were modified from their original versions to ensure that all

resulting questions involved making comparisons (see Appendix A). The AI-outline prompt was also modified to include a 160-word limit and directions to use simpler vocabulary. That prompt was used with GPT-4o to generate a new outline concurrently as the new questions (see Appendix B).

Unlike in Experiment 3, where the AI-elaborate and AI-basic questions did not exhibit strong overlap in idea units, such overlap was present across those question types—and with the human-generated questions—in this experiment (see Online Supplemental Materials for further details). Such overlap was probably unavoidable given the limited number of comparisons between two brake types that are possible given the passage content. Moreover, despite the AI-elaborate prompt retaining instructions to not exceed 15 words per question, GPT-4o repeatedly violated that dictum (an apparent consequence of the fact that LLMs, as of this writing, do not know in advance how long a given response might be). In the interests of maintaining fidelity to GPT-4o's output, we used the longer questions regardless.

The procedure was essentially identical to that of Experiment 3. Scoring involved the same blind scoring procedure as in the prior experiments but was completed by a different rater. To ensure consistency, the rater first rescored 30% of the criterial test responses from the prior experiment, achieving a good rate of interrater agreement (intraclass correlation coefficient of 0.87), before applying the same approach to scoring the criterial test data for Experiment 4.

## Results

### *Training*

Participants in the AI-elaborate and AI-basic groups were rarely able to answer the prequestions correctly (accuracy of $M = 0.14$, $SD = 0.24$; $M = 0.15$, $SD = 0.23$, respectively).

### *Criterial Test*

As with Experiment 3, we performed a 3 (training group: AI-elaborate vs. AI-basic vs. AI-outline) × 3 (question type: elaborate-generated vs. basic-generated vs. human-generated) mixed-design ANOVA on participant-level mean criterial test scores. That analysis revealed a significant main effect of training group, $F(2, 154) = 19.28$, $p < .0001$, $\eta_p^2 = .20$; a significant main effect of question type, $F(2, 308) = 16.93$, $p < .0001$, $\eta_p^2 = .10$; and a significant interaction, $F(4, 308) = 3.52$, $p = .0079$, $\eta_p^2 = .044$. Those patterns are reflected in Figure 5, in which there are indications of prequestioning effects across all question types, replicating the overall results of Experiment 3. Unlike that experiment, however, the highest overall performance and a larger magnitude pre-questioning effect were evident for elaborate-generated questions.

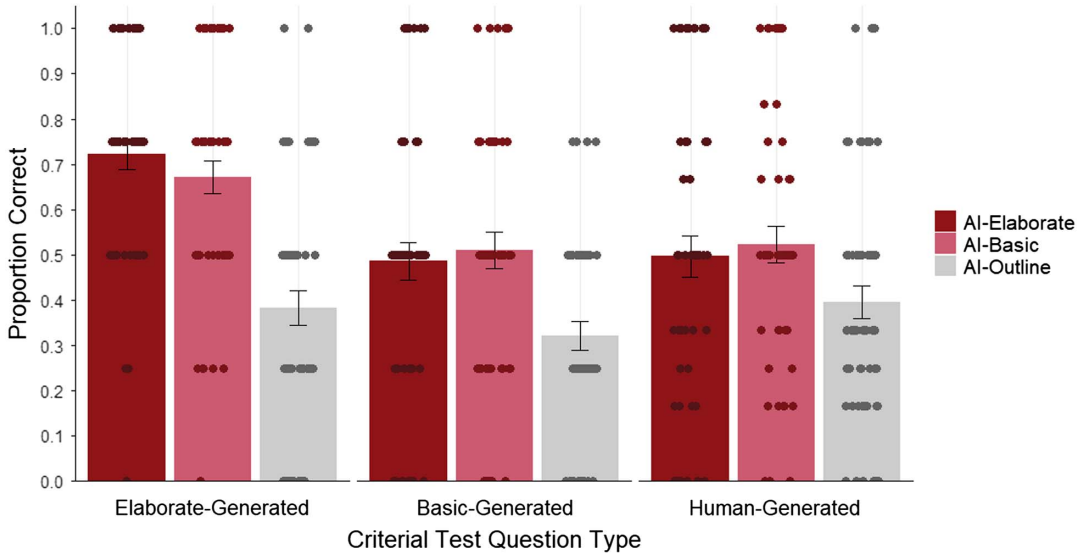
To determine the extent of the observed prequestioning effects, we again compared performance in each prequestioning group versus the AI-outline group using one-tailed tests. The AI-elaborate group outperformed the AI-outline group on all question types, $t$s ≥ 1.73, $p \leq .044$, $d$s ≥ 0.34, $BF_{10} \geq 0.77$. The AI-basic group also outperformed the AI-outline group on all question types, $t$s ≥ 2.35, $p \leq .010$, $d$s ≥ 0.46, $BF_{10} \geq 2.35$. Overall, the prequestioning effect results largely replicate that of Experiment 3, although larger effects were observed for elaborate-generated and basic-generated questions versus human-generated questions (see Figure 5).

We also compared performance among the two prequestioning groups on human-generated questions only. Unlike in Experiment 3, there was no performance advantage for the AI-basic group, $t(102) = 0.42$, $p = .67$, $d = 0.084$, $BF_{01} = 0.67$. Rather, both groups did not perform significantly different from one another, with moderate Bayesian evidence for the null hypothesis.

Finally, we explored the effect of question type by performing pairwise comparisons on data collapsed across training groups.

**Figure 5**

*Experiment 4 Criterial Test Results*




*Note*. Error bars represent standard error of the mean. See the online article for the color version of this figure.

Performance on elaborate-generated questions was significantly higher than performance on basic-generated questions, $t(156) = 6.04$, $p < .0001$, $d = 0.48$, $BF_{10} > 880{,}000$, and on human-generated questions, $t(156) = 4.22$, $p < .0001$, $d = 0.34$, $BF_{10} > 350.50$. Performance on basic-generated and human-generated questions, however, did not significantly differ, $t(156) = 1.10$, $p = .27$, $d = 0.09$, $BF_{01} = 6.20$.

## Readability and Difficulty of Questions and Outlines

Readability measures for all question types are presented in Table 2. All questions required at least an eighth- to ninth-grade reading level and a decade or more of formal education to understand, which is higher than the readability metrics for the "Brakes" passage itself. Elaborate-generated questions had a higher average Flesch–Kincaid grade level and Gunning Fog Index than basic-generated questions, suggesting that the more detailed prompt yielded questions of greater reading difficulty. Possibly due to prompt-imposed word limits, the average word count of ChatGPT-produced questions was lower than human-generated questions in most cases, excepting AI-basic comparison questions.

Readability measures for the outlines used in Experiments 3 and 4 were also obtained. As presented in Table 2, the outline used in Experiment 3 had higher reading difficulty and grade-level ratings than the outline used in Experiment 4, with the former being more appropriate for college-level readers and the latter being suitable for middle- or secondary school-level readers.

The first author and a coauthor of this article independently classified each question used in the entire study according to the cognitive domain levels of Bloom's taxonomy. The first author reviewed the discrepant scores, of which there were three cases, and final classifications were made after discussion between raters. The results are shown in Table 2. The noncomparison elaborate-generated questions ranged from "remember" to "analyze" but were largely at the "apply" level, whereas the basic-generated questions were split between the "remember" and "understand" levels. At least three quarters of the comparison-focused elaborate-generated and basic-generated questions, however, were at the "analyze" level, which may engage more difficult or advanced mental processes. The human-generated questions were also predominantly at the "analyze" level.

## Discussion

The present experiments provide compelling evidence that AI-generated practice questions can be used to generate substantial prequestioning effects. Engaging in AI-generated prequestioning enhanced performance on criterial test questions that were both identical and nonidentical to those presented prior to reading a text passage (Experiments 1, 3, and 4), plus yielded comparable performance as that following human-generated prequestioning (Experiment 2). Together, these results establish the viability of AI-generated questions as a potentially effective way to implement prequestioning in a widely accessible and easily utilized manner. With these results in mind, we next revisit the research questions articulated at the outset of this article.

## Effects of AI-Generated Prequestioning for Identical and Nonidentical Test Questions

Does AI-generated prequestioning produce a prequestioning effect for directly tested information, and moreover, do any benefits of such prequestioning extend to criterial test questions that differ from the practice questions in some way? The answer to both questions is yes. Experiment 1 served as a proof of concept, showing that AI-generated prequestioning, when used prior to reading an educational text, yields better learning than reading the text alone. Yet without controls for time on task and a competitive comparison group, it did not rule out the possibility that the learning gains stemmed from extra practice and exposure to materials rather than prequestioning per se. Experiments 3 and 4 addressed those issues with more robust controls and found that AI-generated prequestioning still yielded better learning outcomes than learning approaches without prequestions. Notably, the effect size magnitudes in Experiments 1 and 3 were comparable ($d \approx 1.0$), although they were reduced in Experiment 4.

Together, these results contribute to the growing evidence that prequestioning engages cognitive processes that uniquely enhance learning. Although the exact nature of these processes remains to be clarified, they likely arise from two sources: (a) the act of making an initial guess, which may create a distinct episodic memory, generate an error signal, or increase attentional focus or curiosity, and (b) the subsequent discovery and processing of the correct answer, which is likely to be more thorough and/or better integrated with prior knowledge or experiences than after non-prequestioning methods (for further discussion, see Pan & Carpenter, 2023).

Not only did prequestioning enhance performance on criterial test questions that were identical to those used during initial practice, but it also benefited performance on other test questions, albeit to a lesser extent in most cases. It should be noted, however, that all criterial test questions addressed the same overarching topic—major

**Table 2**
*Question Readability and Cognitive Level*

| | Readability metric, *M* (*SD*) | | | Bloom's taxonomy cognitive level, % of total | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Type | Word count | Flesch–Kincaid grade level | Gunning Fog index | Remember | Understand | Apply | Analyze | Evaluate | Create |
| Human generated | 19.0 (8.7) | 9.0 (3.9) | 12.3 (5.1) | | | | 75 | 25 | |
| AI-/elaborate generated | 11.8 (2.2) | 10.2 (4.8) | 11.7 (5.9) | 20 | | 60 | 20 | | |
| AI-/elaborate generated (comparison) | 15.3 (4.2) | 12.0 (1.6) | 13.7 (2.3) | | | | 75 | 25 | |
| AI-/basic generated | 13.0 (0.8) | 8.7 (5.0) | 10.4 (6.4) | 50 | 50 | | | | |
| AI-/basic generated (comparison) | 22.0 (0.8) | 11.1 (3.3) | 12.7 (4.6) | | | | 100 | | |

*Note*. AI = artificial intelligence.

types of brakes—and there were instances wherein questions overlapped in idea units (particularly in Experiment 4, which uniquely had multiple cases of overlapping idea units across different question types), which raises the possibility that the observed prequestioning effects may reflect similar or overlapping memories of learned materials. In Experiments 1–3, such overlap was limited. In Experiment 4, however, the similar performance of the AI-elaborate and AI-basic groups can be partly attributed to overlap in idea units across elaborate-generated, basic-generated, and human-generated test questions. Still, such overlap could plausibly occur in real-world scenarios, for instance, when an instructor provides practice questions and avoids reusing the exact same questions on a subsequent exam covering the same topic. Moreover, the observed successful transfer—arguably near in distance (Barnett & Ceci, 2002; Thorndike & Woodworth, 1901; see also Nguyen & McDaniel, 2015)—is notable and adds weight to the possibility that integrative prequestions promote transferable learning (St. Hilaire et al., 2019).

Experiments 3–4 featured one of the first direct comparisons of prequestioning to studying outlines. Both approaches can be viewed as the use of an advance organizer—that is, a learning tool that introduces materials prior to formal learning (Hartley & Davies, 1976). Crucially, prequestioning consistently outperformed studying outlines, which suggests that simply previewing a text passage may not be as potent as making guesses about it beforehand. That result is consistent with prior findings of a prequestioning advantage over such advance organizers as reading learning objectives (Sana et al., 2020), as well as the possibility that specific cognitive mechanisms are triggered by prequestioning—either from the act of guessing or of processing the correct answer afterward—and not other advance organizers (Kornell & Vaughn, 2016; Mera et al., 2021; Pan & Carpenter, 2023). It should be noted, however, that the outlines did not specify comparisons, and moreover, in the present experiments, the prequestions may have been advantaged by their greater similarity with the criterial test (i.e., transfer-appropriate processing). Hence, there may be other circumstances wherein studying an outline is more competitive with prequestioning.

## Emerging Patterns in AI-Enabled Question Generation

How did AI-generated and human-generated prequestioning compare? According to Experiment 2, very favorably with one another. Hence, learners need not necessarily rely on human-generated questions to engage in effective prequestioning. GPT- and human-generated questions were also roughly comparable in reading difficulty. Relatedly, our analysis of GPT-3.5- and GPT-4o-generated questions revealed that a greater proportion of questions addressed the basic cognitive levels of Bloom's taxonomy, in line with Singh et al. (2023). The human-generated questions from St. Hilaire et al. (2019), on the other hand, tended to target higher cognitive levels. Although those patterns imply that AI-generated questions might be unable to promote higher order learning outcomes, the AI-prequestioning groups across experiments were able to perform well on higher order human-generated questions, at least relative to groups that did not practice with AI-generated prequestions (Experiment 2) or prequestions at all (Experiments 3–4). Moreover, when prompts specifying comparisons were included, a larger proportion of the AI-generated questions achieved the higher "analyze" level.

The results of Experiments 3 and 4 further suggest that iterative prompt engineering to generate practice questions (Indran et al., 2024; Lee et al., 2024) may not be essential. In fact, no reliable advantage was found for prequestioning with questions generated using elaborate versus basic prompts. Hence, ChatGPT may have advanced to a point wherein specialized prompts are unnecessary to produce effective practice questions and other learning materials (Wang et al., 2024). However, there were indications that specific prompt language (e.g., to generate comparison questions or constrain complexity) can still matter. Further, a viable alternative given the current capabilities of generative AI may be for learners to use prompts that have been validated across different texts to generate practice questions (see also Zamfirescu-Pereira et al., 2023).

## Interpretative Considerations and Future Research Directions

A limitation of the present experiments is that the criterial test was conducted after a short retention interval. Although the prequestioning effect may persist and even grow in magnitude with longer retention intervals (Kliegl et al., 2024), the durability of AI-generated prequestioning effects has yet to be thoroughly tested. Additionally, the present results are specific to GPT-3.5/GPT-4o and the specific prompts that were used. Future studies might generate more diverse question types from a broader range of texts addressing different domains than that of brakes (e.g., more abstract concepts and/or other subjects), along with practice and test questions that are even more differentiated from one another. One possibility is that prequestions can be engineered to trigger more abstract mental models as opposed to focusing attention on isolated details. The accuracy of AI-generated questions can also be further addressed; although no errors were observed among the questions used in the present experiments, ChatGPT can "hallucinate," confabulating sources and misinterpreting information (e.g., Kıyak & Emekli, 2024), which is clearly undesirable. Learners' responses to prequestions are often ungraded (and are likely often incorrect), but pairing prequestions with accurate and reliable source materials—from which the correct answers can be learned—is evidently crucial.

Future studies could also address using AI-generated practice questions for other forms of practice testing (Pan et al., 2024), such as retrieval practice (Roediger & Butler, 2011) or test-potentiated new learning (interpolated practice questions through a text passage; J. C. K. Chan et al., 2018). The present experiments were conducted on a crowdsourcing platform and with participants who are older than typical students, causing us to employ a limited number of prequestions, strict attention checks, and specific instructions, which may or may not be feasible in an offline setting (note: all participants received attention checks, with similar exclusion rates across groups). We also did not inform participants of an upcoming criterial test (but some instructions may have been interpreted as implying it). Although we suspect that the present findings will replicate in different settings, future research is needed to test that assumption. Arguably the closest real-world scenario to that addressed in the present study is that of an instructor providing AI-generated prequestions to students; whether comparable results would be obtained if students directly generate the prequestions for themselves is still to be determined.

## Pedagogical Implications

The present study reveals that learners can use generative AI to produce practice questions that are effective for prequestioning in educationally relevant circumstances. Thus, the lack of practice questions no longer appears to be an impediment to using this promising learning method. Instead, it appears that students and instructors can provide ChatGPT (and likely other advanced LLM-based chatbots) with text-based materials and receive a steady supply of practice questions in return. Further, the finding that prequestions generated with a basic prompt were as effective as those generated with an elaborate prompt suggests that specialized training in prompt generation may be unnecessary. By contrast, analyses of the cognitive level of ChatGPT-generated questions suggest that students and instructors may need to explicitly specify higher level questions if those are desired. By providing unparalleled accessibility and ease in obtaining practice questions, generative AI has the potential to transform how learners use such questions to engage in prequestioning and other forms of test-enhanced learning.

## References


Anderson, I., & Sheikh, A. (2018). Neurosurgery self-assessment: Questions and answers. *British Journal of Neurosurgery*, *32*(5), Article 582. https://doi.org/10.1080/02688697.2017.1409876

Anderson, L. W. (Ed.). (2009). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives* (Abridged ed.). Longman.

Apostolopoulos, I. D., Tzani, M., & Aznaouridis, S. I. (2023). *ChatGPT: Ascertaining the self-evident*. The use of AI in generating human knowledge. https://doi.org/10.48550/arXiv.2308.06373

Arango-Ibanez, J. P., Posso-Nuñez, J. A., Díaz-Solórzano, J. P., & Cruz-Suárez, G. (2024). Evidence-based learning strategies in medicine using AI. *JMIR Medical Education*, *10*(1), Article e54507. https://doi.org/10.2196/54507

Balabdaoui, F., Dittmann-Domenichini, N., Grosse, H., Schlienger, C., & Kortemeyer, G. (2024). A survey on students' use of AI at a technical university. *Discover Education*, *3*(1), Article 51. https://doi.org/10.1007/s44217-024-00136-4

Barnett, S. M., & Ceci, S. J. (2002). When and where do we apply what we learn? A taxonomy for far transfer. *Psychological Bulletin*, *128*(4), 612–637. https://doi.org/10.1037/0033-2909.128.4.612

Butler, A. C. (2010). Repeated testing produces superior transfer of learning relative to repeated studying. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(5), 1118–1133. https://doi.org/10.1037/a0019902

Carpenter, S. K., Rahman, S., & Perkins, K. (2018). The effects of prequestions on classroom learning. *Journal of Experimental Psychology: Applied*, *24*(1), 34–42. https://doi.org/10.1037/xap0000145

Carpenter, S. K., & Toftness, A. R. (2017). The effect of prequestions on learning from video presentations. *Journal of Applied Research in Memory and Cognition*, *6*(1), 104–109. https://doi.org/10.1016/j.jarmac.2016.07.014

Chan, J. C. K., Meissner, C. A., & Davis, S. D. (2018). Retrieval potentiates new learning: A theoretical and meta-analytic review. *Psychological Bulletin*, *144*(11), 1111–1146. https://doi.org/10.1037/bul0000166

Chan, W., An, A., & Davoudi, H. (2023). A case study on ChatGPT question generation. *2023 IEEE International Conference on Big Data (BigData)* (pp. 1647–1656). https://doi.org/10.1109/BigData59044.2023.10386520

Colliot, T., & Jamet, É. (2018). How does adding versus self-generating a hierarchical outline while learning from a multimedia document influence students' performances? *Computers in Human Behavior*, *80*, 354–361. https://doi.org/10.1016/j.chb.2017.11.037

Dai, Y., Liu, A., & Lim, C. P. (2023). Reconceptualizing ChatGPT and generative AI as a student-driven innovation in higher education. *Procedia CIRP*, *119*, 84–90. https://doi.org/10.1016/j.procir.2023.05.002

Domingo, C., Kleber, J., Miksa, V., & Blickensderfer, E. (2022). Charting a path through the storm: General aviation weather training recommendations. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *66*(1), 1824–1828. https://doi.org/10.1177/1071181322661102

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. https://doi.org/10.3758/BF03193146

Hartley, J., & Davies, I. (1976). Preinstructional strategies: The role of pretests, behavioral objectives, overviews and advance organizers. *Review of Educational Research*, *46*(2), 239–265. https://doi.org/10.3102/00346543046002239

Hausman, H., & Rhodes, M. G. (2018). When pretesting fails to enhance learning concepts from reading texts. *Journal of Experimental Psychology: Applied*, *24*(3), 331–346. https://doi.org/10.1037/xap0000160

Imundo, M. N., Watanabe, M., Potter, A. H., Gong, J., Arner, T., & McNamara, D. S. (2024). Expert thinking with generative chatbots. *Journal of Applied Research in Memory and Cognition*, *13*(4), 465–484. https://doi.org/10.1037/mac0000199

Indran, I. R., Paranthaman, P., Gupta, N., & Mustafa, N. (2024). Twelve tips to leverage AI for efficient and effective medical question generation: A guide for educators using Chat GPT. *Medical Teacher*, *46*(8), 1021–1026. https://doi.org/10.1080/0142159X.2023.2294703

Kasneci, E., Sessler, K., Fischer, F., Gasser, U., & Groh, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, *103*, Article 102274. https://doi.org/10.1016/j.lindif.2023.102274

Kliegl, O., Bartl, J., & Bäuml, K.-H. T. (2024). The pretesting effect comes to full fruition after prolonged retention interval. *Journal of Applied Research in Memory and Cognition*, *13*(1), Article 63. https://doi.org/10.1037/mac0000085

Kornell, N., Hays, M. J., & Bjork, R. A. (2009). Unsuccessful retrieval attempts enhance subsequent learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(4), 989–998. https://doi.org/10.1037/a0015729

Kornell, N., & Vaughn, K. E. (2016). How retrieval attempts affect learning: A review and synthesis. In B. H. Ross (Ed.), *Psychology of learning and motivation* (Vol. 65, pp. 183–215). Academic Press. https://doi.org/10.1016/bs.plm.2016.03.003

Kıyak, Y. S., & Emekli, E. (2024). ChatGPT prompts for generating multiple-choice questions in medical education and evidence on their validity: A literature review. *Postgraduate Medical Journal*, *100*(1189), 858–865. https://doi.org/10.1093/postmj/qgae065

Lee, U., Jung, H., Jeon, Y., Sohn, Y., Hwang, W., Moon, J., & Kim, H. (2024). Few-shot is enough: Exploring ChatGPT prompt engineering method for automatic question generation in english education. *Education and Information Technologies*, *29*(9), 11483–11515. https://doi.org/10.1007/s10639-023-12249-8

Malik, A. R., Pratiwi, Y., Andajani, K., Numertayasa, I. W., Suharti, S., Darwis, A., & Marzuki. (2023). Exploring artificial intelligence in academic essay: Higher education student's perspective. *International Journal of Educational Research Open*, *5*, Article 100296. https://doi.org/10.1016/j.ijedro.2023.100296

Mayer, R. E., & Gallini, J. K. (1990). When is an illustration worth ten thousand words? *Journal of Educational Psychology*, *82*(4), 715–726. https://doi.org/10.1037/0022-0663.82.4.715

Mera, Y., Rodriguez, Gabriel, & Marin-Garcia, E. (2021). Unraveling the benefits of experiencing errors during learning: Definition, modulating factors, and explanatory theories. *Psychonomic Bulletin & Review*, *29*(3), 753–765. https://doi.org/10.3758/s13423-021-02022-8

Metcalfe, J. (2017). Learning from errors. *Annual Review of Psychology*, *68*(1), 465–489. https://doi.org/10.1146/annurev-psych-010416-044022

Morey, R., & Rouder, J. (2023). *BayesFactor: Computation of Bayes factors for common designs*. R package Version 0.9.12-4.6. https://richarddmorey.r-universe.dev/BayesFactor/citation

Motz, B., Chinni, A., Leeuw, J., Jankowski, H., Aggarwal, A., Amato, M., Berlin, K., Britten, K., Brown, A., Cerchiaro, M., Evans, N., Findley, A., Gorman, R., Gregg, K., Hansen, K., Hullinger, R., Larkin, P., Lion, M., Long, R., … Fyfe, E. (2024). *ManyClasses 2: The effects of prequestions on media interactions and learning*. https://doi.org/10.31234/osf.io/3xbma

Nguyen, K., & McDaniel, M. A. (2015). Using quizzing to assist student learning in the classroom: The good, the bad, and the ugly. *Teaching of Psychology*, *42*(1), 87–92. https://doi.org/10.1177/0098628314562685

Pan, S. C., & Carpenter, S. K. (2023). Prequestioning and pretesting effects: A review of empirical research, theoretical perspectives, and implications for educational practice. *Educational Psychology Review*, *35*(4), Article 97. https://doi.org/10.1007/s10648-023-09814-5

Pan, S. C., Dunlosky, J., Xu, K. M., & Ouwehand, K. (2024). Emerging and future directions in test-enhanced learning research. *Educational Psychology Review*, *36*(1), Article 20. https://doi.org/10.1007/s10648-024-09857-2

Pan, S. C., & Rickard, T. C. (2018). Transfer of test-enhanced learning: Meta-analytic review and synthesis. *Psychological Bulletin*, *144*(7), 710–756. https://doi.org/10.1037/bul0000151

Pan, S. C., & Sana, F. (2021). Pretesting versus posttesting: Comparing the pedagogical benefits of errorful generation and retrieval practice. *Journal of Experimental Psychology: Applied*, *27*(2), 237–257. https://doi.org/10.1037/xap0000345

Pan, S. C., Sana, F., Samani, J., Cooke, J., & Kim, J. A. (2020). Learning from errors: Students' and instructors' practices, attitudes, and beliefs. *Memory*, *28*(9), 1105–1122. https://doi.org/10.1080/09658211.2020.1815790

Pan, S. C., Schweppe, J., Wenzel, N., Teo, A. Z. J., & Indrajaya, A. (2025). *Artificial intelligence generated prequestions for enhancing memory and text comprehension*. https://osf.io/x3enc

Pan, S. C., Zung, I., Imundo, M. N., Zhang, X., & Qiu, Y. (2023). User-generated digital flashcards yield better learning than premade flashcards. *Journal of Applied Research in Memory and Cognition*, *12*(4), Article 574. https://doi.org/10.1037/mac0000083

Permut, S., Fisher, M., & Oppenheimer, D. M. (2019). TaskMaster: A tool for determining when subjects are on task. *Advances in Methods and Practices in Psychological Science*, *2*(2), 188–196. https://doi.org/10.1177/2515245919838479

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

Richland, L. E., Kornell, N., & Kao, L. S. (2009). The pretesting effect: Do unsuccessful retrieval attempts enhance learning? *Journal of Experimental Psychology: Applied*, *15*(3), 243–257. https://doi.org/10.1037/a0016496

Rittle-Johnson, B., & Star, J. R. (2011). Chapter seven—The power of comparison in learning and instruction: Learning outcomes supported by different types of comparisons. In J. P. Mestre & B. H. Ross (Eds.), *Psychology of learning and motivation* (Vol. 55, pp. 199–225). Academic Press. https://doi.org/10.1016/B978-0-12-387691-1.00007-7

Roediger, H. L., III, & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences*, *15*(1), 20–27. https://doi.org/10.1016/j.tics.2010.09.003

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian *t* tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*(2), 225–237. https://doi.org/10.3758/PBR.16.2.225

Sana, F., & Carpenter, S. K. (2023). Broader benefits of the pretesting effect: Placement matters. *Psychonomic Bulletin & Review*, *30*(5), 1908–1916. https://doi.org/10.3758/s13423-023-02274-6

Sana, F., Forrin, N. D., Sharma, M., Dubljevic, T., Ho, P., Jalil, E., & Kim, J. A. (2020). Optimizing the efficacy of learning objectives through pretests. *CBE Life Sciences Education*, *19*(3), Article ar43. https://doi.org/10.1187/cbe.19-11-0257

Singh, M., Patvardhan, C., & Lakshmi, C. V. (2023). Does ChatGPT spell the end of automatic question generation research? *2023 IEEE International Conference on Computer Vision and Machine Intelligence (CVMI)* (pp. 1–6). https://doi.org/10.1109/CVMI59935.2023.10464618

St. Hilaire, K. J., & Carpenter, S. K. (2020). Prequestions enhance learning, but only when they are remembered. *Journal of Experimental Psychology: Applied*, *26*(4), 705–716. https://doi.org/10.1037/xap0000296

St. Hilaire, K. J., Carpenter, S. K., & Jennings, J. M. (2019). Using pre-questions to enhance learning from reading passages: The roles of question type and structure building ability. *Memory*, *27*(9), 1204–1213. https://doi.org/10.1080/09658211.2019.1641209

St Hilaire, K. J., Chan, J. C. K., & Ahn, D. (2024). Guessing as a learning intervention: A meta-analytic review of the prequestion effect. *Psychonomic Bulletin & Review*, *31*(2), 411–441. https://doi.org/10.3758/s13423-023-02353-8

Thobroni, M., Zulaeha, I., Mardikantoro, H. B., & Yuniawan, T. (2022). Enrichement books in Indonesia. *ISET: International Conference on Science, Education and Technology* (1018–1026).

Thorndike, E. L., & Woodworth, R. S. (1901). The influence of improvement in one mental function upon the efficiency of other functions. II. The estimation of magnitudes. *Psychological Review*, *8*(4), 384–395. https://doi.org/10.1037/h0071280

Vieriu, A. M., & Petrea, G. (2025). The impact of artificial intelligence (AI) on students' academic development. *Education Sciences*, *15*(3), Article 343. https://doi.org/10.3390/educsci15030343

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of *p* values. *Psychonomic Bulletin & Review*, *14*(5), 779–804. https://doi.org/10.3758/BF03194105

Wang, G., Sun, Z., Gong, Z., Ye, S., Chen, Y., Zhao, Y., Liang, Q., & Hao, D. (2024). *Do advanced language models eliminate the need for prompt engineering in software engineering?* https://doi.org/10.48550/arXiv.2411.02093

Zamfirescu-Pereira, J. D., Wong, R. Y., Hartmann, B., & Yang, Q. (2023). Why Johnny can't prompt: How non-AI experts try (and fail) to design LLM prompts. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (pp. 1–21). https://doi.org/10.1145/3544548.3581388

Zung, I., Imundo, M. N., & Pan, S. C. (2022). How do college students use digital flashcards during self-regulated learning? *Memory*, *30*(8), 923–941. https://doi.org/10.1080/09658211.2022.2058553

(*Appendices follow*)

# Appendix A

## ChatGPT Prompts Used to Generate Questions and Outlines

### AI-Elaborate Prompt Used in Experiments 1–4

You are an expert learning assistant that is tasked with developing practice questions to help students learn a text passage. Please generate [#] open-ended questions drawn from the text enclosed in <> below. All questions must meet the conditions specified in && below.

&

Each question should consist of a single sentence and should be answerable in one sentence.

Each question should have a single clause each. That is, it should be in the form of "Can you explain X," which contains a single clause, and not "Can you explain X, and how does X relate to Y," which contains multiple clauses.

Each question should call for a single answer and not multiple answers.

Do not create questions that are made up of multiple smaller questions.

Do not use the phrase "in the text"

Do not use the phrase "the text"

Do not use the phrase "in the passage"

Do not use the phrase "the passage"

Do not use the text ", and"

Each question should not be longer than 15 words.

All questions should involve recognizing main ideas, evaluating their validity, making inferences, relating the text to other knowledge, or being able to apply the acquired knowledge effectively.

&

Below are some generic example questions enclosed in %%. Use them as inspiration but do not rely on them completely. Generic example questions:

%

1. What is the main concept of the subject of this text?
2. Can you provide a summary of the various categories or elements described in the text?
3. What are the primary themes, ideas, or processes explored in the text?
4. How to differentiate between different aspects or components within the text?
5. Given a problem scenario, can you apply what you have learned from the text to come up with a solution?

%

< [Brakes passage was inserted here].>

Modifications between experiments included the number of questions solicited (Experiment 2) and instructions mandating comparisons (Experiment 4) with different examples, as follows:

All questions should involve comparison. That is, they should require the respondent to identify differences, similarities, or both between two or more concepts or items from the passage.

1. What is the difference between A and B?
2. How are A and B different, and in what ways are they similar?
3. What is the primary difference between A and B?

### AI-Basic Prompt Used in Experiments 3–4

Can you create a series of [#] questions from a passage?

[After an affirmative response, the full text of the passage is provided].

A modification made between experiments involved the addition of phrasing mandating comparison-type questions (Experiment 4), as follows:

Each question involves making a comparison.

### AI-Outline Prompt Used in Experiments 3–4

You are an expert learning assistant that is tasked with helping students to improve their learning by generating a succinct outline of the text enclosed in <> below. You must generate this outline according to the conditions enclosed in %%.

%

The outline should be about [###] words long.

The outline should make use of bullet points.

The outline should have a clear and logical structure that makes it easy for the learner to understand the key contents of the provided text.

The outline should be written in a simple language understandable to the average undergraduate student.

%

< [Brakes passage was inserted here].>

A modification made between experiments involved changing the number of words from 350 to 160 and adding language about concrete rather than abstract words (Experiment 4).

(*Appendices continue*)

## Appendix B

## AI-Generated Outlines Used in Experiments 3–4

### Original Outline (Experiment 3)

I. Introduction to Brake: A device that slows or stops wheel or vehicle movement.

II. Brake Components: Most brakes consist of a fixed part called a brake shoe or block that creates friction against a wheel.

III. Types of Brakes

1. Mechanical Brakes: Operate using levers or cables, commonly found in lightweight bicycles and as emergency or handbrakes in automobiles.
2. Hydraulic Brakes: Use fluids and pistons to transmit force, with brake fluid in cylinders and metal tubes connecting them to wheel cylinders.
    - Types of Hydraulic Brakes: Drum brakes on rear wheels and disk brakes on front wheels, both activated by fluid pressure.
3. Drum Brakes: Comprise a cast-iron drum and semicircular brake shoes that stop or slow down the wheel when pressed against the drum.
4. Disk Brakes: Utilize a cast-iron disk and a caliper assembly with friction pads, similar to rubber brake shoes on bicycles, to slow or stop the wheel.
5. Air Brakes: Found in buses, heavy trucks, and trains, they use compressed air released from a storage unit to push pistons and activate brake shoes against wheels.
6. Train Brake System: Each train car has a compressed air tank, and a main brake pipe enables the train engineer to operate all brakes together.
7. Other Types of Brakes
    - Power Brakes: Enhance brake effectiveness by utilizing engine vacuum and atmospheric pressure difference to drive hydraulic fluid, reducing the need for strong pedal pressure.
    - Electric Brakes: Common in house trailers, these brakes use electromagnets attached to wheels to slow or stop the vehicle when charged with electricity from a battery.

### Simpler Outline (Experiment 4)

I. Definition of Brakes

- Devices that slow or stop wheels or engines by creating friction

II. Types of Brakes

1. Mechanical Brakes
    - Use levers or cables to press brake shoes against wheels.
    - Caliper brakes (bicycles): Levers squeeze cables to press shoes against the wheel rim.
    - Emergency brakes (cars): Prevent rolling, also known as handbrakes or parking brakes.
2. Hydraulic Brakes
    - Use brake fluid pressure to operate brakes.
    - Drum brakes: Shoes press against a rotating drum to slow down or stop.
    - Disk brakes: Pads press against a rotating disk, commonly on front wheels.
3. Air Brakes
    - Use compressed air to activate brakes, found in buses, trucks, and trains.
    - Trains use a main brake pipe to control all brakes simultaneously.

III. Other Brakes

- Power brakes: Use engine vacuum to reduce braking effort.
- Electric brakes: Electromagnets activate brake shoes, used in house trailers.

Received January 13, 2025
Revision received July 21, 2025
Accepted August 17, 2025 ■