

Explanation Timing and Metacognitive Calibration in Retrieval Practice

Conor Caldwell-Gibbons

Advanced Cognitive Psychology: 4145

4/25/2026

Lei Yuan

Abstract

Artificial intelligence tools are getting a lot of use by students as study aids and yet their effects on metacognitive calibration remain poorly understood. Metacognitive calibration refers to the alignment between a learner's subjective confidence and objective performance, and poor calibration can lead learners to disengage from the kind of effortful processing authentic learning requires. This study examined whether the timing of an AI-generated explanation relative to a retrieval attempt affects calibration. Using a within-subjects counterbalanced design, 17 participants completed multiple choice questions on human immunology under two conditions, an AI-first condition in which a labeled AI explanation preceded each question, and a retrieval-first condition in which participants answered and rated confidence before viewing the explanation. The results showed that the AI-first condition produced significantly higher accuracy (65.9% vs. 45.9%, $p < .001$) and significantly higher confidence ratings ($p = .034$) but calibration scores did not differ significantly between conditions ($p = .075$) with both conditions showing underconfidence. These findings suggest that brief AI explanations presented before retrieval may function as effective instruction rather than as a source of illusory comprehension, and can potentially impact or change how AI tools are integrated into educational settings.

Introduction

Artificial intelligence tools are getting more and more integrated into educational settings with students usually consulting AI-generated explanations to support or “confirm” their understanding of complex material. While these tools can provide accurate and detailed content, their effect on metacognitive calibration remains a fairly unexplored question/area of study. Metacognitive calibration refers to the alignment between a learner's subjective confidence and their objective performance (Pieger et al., 2016). When calibration is poor people might overestimate their comprehension and disengage from that effortful processing that true learning cannot be achieved without. One mechanism through which AI explanations may impair calibration is the fluency illusion, which is when a learner encounters a coherent and well organized explanation and the ease of processing creates a subjective sense of understanding that does not reflect actual knowledge (Risko and Gilbert, 2016). Research on cognitive offloading further suggests that delegating explanatory work to an external tool reduces internal encoding and distorts confidence judgments with people commonly reporting feeling more competent than their own memory would justify (Risko and Gilbert, 2016).

A contrasting framework is provided by research on retrieval practice indicating that generating answers from memory before receiving feedback produces stronger and more durable learning than passive exposure to explanations (Karpicke and Roediger, 2008; Rowland, 2014). The testing effect looks like it is driven by the effortful and generative nature of recall attempts opposed to just the act of answering. A meta-analysis by Rowland (2014) found that retrieval practice with feedback yielded effect sizes nearly double those without feedback ($g = 0.73$ versus $g = 0.39$), and the benefit was strongest when learning content was complex and retrieval demands were high. Retrieval attempts also seem to support more accurate metacognitive

monitoring by confronting learners with their knowledge gaps before consulting any external source (Grinschgl et al., 2021). In the retrieval-first condition the explanation arrives after an attempt and functions as corrective feedback and in doing so preserves the productive struggle Rowland (2014) identified as the key mechanism underlying the testing effect. This suggests that the sequence in which explanation and retrieval occur may matter more than simply whether retrieval practice is used at all.

Prior work has documented that AI study tools can harm retention and inflate confidence, but these studies have typically compared AI assisted study to no AI use without manipulating the timing of when you see the AI relative to retrieval. Barcaui (2025) demonstrated in a preregistered randomized controlled trial that undergraduates who studied with unrestricted ChatGPT access scored significantly lower on a surprise retention test 45 days later than those using traditional methods (57.5% vs. 68.5%, $d = 0.68$). Fernandes et al. (2026) found that AI assisted reasoning produced a wider confidence-accuracy gap than unaided reasoning with AI assisted participants overestimating their performance by approximately four percentage points relative to actual accuracy. Neither of these studies examined whether the harm would persist if the AI explanation appeared after retrieval rather than before it.

Participants in the present study completed multiple choice questions under two within-subjects conditions. In the AI-first condition a labeled AI explanation came before each question while in the retrieval-first condition, participants answered and rated confidence before viewing the explanation. If the AI-first condition produces greater overconfidence while raw accuracy remains comparable, this would show that the harm is specifically metacognitive which is a dissociation that prior AI versus no-AI comparisons could or did not provide. It was hypothesized that the AI-first condition would produce greater overconfidence, as pre-exposure

to a fluent explanation was expected to inflate confidence without a corresponding improvement in accuracy.

Method

Participants

I recruited 17 participants through convenience sampling including students enrolled in the same course as me as well as some personal connections for my study. Demographic data including age and gender were not systematically collected as part of the study protocol and no participants were excluded from analysis.

Design

I used a within-subjects counterbalanced design with condition (AI-first vs. retrieval-first) as the independent variable and metacognitive calibration and quiz accuracy as the primary dependent variables. Condition order and question set assignment were counterbalanced across participants using PsyToolkit (Stoet, 2010) so that approximately half of participants began with the AI-first block and the rest began with the retrieval-first block. Two matched question sets were rotated across conditions to control for question set difficulty.

Materials

I developed two sets of reading passages and multiple choice questions on the topic of human immunology covering things from innate and adaptive immunity, T cells, B cells, antigens, immunological memory, to autoimmune responses. I chose this topic because it might be partially familiar for most participants without risking ceiling performance and because it is unlikely to give away or reveal the study hypotheses to the participants. Each set had 10 multiple choice items with four options and AI-generated explanations from Gemini were prepared for each question and were explicitly labeled as AI Explanation to operationalize the independent

variable. The experiment was created and administered using PsyToolkit (Stoet, 2010; Stoet, 2017) in which participants used their own devices and keyboards to provide responses and both accuracy and reaction times were recorded on each trial. Confidence ratings were rescaled from the original 1 to 4 scale to a 0 to 1 scale using a linear transformation to allow direct comparison with accuracy scores in the calibration calculation.

Procedures

Participants accessed the study through a PsyToolkit survey link and after providing consent they completed two experimental blocks in their assigned counterbalanced order. In the AI-first block the participants first read an immunology passage, after which at the start of each trial a labeled AI explanation was displayed on screen. After reading the explanation participants pressed a key to proceed to the question then selected one of four response options then immediately rated their confidence on a 4-point scale ranging from 1 (not at all confident) to 4 (very confident).

In the retrieval-first block participants read the immunology passage and at the start of each trial there was no explanation presented, then they answered the multiple choice question and rated their confidence before the AI explanation appeared on screen as feedback. Each block concluded with an overall confidence rating and a brief condition preference item. The sequence of events for both conditions is illustrated in Table 1.

<i>Step</i>	<i>Block A (AI first)</i>	<i>Block B (Retrieval first)</i>	<i>What was recorded</i>
<i>1</i>	<i>Read passage</i>	<i>Read passage</i>	<i>Nothing</i>
<i>2</i>	<i>See AI explanation</i>	<i>Answer multiple choice</i>	<i>Block B: accuracy and RT</i>
<i>3</i>	<i>Answer multiple choice</i>	<i>See AI explanation</i>	<i>Block A: accuracy and RT</i>
<i>4</i>	<i>Rate confidence 1-4</i>	<i>Rate confidence 1-4</i>	<i>Both: confidence rating</i>
<i>5</i>	<i>Repeat steps 2-4 for all 10 questions</i>	<i>Repeat steps 2-4 for all 10 questions</i>	<i>Nothing</i>
<i>6</i>	<i>End screen</i>	<i>End screen</i>	<i>Both: overall confidence and condition preference</i>

Table 1. Sequence of events in both conditions

Results

All analyses were performed in Posit Cloud (Posit Team, n.d.) using the Tidyverse (Wickham et al., 2019) and nlme (Pinheiro et al., 2017) packages. A blockorder analysis confirmed that condition order did not moderate any dependent variable ($p = .373$), showing that counterbalancing was successful and that observed differences are attributable to the experimental manipulation rather than order effects.

Accuracy.

A linear mixed effects model was used to examine the effect of condition on trial level accuracy with condition entered as a fixed effect and subject as a random intercept. Results showed that condition had a significant main effect on accuracy, with $F(1, 322) = 15.60$, $p < .001$. Participants in the AI-first condition answered 65.9% of questions correctly compared to

45.9% in the retrieval-first condition, representing a large effect ($d = 1.51$). These results are illustrated in Figure 1.

Response Time.

A linear mixed effects model was also used to examine the effect of condition on response time with condition entered as a fixed effect and subject as a random intercept. Results showed that condition did not have a significant effect on response time, $F(1, 308) = 0.03$, $p = .872$. Median response times were similar across conditions with AI-first participants responding in approximately 1100 ms and retrieval-first participants responding in approximately 1030 ms. These results are illustrated in Figure 2.

Calibration.

A paired-samples t-test was conducted to examine the effect of condition on participant-level calibration scores, calculated as mean confidence (rescaled 0 to 1) minus mean accuracy. Results showed that the AI-first condition produced a more negative calibration score ($M = -0.190$, $SD = 0.174$) compared to the retrieval-first condition ($M = -0.079$, $SD = 0.220$), $t(16) = -1.90$, $p = .075$, $d = -0.46$. The difference did not support the predicted direction as both conditions showed underconfidence rather than the predicted overconfidence, though the AI-first condition was more underconfident than retrieval-first. Calibration results are illustrated in Figure 3.

Confidence.

Mean confidence ratings (rescaled 0 to 1) were higher in the AI-first condition ($M = 0.469$) than in the retrieval-first condition ($M = 0.380$), $t(16) = 2.33$, $p = .034$, $d = 0.56$. This difference says that viewing the AI explanation before answering produced reliably higher

subjective confidence, though the magnitude of this confidence increase was outpaced by the accuracy gain, resulting in the more negative calibration score observed in the AI-first condition.

General Discussion

This study asked whether the timing of an AI explanation relative to retrieval affects metacognitive calibration, and I tested this using a counterbalanced within-subjects design in which participants completed immunology questions under both an AI-first and a retrieval-first condition. The AI-first condition produced significantly higher accuracy and significantly higher confidence but the calibration difference between conditions trended while not reaching significance ($p = .075$, $d = -0.46$). The predicted mechanism was the fluency illusion, in which a coherent AI explanation encountered before retrieval was expected to inflate confidence independently of actual learning, producing overconfidence in the AI-first condition (Risko and Gilbert, 2016). In fact both conditions showed underconfidence rather than overconfidence, with AI-first participants being more underconfident than retrieval-first, which is the opposite of what I predicted. One interpretation of this is that the AI explanation in the current study functioned as effective instruction rather than as a heuristic cue to illusory understanding. Under these conditions the explanation might have supported real encoding rather than offloading. Risko and Gilbert (2016) argue that offloading produces competence feelings that exceed actual encoding, but this effect may require sustained or repeated access to the external resource (AI or other offloading tool) rather than a single brief exposure. The retrieval-first condition preserved the generative retrieval attempt that Karpicke and Roediger (2008) and Rowland (2014) identified as the core mechanism of the testing effect, although the calibration advantage this should have produced was offset by much lower accuracy.

Fernandes et al. (2026) found that AI assisted reasoning produced a four-point overconfidence gap relative to unaided performance. They gave participants ongoing access to an AI assistant throughout the reasoning task, enabling consistent offloading of the cognitive work itself. In my study participants received a single brief explanation before each question and then answered from memory with no AI available. This design may have operationalized AI as instruction rather than AI as cognitive substitute, which may have affected encoding and calibration in a fundamentally different way. A trending calibration difference ($p = .075$, $d = -0.46$) with a medium effect size suggests the present study was too small to reliably detect the effect rather than that no effect exists. The data points towards the idea that cognitive offloading theory as applied to AI study tools may need to account for the degree and duration of offloading rather than treating all AI exposure as equivalent. A single brief explanation that disappears before retrieval may occupy a fundamentally different position on the offloading spectrum than ongoing AI consultation and is a distinction future designs should account for.

For students the present findings suggest that the real concern about AI study tools might not be whether students use them but how those tools are set up relative to when retrieval happens. A brief AI explanation used as pre-study material before a self-test may improve immediate performance without the severe calibration costs that ongoing AI use produces. For instructors these findings support the value of designing assignments integrated with AI that require students to attempt retrieval before consulting AI explanations, preserving the productive struggle that Rowland (2014) identified as the key mechanism of the testing effect.

The study has a few limitations worth noting for future research, starting with the fact that a sample of 17 participants is quite small, and the trending calibration effect ($p = .075$) might have reached significance if n was larger. The study only measured immediate performance with

no delayed retention test therefore leaving whether the accuracy advantage of the AI-first condition persists or reverses at a one week follow up unanswered. Given Barcaui's (2025) finding that AI study costs emerge most clearly at 45 days, a delayed test seems like the most important addition for future work or replications. The immunology content was also generally unfamiliar to most participants which may have suppressed confidence across both conditions and made the calibration difference harder to detect. Future work should examine whether effects differ with content participants have partial, complete, or no prior knowledge of. It remains unclear whether the effects observed here depend specifically on participants knowing the explanation was AI generated or whether any fluent pre-question explanation would produce similar patterns.

The central contribution of this study is the reframing of the question from whether students should use AI to when they should use it. Brief AI explanations presented immediately before retrieval seem to work more like actual teaching than like a shortcut that creates false confidence. Explanation timing relative to retrieval, rather than AI use may be what is actually causing the metacognitive costs documented in prior work.

Figures

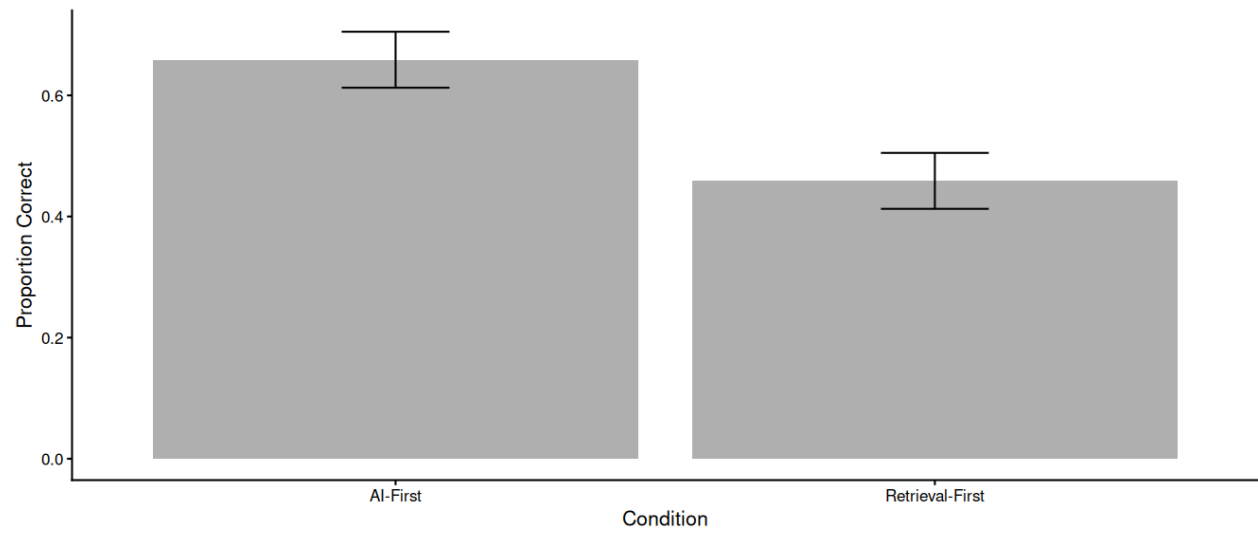


Figure 1. Mean proportion correct by condition. Error bars represent standard error of the mean.

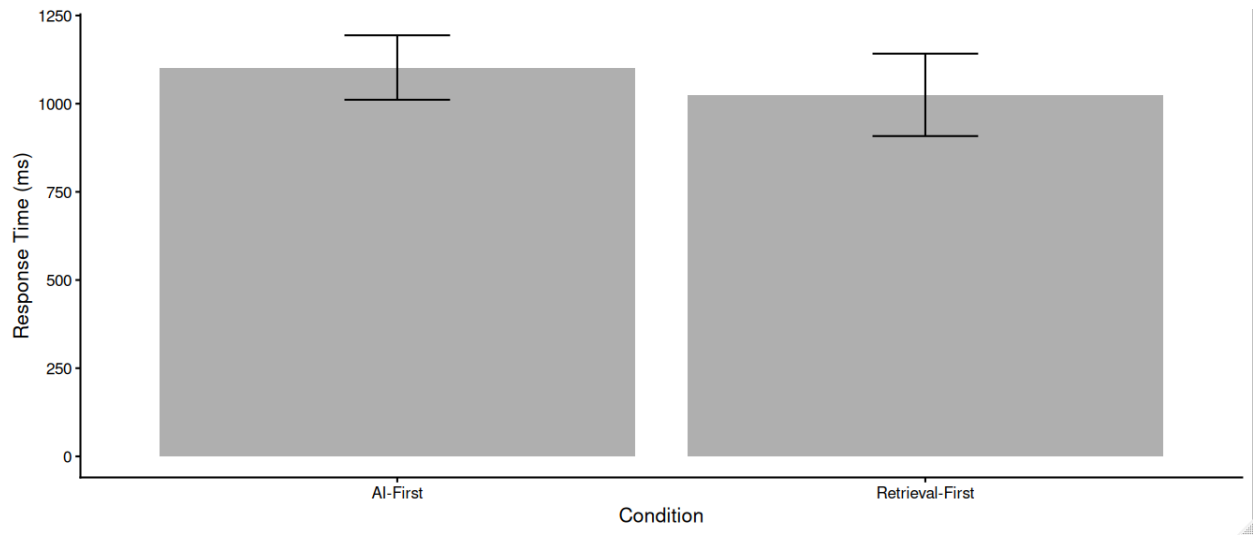


Figure 2. Median response time in milliseconds by condition with error bars representing standard error of the mean.

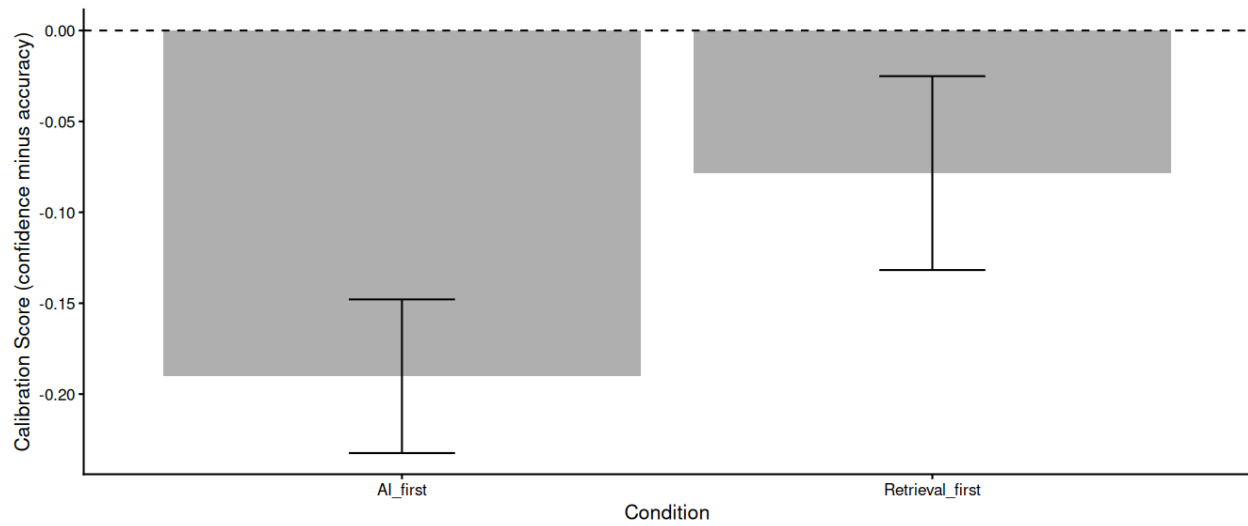


Figure 3. Mean calibration score by condition with scores below zero indicating underconfidence. Error bars represent standard error of the mean and the dashed line represents perfect calibration.

References

- Barcaui, A. (2025). ChatGPT as a cognitive crutch: Effects of AI-assisted study on long-term retention. *Social Sciences and Humanities Open*, 11, 102287. <https://doi.org/10.1016/j.ssaho.2025.102287>
- Fernandes, D., Villa, S., Nicholls, S., Haavisto, O., Buschek, D., Schmidt, A., Kosch, T., Shen, C., and Welsch, R. (2026). AI makes you smarter but none the wiser: The disconnect between performance and metacognition. *Computers in Human Behavior*, 175, 108779. <https://doi.org/10.1016/j.chb.2025.108779>
- Grinschgl, S., Papenmeier, F., and Meyerhoff, H. S. (2021). Consequences of cognitive offloading: Boosting performance but diminishing memory. *Quarterly Journal of Experimental Psychology*, 74(9), 1–15. <https://doi.org/10.1177/17470218211008060>
- Karpicke, J. D., and Roediger, H. L. (2008). The critical importance of retrieval for learning. *Science*, 319(5865), 966-968. <https://doi.org/10.1126/science.1152408>
- Pieger, E., Mengelkamp, C., and Bannert, M. (2016). Metacognitive judgments and disfluency: Does it matter whether judgments refer to the current or a future test? *Learning and Instruction*, 44, 13-21. <https://doi.org/10.1016/j.learninstruc.2016.01.012>
- Pinheiro, J., Bates, D., DebRoy, S., and Sarkar, D. (2017). *nlme: Linear and nonlinear mixed effects models*. R package version 3.1-131. <https://CRAN.R-project.org/package=nlme>
- Posit Team. (n.d.). *Posit Cloud* [Web application]. Posit Software, PBC. <https://posit.cloud/>
- Risko, E. F., and Gilbert, S. J. (2016). Cognitive offloading. *Trends in Cognitive Sciences*, 20(9), 676-688. <https://doi.org/10.1016/j.tics.2016.07.002>
- Rowland, C. A. (2014). The effect of testing versus restudy on retention: A meta-analytic review of the testing effect. *Psychological Bulletin*, 140(6), 1432-1463.
- Stoet, G. (2010). PsyToolkit: A software package for programming psychological experiments using Linux. *Behavior Research Methods*, 42(4), 1096-1104. <https://doi.org/10.3758/BRM.42.4.1096>
- Stoet, G. (2017). PsyToolkit: A novel web-based method for running online questionnaires and reaction-time experiments. *Teaching of Psychology*, 44(1), 24-31. <https://doi.org/10.1177/0098628316677643>

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D. A., François, R., and Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, 4(43), 1686.
<https://doi.org/10.21105/joss.01686>