

Large Language Models as Study Aids: Promises, Pitfalls, and the Risk to Long-Term Knowledge Retention

Conor Liam Caldwell-Gibbons

SN =110596707

Department of Psychology CU Boulder

PSYC 4541

Chandra Reynolds

March 10, 2026

Large language models such as ChatGPT, Claude, and Gemini have become widely accessible to undergraduate students, and many now use them regularly as on-demand study aids. A growing concern in cognitive and educational psychology is whether outsourcing thinking to AI tools compromises the internal knowledge structures students build over time. Cognitive offloading theory (Risko & Gilbert, 2016) holds that when people delegate cognitive tasks to an external tool they reduce internal processing demands and in doing so they also reduce how much they retain.

This paper argues that although LLMs offer real benefits for immediate task performance, their use as primary study tools poses a significant risk to long-term knowledge retention by reducing the effortful cognitive processing that durable learning requires. Durable learning refers to the formation of stable memory traces that persist over time and transfer successfully to new contexts as opposed to surface level familiarity that fades shortly after encoding. Three aspects of this risk are examined, the first being the direct effects of LLM use on delayed retention (Barcaui, 2025), followed by changes to higher-order cognition during AI-assisted study (Essel et al., 2024) and the effects of LLM-based information gathering on the depth of knowledge that students build (Melumad & Yun, 2025). These studies differ in design with Barcaui (2025) using a randomized controlled trial, Essel et al. (2024) using a pretest-posttest approach, and Melumad and Yun (2025) a multiexperiment series, together addressing complementary questions about retention, cognitive engagement, and knowledge quality.

Critical Issues

The most direct evidence on this issue comes from a randomized controlled trial in which undergraduate students were assigned to either study with ChatGPT or use traditional non-AI study methods. The trial enrolled 120 undergraduate business administration students (68 male, 52 female, ages 18 to 24), randomly divided into two equal groups of 60, with retention assessed via a surprise test administered 45 days after the study period and group differences analyzed using an independent samples t-test. Students in the ChatGPT condition performed significantly worse than those in the traditional study condition (Barcaui, 2025). This pattern is consistent with what cognitive offloading theory would predict, showing that when students delegate the work of processing and explaining material to an LLM they reduce the internal encoding that supports later retrieval. The act of working through material independently even when it is effortful or slow, is precisely what commits that material to long-term memory (Risko & Gilbert, 2016). A well organized LLM response removes that effort and with it the encoding benefit. Even when students feel prepared after using an LLM that sense of readiness may reflect familiarity with the output rather than genuine internalization of the content.

The second body of evidence shifts attention from what students retain to how they engage with material during the learning process itself. Essel et al. (2024) conducted a pretest-posttest study examining changes in critical thinking, creative thinking, and reflective thinking as students were exposed to instant LLM answers during academic tasks. The study enrolled 125 undergraduate students in Ghana,

predominantly male (around 58 to 63 percent across groups), randomly allocated to a ChatGPT condition (n = 60) or a control condition using traditional databases and search engines (n = 65), with posttest scores on each cognitive skill scale analyzed using ANCOVA while controlling for pretest scores. The findings showed that these cognitive skills declined significantly when LLM responses were immediately available, with qualitative data suggesting that students generated less independent reasoning when they could simply receive an answer on demand. This matters because deep cognitive engagement including questioning and making connections as well as elaborating on new ideas, is not just a byproduct of learning but one of its primary drivers. Grinschgl and Neubauer (2022) argue that LLMs represent a qualitatively different form of cognitive offloading than earlier tools precisely because they can externalize not just information storage but reasoning, synthesis, and evaluation. When those processes are consistently outsourced students are not just learning less efficiently but perhaps they may also be gradually reducing their capacity and inclination for independent intellectual effort.

The third line of evidence concerns the depth and quality of the knowledge students build when they rely on LLM-generated summaries rather than engaging in active information search. Melumad and Yun (2025) conducted experiments comparing knowledge outcomes for people who learned from an LLM summary against those who searched the web independently and integrated information from multiple sources. Across seven online and laboratory experiments totaling 10,462 participants, participants were randomly assigned to learn via either ChatGPT or standard Google web search with depth of knowledge assessed through measures of advice quality

including length, originality, and semantic uniqueness of responses. The results showed that learners who used LLM summaries built shallower knowledge structures overall even when they reported feeling well informed at the end of the task. This reveals a gap between perceived and actual knowledge, meaning students relying on LLMs may not recognize that they are underlearning. When students search independently they decide what is relevant, compare sources, and synthesize what they find, and receiving a coherent summary from an LLM skips that process entirely. As cognitive offloading theory would predict, reducing the difficulty of acquiring information also reduces how richly it is encoded (Risko & Gilbert, 2016).

Conclusions

The convergence of evidence does not argue for dismissing LLMs entirely, as strategic use can reduce cognitive load on routine tasks and preserve attention for higher-order thinking (Grinschgl & Neubauer, 2022). The key evaluative question then is therefore not whether LLMs are harmful in an absolute sense but whether the way students currently use them is compatible with the cognitive conditions that support durable learning. Pedagogically sound integration such as using LLMs to check understanding after an initial attempt rather than to generate initial answers, may preserve the benefits of these tools while limiting the retention costs the literature identifies. Working through this literature has made me more deliberate about how I use LLMs in my own studying. The gap between perceived and actual learning described by Melumad and Yun (2025) is particularly difficult to dismiss given how naturally confident an LLM-generated explanation can make a reader feel, and I now use these tools

primarily to check understanding after working through material on my own rather than as a first stop.

References

- Barcaui, A. (2025). ChatGPT as a cognitive crutch: Evidence from a randomized controlled trial on knowledge retention. *Social Sciences and Humanities Open*, 12, 102287. <https://doi.org/10.1016/j.ssaho.2025.102287>
- Essel, H. B., Vlachopoulos, D., Essuman, A. B., & Amankwa, J. O. (2024). ChatGPT effects on cognitive skills of undergraduate students: Receiving instant responses from AI based conversational large language models. *Computers and Education: Artificial Intelligence*, 6, 100198. <https://doi.org/10.1016/j.caeai.2023.100198>
- Grinschgl, S., & Neubauer, A. C. (2022). Supporting cognition with modern technology: Distributed cognition today and in an AI enhanced future. *Frontiers in Artificial Intelligence*, 5, 908261. <https://doi.org/10.3389/frai.2022.908261>
- Melumad, S., & Yun, J. H. (2025). Experimental evidence of the effects of large language models versus web search on depth of learning. *PNAS Nexus*, 4(10), pgaf316. <https://doi.org/10.1093/pnasnexus/pgaf316>
- Risko, E. F., & Gilbert, S. J. (2016). Cognitive offloading. *Trends in Cognitive Sciences*, 20(9), 676–688. <https://doi.org/10.1016/j.tics.2016.07.002>