Social Sciences & Humanities Open 12 (2025) 102287

Contents lists available at ScienceDirect

Social Sciences & Humanities Open

journal homepage: www.sciencedirect.com/journal/social-sciences-and-humanities-open

Regular Article

ChatGPT as a cognitive crutch: Evidence from a randomized controlled trial on knowledge retention

André Barcaui

Universidade Federal Do Rio de Janeiro (UFRJ), Av.Pasteur, 250 – Botafogo, Rio de Janeiro, Cep: 21941-901, Brazil

ARTICLE INFO

Keywords:
ChatGPT
Knowledge retention
Cognitive offloading
Higher education
Experimental learning

ABSTRACT

The rapid integration of generative artificial intelligence into higher education has outpaced empirical understanding of its effects on fundamental learning processes. To address this gap, this randomized controlled trial (n = 120) tested ChatGPT's impact on long-term knowledge retention in undergraduates learning AI. Participants were randomly assigned either to use ChatGPT as a study aid (AI-Assisted Group) or to use only traditional, non-AI study methods (traditional learning group). Knowledge retention was assessed with a surprise test 45 days after learning. Students who used ChatGPT scored significantly lower on the retention test (57.5 % correct) compared to those who studied traditionally (68.5 % correct), t (83) = −3.19, p = .002, Cohen's d = 0.68. This suggests that unrestricted ChatGPT use impaired long-term retention, likely by reducing the cognitive effort that supports durable memory. The findings align with cognitive offloading theory and the 'desirable difficulties' principle: while AI assistance may ease initial learning, it appears to undermine the effortful processes needed for robust learning. These results have important implications for how generative AI tools should be integrated into higher education.

1. Introduction

The rapid diffusion of generative AI into higher education has created a fundamental tension. On one hand, tools like ChatGPT promise scalable tutoring, instant feedback, and personalized support; on the other, they may short-circuit the effortful cognitive processes that make learning stick, especially memory consolidation over time. This tension has become particularly urgent given the unprecedented speed of adoption: within two months of its launch in November 2022, ChatGPT had reached 100 million users, making it the fastest-growing consumer application in history and fundamentally altering the educational landscape (Ansari et al., 2023; Memarian & Doleck, 2023; Ngo, 2023; Strzelecki, 2023).

This paper offers an empirical test of that tension. Grounded in cognitive offloading and desirable difficulties, we ask whether unrestricted access to ChatGPT during self-directed study helps or harms long-term knowledge retention. Cognitive offloading predicts that when external tools shoulder core mental work, learners expend less effort during encoding; desirable difficulties predict that removing productive struggle reduces durable memory. Together, these perspectives yield a clear, testable prediction about retention measured well after initial learning.

Recent studies have begun documenting concerning patterns (Gerlich, 2025; Farrokhnia et al., 2024; Jošt et al., 2024; Lee et al., 2025; Memarian & Doleck, 2023). When students have unrestricted access to AI assistants, they often exhibit reduced critical thinking, decreased problem-solving effort, and potential over-reliance on automated responses (Bastani et al., 2024; Niloy et al., 2023; Cotton et al., 2023). This phenomenon aligns with established theories of cognitive offloading, the tendency to rely on external resources as a substitute for internal memory, which may fundamentally alter how students encode and retain information (Niloy et al., 2023; Cotton et al., 2023). Cognitive offloading refers to delegating mental operations (e.g., retrieval, synthesis, planning) to external tools, thereby reducing internal effort during encoding. The ease of obtaining instant, comprehensive answers from ChatGPT could be creating a generation of students who remember where to find information rather than the information itself, echoing concerns raised about the "Google effect" on memory (Sparrow et al., 2011).

The theoretical framework of desirable difficulties provides a crucial lens for understanding these effects. Desirable difficulties are purposeful challenges at study time (e.g., retrieval, spacing, generation) that lower short-term fluency but improve long-term retention. This well-established principle demonstrates that certain challenges during

E-mail addresses: barcaui@facc.ufrj.br, andre.barcaui@gmail.com.

https://doi.org/10.1016/j.ssaho.2025.102287
Received 11 July 2025; Received in revised form 20 November 2025; Accepted 25 November 2025
Available online 29 November 2025

learning, such as effortful retrieval, generation of answers, and productive struggle, lead to stronger long-term retention despite causing short-term performance costs (Bjork & Bjork, 2011). By potentially eliminating these beneficial difficulties, AI tools might optimize for immediate task completion while undermining the deeper learning processes necessary for durable knowledge construction.

Despite the urgency of these questions, empirical evidence on AI's impact on learning outcomes remains limited. Most existing studies are either theoretical, survey-based (Lo, 2023), or focus on immediate performance instead of long-term retention (Cefa et al., 2025; Rokhsari, 2025). This gap in experimental evidence is particularly concerning given the rapid, widespread adoption of these tools in educational settings. Without rigorous empirical data, institutions are making consequential decisions about AI integration based on speculation in place of evidence.

This study addresses this critical gap through a randomized controlled trial examining how ChatGPT use affects knowledge retention in higher education. By comparing students who studied AI and machine learning concepts with ChatGPT assistance against those using traditional methods, and measuring retention after a 45-day delay, we provide experimental evidence on the long-term cognitive impacts of generative AI in learning contexts. We also prioritized **ecological validity,** by which we mean that key aspects of the learning context mirrored authentic student behavior, real course topics, naturalistic access to tools, and a delayed surprise assessment after 45 days.

Based on cognitive offloading, desirable difficulties, and memory consolidation, we derive the following pre-registered hypotheses and one exploratory question.

- H1 (Directional). Students who use ChatGPT for self-directed learning will exhibit significantly lower long-term knowledge retention than students who use traditional, non-AI study methods. This follows from cognitive offloading and desirable difficulties, both of which predict detrimental effects when effortful processing is reduced (Risko & Gilbert, 2016; Bjork & Bjork, 2011).
- H2 (Forgetting pattern). The forgetting curve will differ by condition: the AI-assisted group will show a steeper decline over time than the traditional learning group, consistent with memory consolidation theory (Radvansky et al., 2022) and prior evidence that bypassing effort weakens encoding (Lee et al., 2025; Sparrow et al., 2011).
- RQ (Exploratory). Does prior AI experience moderate the relationship between study condition and long-term retention? Competing mechanisms make the direction ambiguous: greater experience could enable more effective, less detrimental use (Heung & Chiu, 2025), or it could reflect more ingrained offloading habits that exacerbate shallow encoding (Gerlich, 2025). We therefore treat moderation as an exploratory question rather than a confirmatory hypothesis.

We expect a steeper forgetting curve in the AI condition because reduced effort during initial learning should impair systems-level consolidation: with less retrieval, generation, and elaboration at study time, hippocampus-dependent traces remain weaker and less integrated (Radvansky et al., 2022). Consistent with the "Google effect" (Sparrow et al., 2011) and recent work on AI-related offloading (Gerlich, 2025; Lee et al., 2025), weaker initial encoding should accelerate decay, not merely lower the delayed endpoint. Thus, condition differences should be temporal (different trajectories) rather than purely static (different means at Day 45).

Recent empirical work in AI-augmented blended learning shows that over-reliance on ChatGPT can dampen both higher-order thinking and self-regulated learning. For example, Lee et al. (2024) propose a guidance-based ChatGPT approach and find that giving students hints rather than direct answers help preserve high-order cognition and self-regulation. Similarly, Wu et al. (2024) compare convergent ChatGPT responses with divergent search strategies and observe that ChatGPT-based convergent information influences knowledge construction in distinct ways. Other investigations further highlight benefits and risks: Kasneci et al. (2023) document performance gains but also caution about dependency, and Baidoo-Anu and OwusuAnsah (2023) explore equity and access vulnerabilities in AI use. In contrast to these, our study's unique contribution is a pre-registered randomized design that isolates the causal effect of AI assistance on long-term retention, measured via a surprise delayed test at 45 days. Whereas much prior work emphasizes immediate performance or metacognitive judgments, our focus centers on the **durability of learning.**

This investigation is particularly timely as educational institutions worldwide grapple with policies regarding AI use. Our findings aim to inform evidence-based decisions about how to integrate these powerful tools while preserving the cognitive engagement necessary for meaningful learning.

2. Generative AI adoption in higher education

The emergence of generative AI in educational settings represents a paradigm shift in how students access and process information. Since ChatGPT's public release in November 2022, its adoption in academic contexts has been unprecedented, with studies documenting widespread usage among university students globally (Strzelecki, 2023; Strzelecki et al., 2024). Recent surveys indicate that between 60 and 80 % of university students have experimented with ChatGPT for academic purposes, with regular usage patterns emerging across diverse disciplines (Ahmed et al., 2024; Liao et al., 2024).

The rapid integration of these tools has generated both enthusiasm and concern within the academic community. Proponents highlight the democratizing potential of AI-powered learning assistance, noting its capacity to provide personalized tutoring, instant feedback, and adaptive learning experiences previously available only through one-on-one instruction (Kasneci et al., 2023; Jeon & Lee, 2023). Evidence suggests that when strategically employed, ChatGPT can enhance student engagement, support differentiated instruction and facilitate deeper exploration of complex topics (Zhang & Tur, 2024).

However, this technological integration has proceeded largely without empirical validation of its effects on fundamental learning processes. While initial studies have focused on immediate performance metrics and user satisfaction (Lim et al., 2023), critical questions about **long-term** learning outcomes remain largely unexplored. This gap is particularly concerning given the scale and speed of adoption, with educational institutions making policy decisions based on limited evidence about cognitive impacts.

3. Cognitive offloading and digital dependency

The concept of cognitive offloading provides a lens for understanding AI's impact on learning. This phenomenon, extensively studied in the context of digital technologies, suggests that when individuals can rely on external sources for information storage and processing, they tend to reduce their own cognitive effort (Risko & Gilbert, 2016). The implications for education are profound: while offloading can enhance immediate task performance, it may simultaneously undermine the cognitive processes necessary for deep learning and retention.

Recent investigations into AI-mediated cognitive offloading have yielded concerning findings. Gerlich (2025) documented that frequent users of generative AI report decreased confidence in their critical thinking abilities and reduced cognitive effort in problem-solving tasks. This aligns with earlier work on the "Google effect" (Sparrow et al., 2011), which demonstrated that easy access to information leads individuals to remember where to find information rather than the information itself. However, AI assistants like ChatGPT represent a more extreme form of offloading, providing not just information retrieval but also synthesis, analysis, and problem-solving capabilities.

The educational implications are particularly troubling. Lee et al. (2025) found that knowledge workers who regularly used generative AI

for analytical tasks showed significant reductions in self-reported critical thinking effort and problem-solving confidence. Similarly, Yang et al. (2025) demonstrated that high school students using ChatGPT for programming exercises exhibited lower learning gains and reduced engagement compared to those using traditional instructional methods. These findings suggest that while AI tools may facilitate task completion, they potentially inhibit the effortful processing required for meaningful learning.

4. Desirable difficulties and the learning process

The framework of desirable difficulties, developed by Bjork and Bjork (2011), posits that certain challenges during learning, while impeding short-term performance, enhance long-term retention and transfer. These beneficial difficulties include spacing practice sessions, interleaving different topics, testing instead of re-studying, and generating answers rather than recognizing them. The cognitive effort required to overcome these challenges strengthens memory traces and promotes deeper understanding.

Contemporary research continues to validate the importance of effortful processing in learning. Agarwal et al. (2021) conducted a comprehensive review of retrieval practice studies, finding consistent benefits across 50 real-world educational experiments. The act of actively retrieving information from memory, even when difficult, produces superior long-term retention compared to passive review. Similarly, Maceiras et al. (2025) demonstrated that engineering students who engaged in active learning methods showed over 50 % better retention of key concepts after several weeks compared to those in traditional lecture formats.

The intersection of AI assistance and desirable difficulties presents a fundamental tension. When students use ChatGPT to immediately obtain explanations, examples, or solutions, they bypass the productive struggle that characterizes effective learning. This convenience may optimize for immediate understanding but potentially undermines the formation of durable memories. As Bjork and Bjork (2011) emphasize, the conditions that produce optimal performance during learning are often directly opposed to those that produce optimal long-term retention.

5. Memory consolidation and the forgetting curve

Memory consolidation denotes the time-dependent stabilization and reorganization of new memories from fragile traces into more durable representations. Ebbinghaus's forgetting curve, first described in 1885, remains one of the most robust findings in memory research. The curve demonstrates that without active rehearsal or retrieval, memory for newly learned information decays exponentially, with the steepest decline occurring in the first hours and days after learning (Radvansky et al., 2022). This fundamental principle has profound implications for educational practice and the evaluation of learning interventions.

Modern neuroscience has elucidated the mechanisms underlying the forgetting curve, revealing that memory consolidation requires time and cognitive effort. During consolidation, memories transition from fragile, hippocampus-dependent representations to more stable, cortically distributed networks. This process is facilitated by active retrieval, elaborative processing, and meaningful connections to existing knowledge; precisely the types of cognitive activities that may be diminished when students rely heavily on AI assistance.

The temporal dynamics of forgetting become particularly relevant when evaluating educational technologies. While immediate post-learning assessments may show minimal differences between AI-assisted and traditional learning, the true test lies in long-term retention. Studies that fail to include delayed testing may miss critical differences in memory durability. This methodological consideration informed our decision to assess retention after a 45-day interval, allowing sufficient time for differential forgetting patterns to emerge.

6. Empirical evidence on AI and learning outcomes

The empirical literature on AI's impact on learning outcomes, while growing rapidly, presents a complex and sometimes contradictory picture. Several studies have documented potential benefits when AI is integrated thoughtfully into educational practices. For instance, Heung and Chiu (2025) found that structured use of ChatGPT for specific learning activities could enhance student engagement without compromising outcomes. Similarly, Sun et al. (2024) reported no significant differences in programming performance between students who received ChatGPT assistance and those who worked independently, suggesting that under certain conditions, AI tools need not impair learning.

However, a growing body of evidence points to potential negative effects, particularly when AI use is unrestricted or replaces rather than supplement traditional learning activities. Niloy et al. (2023) demonstrated that students using ChatGPT for creative writing tasks showed decreased originality and quality over time compared to those relying on their own abilities. Jošt et al. (2024) found that programming students who became dependent on AI assistance struggled more with novel problems requiring conceptual understanding. These findings suggest that while AI can provide immediate support, over-reliance may inhibit the development of independent problem-solving skills.

The mixed findings in the literature likely reflect the critical importance of implementation context. Systematic reviews by Abu Khurma et al. (2024) and Ansari et al. (2023) highlight that outcomes depend heavily on how AI tools are integrated into learning activities, the level of guidance provided, and the specific learning objectives. This underscores the need for controlled experimental studies that isolate the effects of AI use while holding other factors constant.

7. Theoretical integration and research gap

Together, these perspectives provide a framework for understanding how generative AI might impact learning (Fig. 2) (see Fig. 1). When students use ChatGPT to circumvent cognitive challenges, they engage in a form of offloading that eliminates desirable difficulties and potentially weakens memory consolidation processes. This theoretical synthesis predicts that while AI assistance might facilitate immediate task completion and comprehension, it could simultaneously undermine the formation of durable, transferable knowledge.

Despite this strong theoretical foundation, empirical evidence directly testing these predictions remains scarce. Most existing studies have focused on immediate performance, user satisfaction, or qualitative experiences rather than long-term retention. The few studies examining retention have typically used short intervals or failed to control for confounding variables. This gap is particularly problematic given the rapid adoption of AI tools in education and the high stakes involved in educational policy decisions.

The present study addresses this critical gap by providing experimental evidence on how ChatGPT use affects knowledge retention over a meaningful time interval. By randomly assigning participants to AI-assisted versus traditional study conditions and measuring retention after 45 days, we test whether the theoretical predictions about cognitive offloading and reduced effortful processing translate into measurable differences in learning outcomes. This evidence is essential for informing evidence-based decisions about the appropriate role of generative AI in educational settings.

8. Methods

This section first orients the reader to what follows: the study design and general framework, participants and recruitment, conditions and materials (including the common topic set), procedure and timeline, the retention test, and validation and fidelity. Fig. 3 summarizes the three phases; details appear in the subsections below.

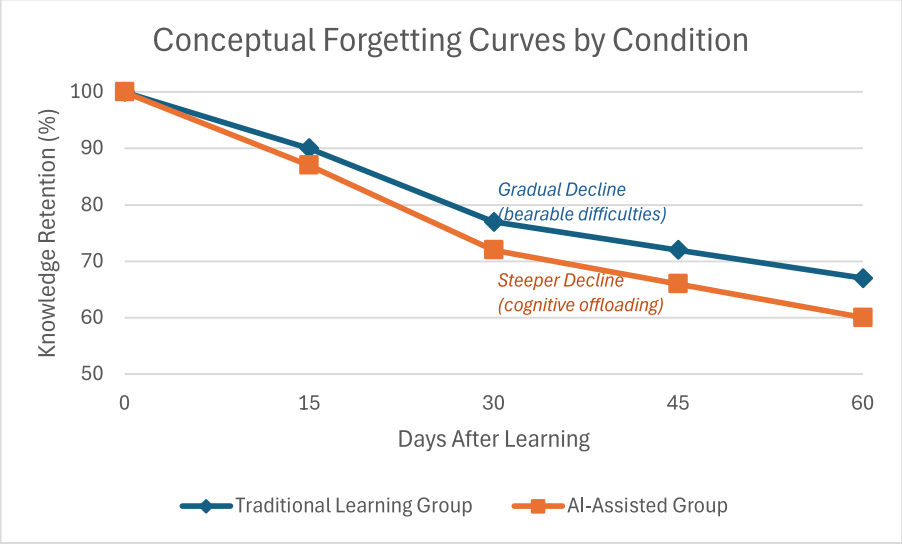

Fig. 1. Hypothesized forgetting curves: AI-assisted learning shows steeper decline (cognitive offloading), while traditional learning shows more gradual decay (desirable difficulties).

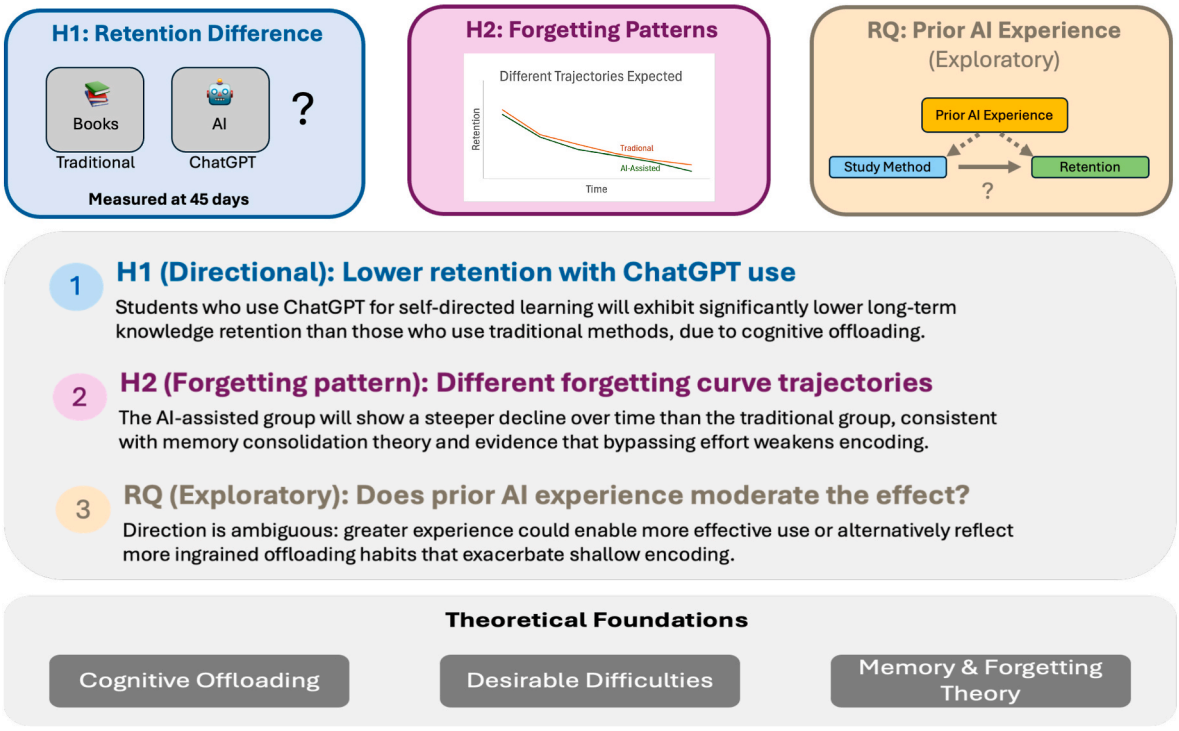

Fig. 2. Conceptual integration of cognitive offloading, desirable difficulties, and consolidation motivating our predictions; see Introduction for rationale.

8.1. *Study design and general framework*

This study employed a randomized controlled trial (RCT) design to investigate the impact of generative AI use on long-term knowledge retention. The experiment was conducted over a three-month period from October 2024 to January 2025, structured in three distinct phases.

- **Phase 1 (October 2024):** Baseline assessment and participant characterization
- **Phase 2 (November 2024):** Learning intervention with randomized conditions
- **Phase 3 (January 2025):** Delayed retention testing (45 days post-intervention)

8.2. *Participants and recruitment*

The extended timeline between intervention and testing was specifically designed to assess long-term retention rather than immediate recall, addressing a critical gap in the existing literature. To further support ecological validity, participants engaged in self-directed study with minimal guidance in the AI condition, and outcomes were assessed after a naturalistic 45-day interval without advance notice.

Participants were undergraduate business administration students

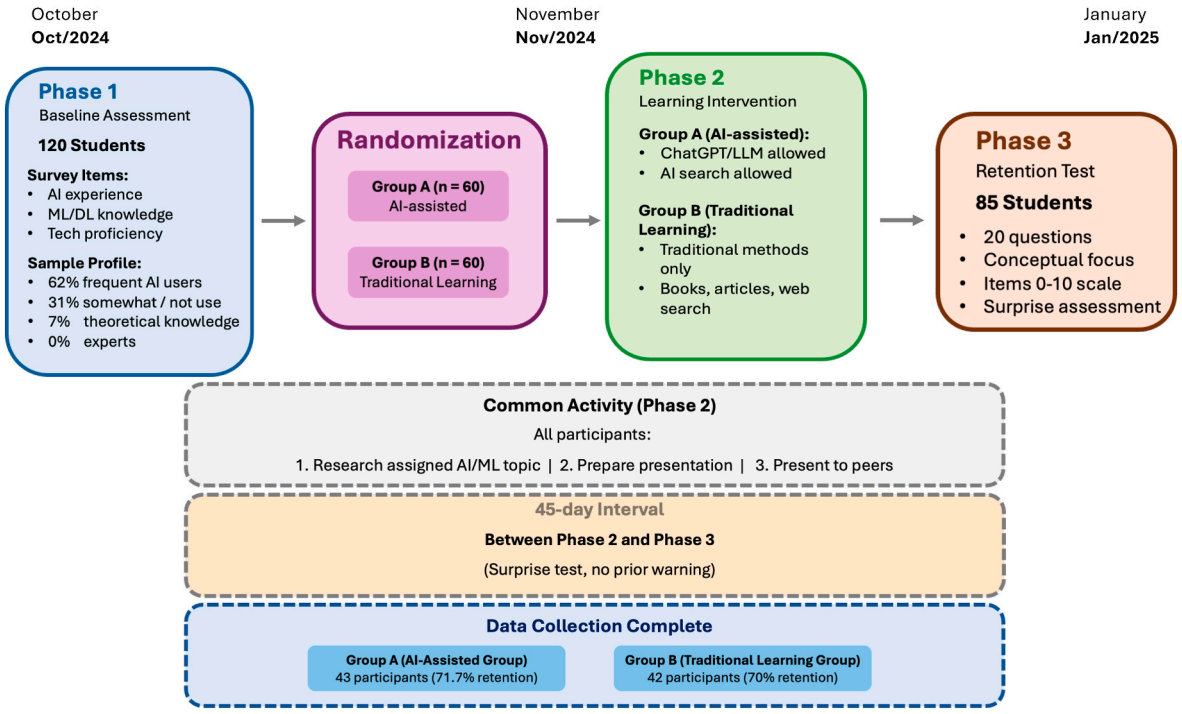

Fig. 3. Experimental design flowchart showing the three phases of the study, randomization process, and participant retention across groups. The 45-day interval between the learning intervention (Phase 2) and surprise retention test (Phase 3) was designed to assess long-term knowledge retention under naturalistic forgetting conditions.

recruited through convenience sampling from a large Brazilian university. This approach facilitated high participation rates and study completion, as students were already engaged with the professor in an academic context. While convenience sampling may limit generalizability, it offered important advantages for this initial experimental investigation, including enhanced compliance with study protocols and reduced attrition. The recruitment process emphasized voluntary participation and made clear that study involvement would not affect course grades or academic standing. Eligibility criteria included current enrollment in business administration programs, no prior formal training in artificial intelligence or machine learning, and commitment to complete all study phases. These criteria ensured a relatively homogeneous sample in terms of prior technical knowledge while maintaining ecological validity for typical undergraduate populations.

Rationale for sampling frame and convenience. We used a convenience sample of Brazilian business undergraduates because the researcher teaches in this context, enabling ecologically valid, low-friction recruitment and compliance. Management courses require conceptual retention and transfer—precisely the outcomes theorized to be vulnerable to cognitive offloading and the removal of desirable difficulties—making this population theoretically appropriate. While the underlying mechanism should generalize beyond business programs, effect magnitudes may vary across disciplines; we therefore outline transportability considerations for future replications.

The final sample comprised 120 students with balanced gender representation (68 males, 52 females) and ages ranging from 18 to 24 years, typical of undergraduate cohorts. All participants were native Portuguese speakers completing their studies in Portuguese, eliminating potential confounds from language barriers. Baseline assessment of AI familiarity revealed a diverse range of prior experience: while no participants (0 %) identified as complete novices to AI, the majority (62 %) reported frequent use of ChatGPT or equivalent tools, 31 % indicated recent or initial use of these technologies, 7 % possessed theoretical knowledge of machine learning or deep learning without professional practice, and none (0 %) identified as experts or professionals in the field. This distribution reflects the current reality of AI tool adoption among undergraduate students, where casual use has become commonplace even without formal training (see Table 1) (see Table 2).

Participants were randomly assigned to experimental conditions using a simple randomization procedure. After completing the baseline assessment, each participant was allocated to either the AI-assisted or traditional learning group through random number generation, ensuring each student had an equal probability of assignment to either condition. This straightforward randomization approach resulted in balanced groups (n = 60 each) and minimized selection bias.

Given the nature of the intervention, participants could not be blinded to their assigned condition as they necessarily knew whether they were using AI tools. However, important methodological safeguards were implemented to preserve study integrity. The retention test was administered uniformly to all participants without revealing its connection to the earlier learning phase, and it was framed as a general knowledge assessment instead of a measure of retention from their presentations. Furthermore, all data were analyzed using coded group identifiers, with statistical analyses conducted without knowledge of which code represented which experimental condition. This partial

Table 1
Summary of participant demographics and baseline characteristics by experimental condition.

Characteristic	AI-Assisted Group (N = 60)	Traditional Learning Group (N = 60)
Gender		
Males	34 (56.7 %)	34 (56.7 %)
Females	26 (43.3 %)	26 (43.3 %)
Age Range	18–24 years	18–24 years
Prior AI Familiarity		
Complete Novice	0 (0 %)	0 (0 %)
Recent/Initial User	19 (31.7 %)	18 (30.0 %)
Frequent User	37 (61.7 %)	37 (61.7 %)
Theoretical Knowledge	4 (6.7 %)	5 (8.3 %)
Expert/Professional	0 (0 %)	0 (0 %)

Table 2
Key statistical outcomes for knowledge retention.

Group	N (Completed Test)	Mean Score (10-point scale)	Standard Deviation	t-statistic	Degrees of Freedom (df)	p-value	Cohen's d	95 % CI for Cohen's d
Traditional Learning	42	6.85	1.7	−3.19	83	0.002	0.68	[0.24, 1.12]
AI-Assisted Learning	43	5.75	1.5					

Fig. 5, a box plot of retention scores, further illustrates the higher median score and less variability in the traditional learning group compared to the AI-assisted group, which showed a wider spread of scores.

blinding approach, while not eliminating all potential biases, reduced the risk of experimenter effects influencing the results.

8.3. Conditions and materials

The study compared two learning approaches that reflect realistic educational scenarios. A curated common topic set was used across conditions, standardized for scope and workload (≈8–12 core ideas; ≈25–35 min of study).

Topic assignment and fairness. Each participant was randomly assigned one topic from the common set using a reproducible random-number procedure. To ensure equity, assignment was block-randomized by difficulty band (Appendix A4), yielding the same distribution of Band-1/2/3 topics in both conditions. Crucially, the Day-45 retention test sampled evenly across the full topic set, preventing students from tailoring preparation to a known test topic; topic-specific advantages therefore average out at the cohort level, and our subgroup analysis indicates the condition effect holds across domains.

Group A (AI-assisted group, n = 60). Participants used OpenAI ChatGPT via the free web interface (no API, no plugins). The available model during the study window was GPT-4, web app; the UI displayed a GPT-4 model label; no plugins/browsing; web interface only. Default system settings were used; web browsing, third-party tools, and custom instructions/memory were disabled. No structured tutorial or prompt-engineering guidance was provided; the intention was to capture AI use as it naturally occurs, with participants free to interact as they typically would. Permitted uses included obtaining conceptual explanations, generating examples and analogies, structuring presentations, and synthesizing information across sources. To clarify academic integrity and privacy, participants received a brief guardrails notice reminding them to verify claims with independent sources, avoid verbatim copy-paste, and refrain from entering personal or confidential data; the exact text appears in Appendix A6.

Group B (traditional learning group, n = 60). Participants were instructed not to use AI chatbots or generative assistants during preparation. Permitted resources included (a) course notes and instructor slides (b) university library databases (e.g., Scopus, Web of Science), and open-access repositories, (c) standard web search engines without AI chat features, and (d) textbooks and peer-reviewed articles relevant to the assigned topic. Prohibited resources included ChatGPT, Copilot, Gemini, Perplexity, Claude, and any tool marketed as an "AI assistant," "chatbot," or "generative" tool. Fidelity note: We did not directly monitor AI non-use in the traditional learning group; to aid replication, Appendix A2 provides a minimal two-item post-task compliance check and a brief honesty declaration template.

8.3.1. Topic set and materials (common to both groups)

All participants studied the same curated set of AI/ML topics spanning foundations, methods, applications, and ethics; the full list with brief learning objectives appears in Appendix A4. Using a common set fixed content scope and depth for everyone, ensuring parity across conditions and isolating the learning approach as the experimental contrast. Topics were chosen for beginners to cover about 8–12 core ideas and a reading or study load of roughly 25–35 min per topic. To support the traditional condition without introducing AI coaching, participants received a one-page Start-here guide and brief citation guidance, as documented in Appendix A3.

8.4. Procedure and timeline

Participants had two weeks to research the material and prepare a 10-min presentation with visual aids. The presentation requirement encouraged meaningful engagement, while presentation quality was not formally assessed to avoid confounding retention with public-speaking ability or anxiety. Presentations were delivered to small groups of 8–10 peers, creating a low-stakes but authentic learning environment that encouraged thorough preparation without excessive performance pressure.

The study employed multiple assessment points to capture both baseline characteristics and learning outcomes. During Phase 1, participants completed a baseline assessment of prior experience with AI technologies, general technological proficiency, academic performance indicators, and learning preferences. AI familiarity was measured on a 5-point ordinal scale ranging from complete novice to expert user, with behavioral anchors for consistent interpretation. These data allowed us to examine whether prior AI experience moderated the relationship between study condition and retention outcomes.

8.5. Outcome measure: retention test

The primary outcome measure was a knowledge retention test administered approximately 45 days after the learning intervention. The test used a single best answer format with five options and one keyed correct alternative. Conceptual understanding was defined as correct application of a concept to a novel scenario or discrimination among plausible alternatives rather than verbatim definition recall. This instrument consisted of 20 multiple-choice questions (Appendix A5) carefully crafted to assess conceptual understanding rather than rote memorization. The questions were developed through an iterative process involving subject-matter experts in AI and educational assessment. They underwent pilot testing with a separate sample of 30 students and were refined based on item analysis to ensure appropriate difficulty and discrimination. The final instrument demonstrated good internal consistency reliability (Cronbach's α = .82) and content validity across all topic areas.

8.6. Validation and fidelity

The retention test was administered as a surprise assessment to prevent preparatory studying that could mask natural forgetting. Participants were contacted and invited to a follow-up session without being told the specific purpose of the session. Upon arrival, they completed the knowledge-retention test under uniform, proctored conditions. Testing was individual and timed, no feedback was provided during or after the test.

Given the nature of the intervention, participants could not be blinded to condition. Outcome scoring was independent of condition because the test used a single-best-answer format with five options and

machine scoring. Traditional learning group non-use of AI was not directly monitored; Appendix A2 provides a minimal two-item compliance-check template for future replications. To preserve academic integrity and privacy, the AI-assisted group received only a short guardrail notice rather than instructional coaching, as documented in Appendix A6.

Missing data were handled using a principled approach. The primary analyses used complete cases; sensitivity analyses under alternative missing-data assumptions were conducted to check robustness. Of the 120 participants who began the study, 85 completed the retention test (overall follow-up rate 70.8 percent). Completion was balanced across conditions, with 43 in the AI-assisted group and 42 in the traditional learning group, suggesting that attrition was not systematically related to experimental condition.

Assumption checks preceded inference. Normality was examined with Shapiro–Wilk and visual Q–Q inspection, homogeneity of variance with Levene's test, and robust alternatives such as Welch's *t*-test were used when assumptions were violated. Analyses were run in Python (v3.13) with standard libraries such as SciPy and StatsModels at $\alpha = .05$, two-tailed. For the ANCOVA, retention was modeled with group (AI vs. traditional) as a fixed factor and self-reported study time (hours) as a mean-centered covariate; we verified homogeneity of regression slopes via the group × study-time interaction and report adjusted (marginal) means at the sample mean of the covariate. For visualization only, we plot exponential trajectories between Day 0 and Day 45; all statistical inference is based on the observed Day-45 endpoints, not on the interpolated curves.

This study was reviewed by the Department of Business Administration and the Graduate Program in Business Administration at Universidade Federal do Rio de Janeiro (UFRJ) and formal approval was waived as the research constitutes a classroom-based pedagogical study posing no risk to participants, dated October 22, 2025. Participation was voluntary, and written informed consent was obtained from all individual participants included in this study.

9. Results

The primary analysis revealed a statistically significant difference in knowledge retention between the conditions. Participants in the traditional learning group ($M = 6.85$, $SD = 1.7$) scored significantly higher on the retention test than those in the AI-assisted group ($M = 5.75$, $SD = 1.5$), $t(83) = -3.19$, $p = .002$. This difference of approximately 1.1 points on a 10-point test represents a meaningful effect. The traditional learners' mean was roughly 11 % of the test score higher than the AI-assisted learners' mean. An 11 % gap on a course assessment could easily translate to a full letter-grade difference in many academic settings, underscoring the practical significance of this finding.

Assumption checks supported the use of the independent samples *t*-test. Shapiro–Wilk tests indicated no meaningful departures from normality within either group (AI-assisted: $W = 0.98$, $p = .30$; traditional: $W = 0.97$, $p = .19$), and Levene's test indicated homogeneity of variance ($F(1, 83) = 0.62$, $p = .43$). Visual inspection of Q–Q plots likewise showed no major deviations. Together, these results justify the use of the standard independent-samples *t*-test reported above.

The score distributions, visualized in Fig. 4, provided further insights. The traditional learning group's scores were clustered toward the high end, with approximately 73.8 % of those students scoring 6 or above. In contrast, the AI-assisted group's scores were more widely spread out, with only about 51.2 % attaining a score of 6 or higher, suggesting greater inconsistency in outcomes for the AI-assisted learners.

The magnitude of this difference was contextualized by Cohen's effect size, which was calculated as $d = 0.68$ (95 % CI [0.24, 1.12]). This is conventionally interpreted as medium-to-large, meaning that roughly 75 % of the students in the traditional condition scored higher than the average student in the AI-assisted condition. This corresponds to a probability of superiority of about 0.66, indicating a 66 % chance that a randomly chosen traditional learner would have a higher score than a randomly chosen AI-assisted learner.

The following table summarizes the key statistical outcomes for knowledge retention, providing a consolidated view of the primary findings.

The study also investigated whether prior experience with AI moderated the effect of study condition, as per RQ. Contrary to expectation, the correlation between AI familiarity and retention was weak and non-significant ($r = 0.18$, $p = .10$). This non-significant finding is important as it indicates that prior experience with AI tools did not substantially influence learning outcomes regardless of experimental condition. This suggests that the detrimental effect is robust and not merely due to unfamiliarity, reinforcing the later concept of "borrowed competence". Fig. 6 visually depicts this relationship, showing no clear predictive pattern between AI familiarity levels and retention scores.

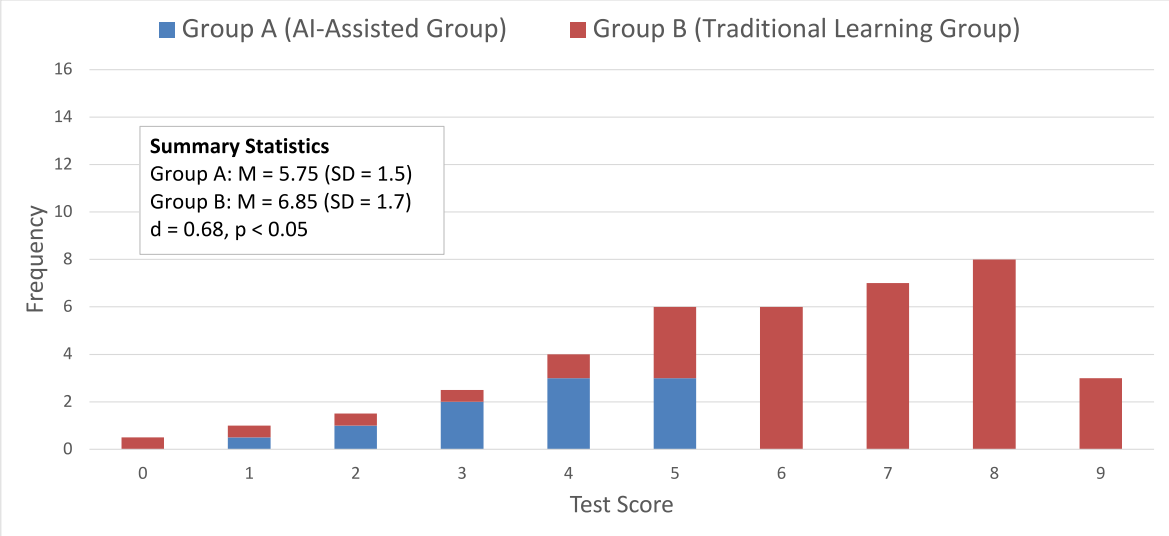

Fig. 4. Distribution of retention test scores by experimental condition. Each histogram shows the score distribution (0–10 scale) for the AI-assisted group (orange) and the traditional learning group (blue). The traditional learning group's distribution is skewed toward higher scores, while the AI-assisted group's scores are more broadly distributed across the scale.

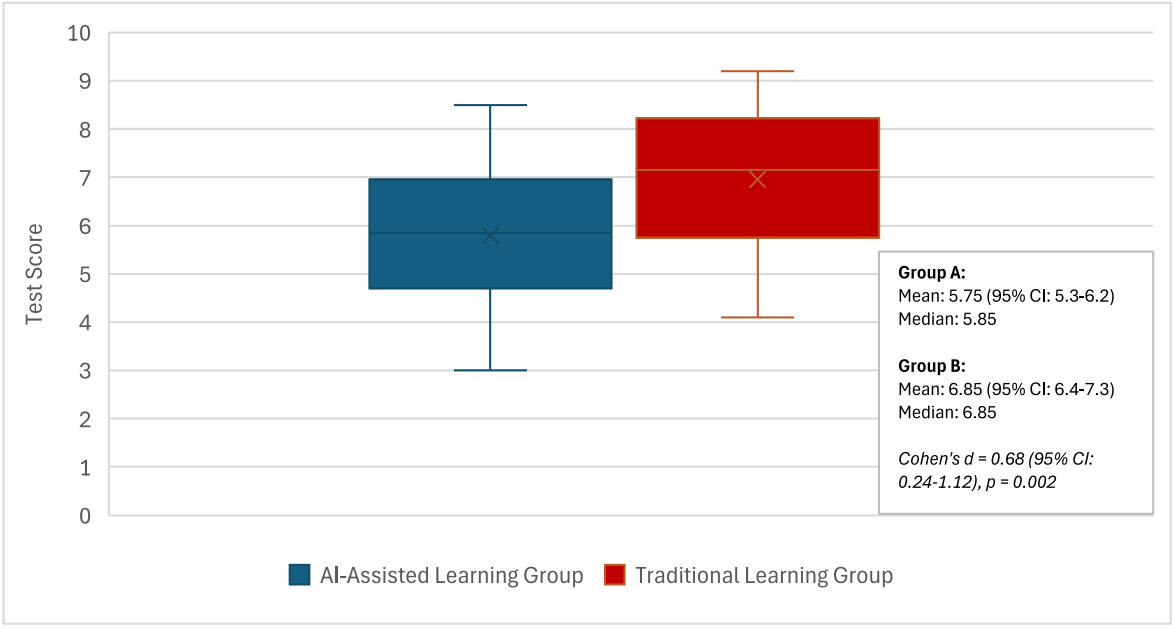

Fig. 5. Retention scores by condition (box plots). Box plots of the retention test scores for the traditional (blue) and AI-assisted (red) groups. The traditional learning group shows a higher median score and less variability.

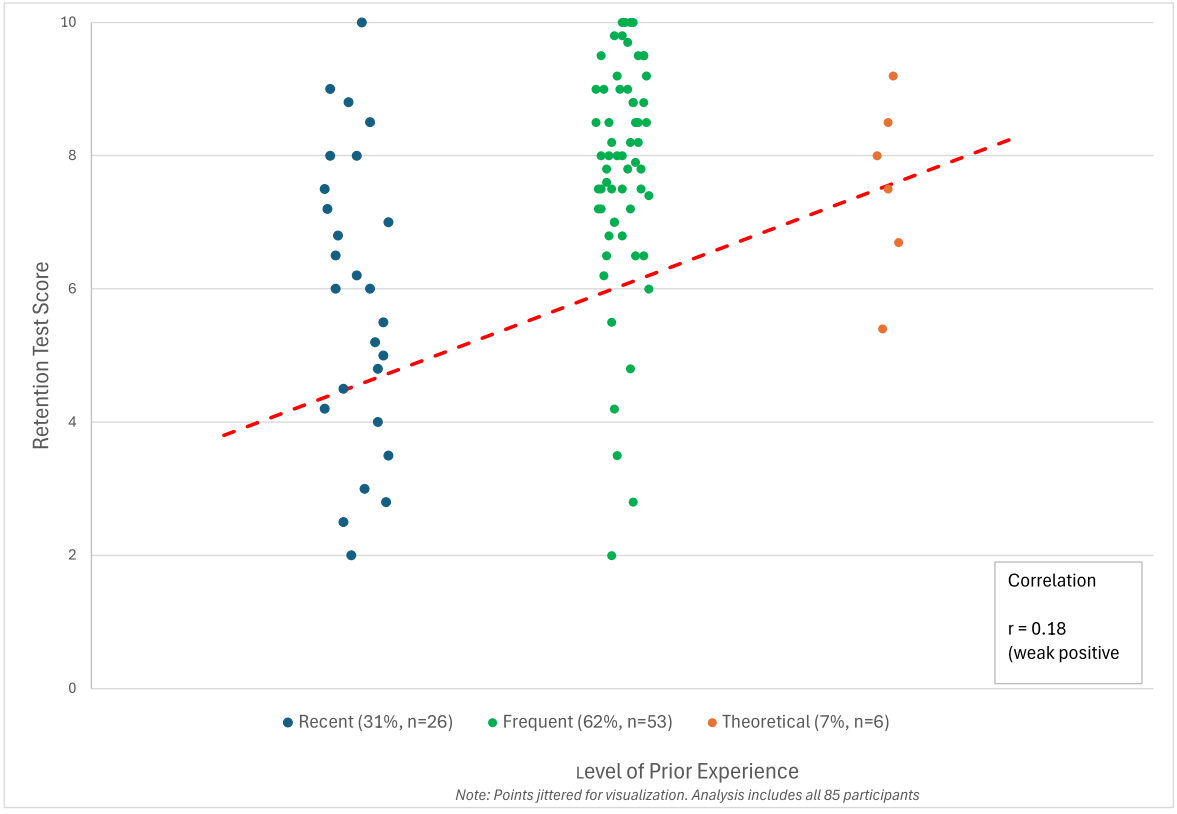

Fig. 6. Relationship between prior AI experience level and retention test scores. Scatter plot displaying individual scores for participants grouped by their self-reported AI familiarity: Recent/Initial Users (blue, 31 % of sample), Frequent Users (green, 62 % of sample), and those with Theoretical Knowledge only (orange, 7 % of sample). The dashed red line indicates the overall linear trend across all groups. Points are jittered horizontally to improve visibility of overlapping values. Despite the apparent gradient in experience levels, the correlation between prior AI experience and retention performance was weak and not statistically significant ($r = 0.18$, $p = .10$), suggesting that familiarity with AI tools did not predict learning outcomes in this study.

Subgroup analyses by topic area revealed some variation in the effect of learning condition. Technical topics showed the largest disadvantage for AI-assisted learning ($d = 0.92$), while ethics and society topics showed a smaller, though still meaningful, effect ($d = 0.45$). However, these differences were not statistically significant ($Q = 2.84$, $p = .42$), suggesting a relatively consistent pattern across content areas. Fig. 7, a

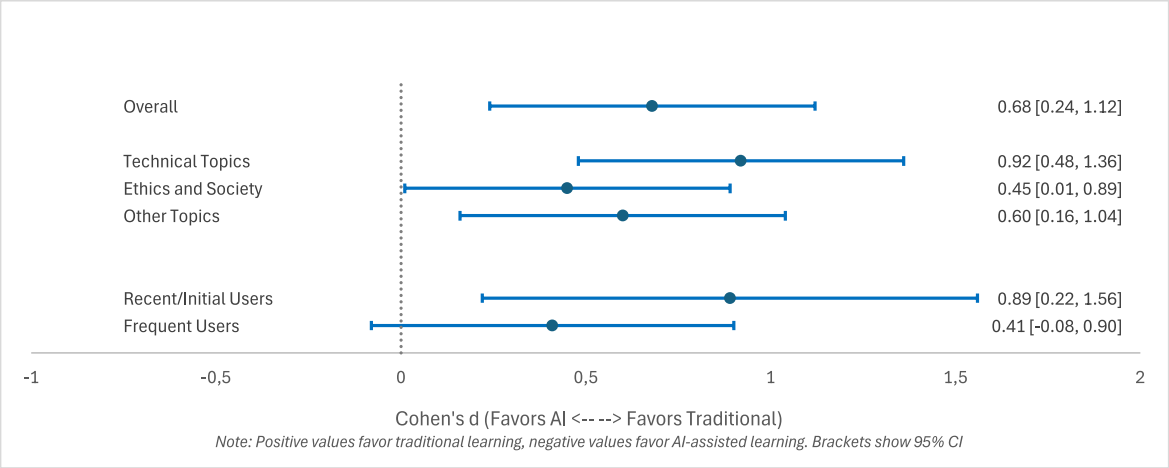

Fig. 7. Forest plot of subgroup effect sizes (Cohen's *d*) for AI assistance on retention. Positive values indicate higher retention in the traditional condition. Horizontal bars are 95 % CIs. The effects consistently favor traditional learning across subgroups (largest in technical topics, $d = 0.92$, and among newer AI users, $d = 0.89$); heterogeneity was low ($I^2 = 28$ %, $p = .21$).

forest plot, illustrates this subgroup effect sizes, consistently favoring traditional learning. The heterogeneity was low ($I^2 = 28$ %, p = .21), further supporting the consistency of the effect across different content domains. Table 3 presents the subgroup analysis of effect sizes. (see Table 4).

An analysis of study habits showed that AI-assisted learners spent significantly less time on the learning task (M = 3.2 h) than traditional learners (M = 5.8 h), t (83) = −4.92, p < .001. This ≈45 % reduction in time-on-task could partly explain the poorer retention in the AI-assisted group, as reduced engagement likely led to more superficial processing of the material. When time-on-task was included as a covariate in an ANCOVA, the effect of AI assistance on retention remained statistically significant, F (1, 82) = 7.89, p = .006, indicating that AI assistance has an independent detrimental effect beyond mere time spent.

Adjusted (marginal) means from the ANCOVA, evaluated at the sample mean of study time (≈4.5 h), indicated that traditional learners still outperformed AI-assisted learners (Traditional: M_adj = 6.50, SE = 0.24; AI-assisted: M_adj = 5.85, SE = 0.23), an adjusted difference of 0.65 points (95 % CI [0.19, 1.11]). Study time was positively associated with retention (β > 0, p < .01), and the group × study-time interaction was not significant, F (1, 81) = 1.02, p = .31, supporting the homogeneity-of-slopes assumption. Thus, even holding time-on-task constant, the traditional condition retained an advantage, consistent with a qualitative difference in depth of processing rather than a purely quantitative time effect.

Fig. 8 plots illustrative forgetting curves for both groups. The traditional learning group appears to decline more gradually than the AI-assisted group, yielding a steeper curve for the latter. Inference rests on the Day-45 comparison: traditional learners outperformed AI-assisted learners by ~11 percentage points (t (83) = −3.19, p = .002, d = 0.68). The continuous paths are visualization-only exponential interpolations anchored at Day 0 (set to 100 %) and the observed Day-45 means, and they are consistent with the interpretation that AI assistance reduced effortful encoding, producing less durable memory traces.

Table 3
Subgroup Analysis of Effect Sizes (Cohen's d) for AI Assistance on Retention.

Subgroup/Topic Area	Cohen's d	95 % CI for Cohen's d
Technical Topics	0.92	[0.48, 1.36]
Ethics and Society	0.45	[0.01, 0.89]
Other Topics	0.60	[0.16, 1.04]
Overall	**0.68**	**[0.24, 1.12]**
Heterogeneity	$I^2 = 28$ %	*p =.21*

10. Discussion

The present study provides compelling evidence that AI assistance during learning can significantly impair long-term knowledge retention. Students who learned using traditional methods outperformed their AI-assisted counterparts by a substantial margin (d = 0.68), directly supporting our first hypothesis (H1), confirming a significant difference in learning outcomes based on the study method employed. This clarifies how emerging technologies interact with core cognitive processes and extends current frameworks to AI.

Our findings also extend cognitive offloading theory into the AI domain. Risko and Gilbert (2016) established that people naturally offload cognitive demands onto external aids; our results demonstrate that generative AI represents a qualitatively different form of offloading. Unlike a calculator or a notebook that offloads specific tasks or memory storage, AI can offload entire cognitive processes – including comprehension, synthesis of information, and even aspects of critical thinking. The medium-to-large effect size ($d = 0.68$) provides empirical evidence that this comprehensive offloading has tangible consequences for learning outcomes. Notably, the forgetting curve for AI-assisted learners was significantly steeper (about an 11 % greater loss of knowledge by day 45), illustrating how offloading to AI disrupted the memory consolidation process.

This differential pattern of knowledge decay offers strong support for our second hypothesis (H2), consistent with different forgetting trajectories between AI-assisted and traditional learning conditions. This observation aligns with the collaborative inhibition framework (Basden et al., 1997), where relying on external support during encoding can create retrieval dependencies that impair independent recall later. However, our data reveal that AI presents a unique case: the especially steep decay for the AI-assisted group suggests not just a retrieval dependency, but also fundamentally weaker initial encoding. In other words, AI assistance affected the **quality** of learning at the outset, not just students' ability to retrieve information later.

The results provide strong empirical support for extending the desirable difficulties framework (Bjork & Bjork, 2020) to AI-mediated learning contexts. The subgroup analysis revealing that technical concepts showed the largest impairment (d = 0.92) is particularly telling. These are precisely the areas where AI could provide the most sophisticated assistance, such as generating code, explaining complex algorithms, or breaking down theoretical frameworks. Yet this is where we observed the greatest learning deficit, reinforcing that difficulty serves a crucial pedagogical function that cannot be circumvented without cost. The heterogeneity analysis ($I^2 = 28$ %, p = .21) indicates relatively

Table 4
Comparison of study habits and time-on-task by experimental condition.

Group	N (Completed Test)	Mean Time-on-Task (Hours)	Standard Deviation	t-statistic	Degrees of Freedom (df)	p-value
Traditional Learning	42	5.8	1.2	−4.92	83	<0.001
AI-Assisted Learning	43	3.2	1.0			

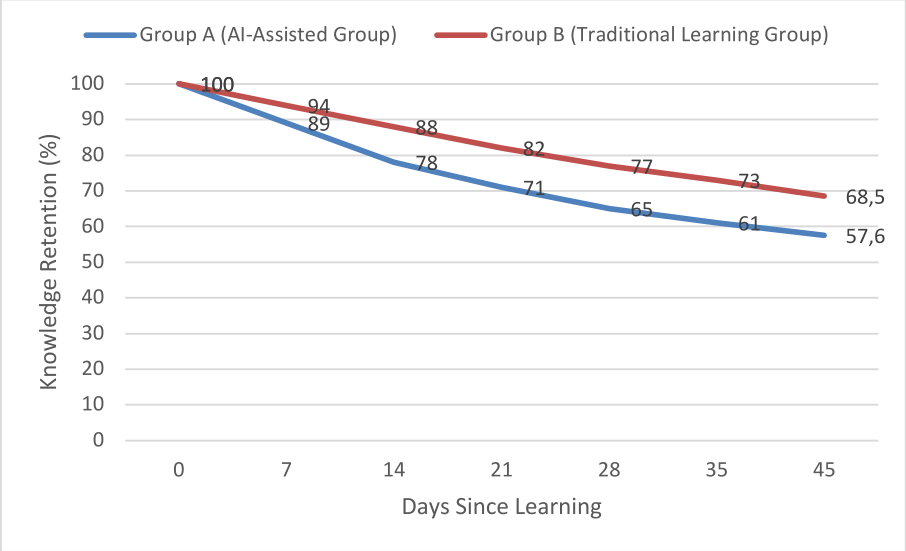

Fig. 8. Illustrative forgetting-curve visualization over the 45 days after learning. Only Day 0 (set to 100 %) and Day 45 (observed means: 68.5 % traditional learning group; 57.5 % AI-assisted group) are data-anchored. The continuous paths interpolate exponential decay between these anchors and are for visualization only. Shaded regions reflect uncertainty around the Day-45 means.

consistent effects across content domains, suggesting that the benefits of struggle extend beyond technical material.

Even in ethics and society topics, where one might expect discussion and reflection to dominate over problem-solving, we still observed a meaningful effect size ($d = 0.45$). This consistency across domains indicates that AI's impact on learning operates through fundamental cognitive mechanisms rather than domain-specific processes. The seminal study by Kapur (2008) on productive failure demonstrates that struggling with problems before receiving instruction enhances learning. Our data extend this finding by showing that when AI provides immediate, articulate solutions, it obscures these productive struggle phases that are essential for deep encoding and long-term retention.

To address our third hypothesis (RQ), we examined whether prior experience with AI would moderate learning outcomes. We found no such moderating effect: the relationship between AI familiarity and retention was weak and non-significant ($r = 0.18$, $p = .10$). In other words, even students who frequently use AI tools did not show improved retention under AI assistance. This result fails to support RQ and carries important metacognitive implications. It suggests that even experienced AI users may not recognize when AI assistance is undermining their learning.

Our data do not directly measure metacognitive calibration or reliance on externalized explanations. Nevertheless, a plausible interpretation consistent with the observed pattern is what we call "borrowed competence." AI can supply structure, vocabulary, and reasoning scaffolds that inflate the feeling of mastery during study without necessarily strengthening memory traces through retrieval or learner-generated elaboration. We therefore present borrowed competence as an interpretive, testable hypothesis rather than a definitive conclusion, and we outline in Future Directions how it could be assessed directly (e.g., metacognitive judgments, process tracing/telemetry, delayed transfer).

Although the AI effect on retention persists after controlling for total study time, the *type* of time students spent likely differed across conditions. Traditional learners' longer sessions may have included more rereading, self-quizzing, spaced review, and elaborative rehearsal—activities that strengthen consolidation via effortful retrieval. In contrast, AI-assisted learners' (shorter) sessions may have focused on prompt refinement and evaluation/curation of model outputs, which can increase fluency and task completion but afford fewer retrieval opportunities. We did not directly measure study strategies, so this account is inferential; nevertheless, acknowledging this qualitative difference offers a plausible explanation for why an AI disadvantage remains even at comparable durations, in line with desirable-difficulties and cognitive offloading perspectives.

Qualitative differences in how the two groups engaged with the material, alongside the statistically significant reduction in time-on-task for the AI-assisted group (3.2 h vs. 5.8 h), shed light on possible mechanisms. While reduced study time clearly contributes to the observed retention deficit, our ANCOVA analysis demonstrated that AI assistance had an independent detrimental effect beyond simply spending less time. According to self-determination theory (Ryan & Deci, 2020), intrinsic motivation is driven by autonomy, competence, and relatedness. Our results imply that AI assistance may undermine these factors: it can reduce autonomy (by guiding learners down AI-chosen paths), create an illusion of competence (through readily available answers), and remove collaborative aspects of learning. Prior eye-tracking research (Alemdag & Cagiltay, 2018) shows that students who grapple with difficult material tend to re-read and revisit concepts, indicating deep processing. We did not track attention directly, but the much shorter study times and lower retention scores in the AI-assisted group suggest their processing was more superficial and linear.

In contrast, the traditional learning group's longer study time likely reflects iterative re-engagement with the content, the kind of effortful process essential for deep learning. The findings call for a fundamental rethinking of how we integrate AI into education. Concretely, two sequencing choices follow from our results: first, delay AI until after an

initial AI-free encoding and quick self-quiz, then use AI to compare answers, surface gaps, and request targeted explanations ("AI-after-attempt"); second, use AI as a retrieval coach rather than an answer engine by having students respond before seeing AI output and then receiving graded feedback, a hinted explanation, and a follow-up question that supports spaced/interleaved review. These tactics preserve effortful retrieval and elaboration while still leveraging AI for feedback and scheduling. First, the substantial 45-day retention deficit (≈57.5 % vs. 68.5 %) shows that unrestricted AI assistance can have lasting negative effects on what students remember.

Second, this effect was largest for technical topics (d ≈ 0.92), suggesting AI help is especially detrimental when learning conceptually demanding material – precisely when productive struggle is most needed. Third, the impact was consistent regardless of students' prior familiarity with AI; mere experience with these tools did not inoculate learners against the offloading effect. In sum, educators should be cautious about allowing unrestricted AI use, especially during initial learning of new or complex material. Careful scaffolding is needed to ensure that AI serves as a supplement to (not a replacement for) the desirable difficulties that foster durable learning.

Taken together, several features strengthen the inference here: the randomized controlled design (with topic assignment block-randomized by difficulty band) supports causal interpretation; the outcome targeted long-term memory via a 45-day delay, addressing a common gap; the surprise test reduces contamination from last-minute cramming; and the account is anchored in established theory (desirable difficulties, cognitive offloading, self-determination). These elements increase confidence that the observed retention deficit reflects changes in learning processes under AI use rather than topical idiosyncrasies or short-term preparation.

10.1. Limitations and future directions

A primary limitation is attrition. Overall loss to follow-up was 29.2 percent, with 120 enrolled and 85 tested after 45 days. Attrition was similar across conditions, and completers did not differ from non-completers on baseline age, gender, or initial knowledge scores, which reduces concerns about differential dropout. Even so, this level of loss can bias estimates if missingness relates to unobserved factors such as motivation, time pressure, or post-treatment experiences. Because outcomes for non-completers are not observed, missing-not-at-random mechanisms cannot be ruled out, so the true effect could be somewhat larger or smaller than reported. Future work should plan intent-to-treat analyses, reduce loss with stronger reminders and modest incentives, and use principled handling of missing data, including multiple imputation accompanied by sensitivity analyses.

The controlled experimental setting strengthened internal validity but likely does not capture the full texture of real-world AI use across an entire semester. Students may develop different usage patterns, metacognitive habits, and norms of collaboration over longer horizons than the six-week intervention studied here. We also focused on a single system, ChatGPT-4 accessed via the web interface during a defined calendar window. Reporting the exact label and window in Appendix A1 supports reproducibility, yet results could vary with other models, interfaces, or future versions.

Generalizability is also constrained by the sampling frame. Participants were drawn from a single university context in a Brazilian business program through convenience recruitment. This improves ecological fit to the target population we teach but narrows external validity. Replications across institutions, educational levels, cultural settings, and disciplines will be important to determine whether effect magnitudes differ, for example between STEM and humanities cohorts or between introductory and advanced courses. In addition, the learning activity involved preparing a brief presentation for peers, which may engage "learning-by-teaching" processes (organizing, explaining for an audience) that differ from exam-oriented studying. This choice fits authentic coursework but narrows generalizability; effects could differ when the goal is performance on a known test, so future work should vary the goal (teach vs. be tested) under matched content and time.

The study deliberately emphasized durable memory by using a surprise assessment 45 days after study to minimize last-minute review. In other words, the design privileges retention as a primary learning outcome. This is a legitimate goal, but it does not exhaust the space of valued outcomes. AI tools may foster other competencies that we did not assess, such as information synthesis, prompt design and iteration, workflow efficiency, rapid drafting, and collaborative and higher-order reasoning. Future work should pair delayed retention with a broader battery that also measures transfer, problem solving, collaboration, efficiency, and study-process quality.

Two key variables were **self-reported**: prior AI experience and time-on-task, which introduces potential recall and social-desirability error. Such noise likely **attenuates** associations (and can under-adjust covariates), so estimates linking time-on-task to retention should be read as **conservative**; future work should use passive logs/telemetry, timers, or ecological momentary prompts to validate or replace self-reports.

Finally, we recorded total time-on-task but did not directly capture strategies. As discussed in the main text, it is plausible that the nature of time differed by condition, with traditional learners engaging in rereading, self-quizzing, spacing, and elaborative rehearsal, and AI-assisted learners spending more effort on prompt refinement and evaluation of outputs. Process data that can adjudicate this mechanism would be valuable, for example brief self-quizzing logs, prompted think-alouds, clickstream and prompt telemetry, and metacognitive judgments that assess calibration. Beyond behavioral evidence, targeted cognitive and neural work that probes retrieval opportunities, prediction-error signals during study, and hippocampal–cortical engagement with and without AI could clarify how assistance changes encoding and consolidation, and in turn guide interventions that preserve effortful retrieval while using AI for feedback and spacing.

11. Conclusion

In summary, this randomized controlled trial provides strong evidence that unrestricted use of generative AI as a study aid can significantly impair long-term knowledge retention. Students who learned without AI retained substantially more information after 45 days than those who used ChatGPT. The effect size ($d = 0.68$) corresponds to an ~11 percentage-point performance gap, underscoring the practical significance of this finding.

These results strongly support our theoretical framework. By providing immediate, comprehensive answers, the AI tool facilitated a form of cognitive offloading that eliminated the desirable difficulties needed for deep learning. Skipping those effortful processes likely led to weaker memory encoding, as evidenced by the steeper forgetting curve in the AI-assisted group. Notably, this detrimental effect occurred across all topic types and was not reduced even for students already familiar with AI, suggesting a robust phenomenon. It may even create a metacognitive blind spot, where students confuse the AI's fluency with their own understanding.

While this study has limitations, including its specific sample and controlled setting, its findings serve as a crucial, evidence-based caution for educators, policymakers, and students. As generative AI tools are rapidly integrated into higher education, it is critical to remain mindful of their potential cognitive costs. Future teaching strategies should aim to harness the benefits of AI without sacrificing the cognitive engagement and productive struggle required for durable learning. This means using AI to complement, not replace, challenging learning activities, avoiding scenarios where the AI becomes a mere cognitive crutch. In the age of AI, the core principles of human learning are not outdated; in fact, they are more important than ever to uphold.

Data availability statement

The data are not publicly available due to privacy or ethical restrictions.

Ethical statement

This study involved human participants (undergraduate business administration students). All participants provided electronic informed consent via email, confirming their voluntary participation and understanding that their involvement would not affect their course grades or academic standing. The study protocol was designed to protect participant anonymity by using study codes from the outset for data collection. The research adhered to ethical guidelines for studies involving human subjects and was conducted with respect for the privacy and rights of all participants.

Declaration of the use of AI assisted technologies

During the preparation of this work, the author utilized Claude Sonnet 4 to assist with reviewing the translation from Portuguese (Brazil) to English. Following the use of this tool/service, the author thoroughly reviewed and edited the content as needed and takes full responsibility for the content and accuracy of the publication. The generative AI tool was used solely for language refinement and did not contribute to the scientific content, data analysis, or conclusions of the study.

Funding statement

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. The study was conducted without external financial support.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ssaho.2025.102287.

References

Abu Khurma, O., Albahti, F., Ali, N., & Bustanji, A. (2024). AI ChatGPT and student engagement: Unraveling dimensions through PRISMA analysis for enhanced learning experiences. *Contemporary Educational Technology, 16*(2), ep503. https://doi.org/10/309935/cedtech/14334.

Agarwal, P. K., Nunes, L. D., & Blunt, J. R. (2021). Retrieval practice consistently benefits student learning: A systematic review of applied research in schools and classrooms. *Educational Psychology Review, 33*(4), 1409–1453.

Ahmed, Z., Shanto, S., Rime, M., Morol, M., Fahad, N., Hossen, M., & Abdullah-Al-Jubair, M. (2024). The generative AI landscape in education: Mapping the terrain of opportunities, challenges, and student perception. *IEEE Access, 12*, 147023–147050. https://doi.org/10.1109/ACCESS.2024.3461874

Alemdag, E., & Cagiltay, K. (2018). A systematic review of eye tracking research on multimedia learning. *Computers & Education, 125*, 413–428. https://doi.org/10.1016/j.compedu.2018.06.023

Ansari, A., Ahmad, S., & Bhutta, S. (2023). Mapping the global evidence around the use of ChatGPT in higher education: A systematic scoping review. *Education and Information Technologies, 29*, 11281–11321. https://doi.org/10.1007/s10639-023-12223-4

Baidoo-Anu, D., & Owusu Ansah, L. (2023). Education in the era of generative artificial intelligence (AI): Understanding the potential benefits of ChatGPT in promoting teaching and learning. *SSRN Electronic Journal, 7*(1), 52–62. https://doi.org/10.61969/jai.1337500

Basden, D. R., Basden, B. H., Bryner, S., & Thomas, R. L., III (1997). A comparison of group and individual remembering: Does collaboration disrupt retrieval strategies? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23*(5), 1176–1188. https://doi.org/10.1037/0278-7393.23.5.1176

Bastani, H., Bastani, O., Sungu, A., Ge, H., Kabakcı, Ö., & Mariman, R. (2024). Generative AI can harm learning [The Wharton School research paper]. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4895486

Bjork, E. L., & Bjork, R. A. (2011). Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. In M. A. Gernsbacher, R. W. Pew, L. M. Hough, & J. R. Pomerantz (Eds.), *Psychology and the real world: Essays illustrating fundamental contributions to society* (pp. 56–64). Worth Publishers.

Bjork, R. A., & Bjork, E. L. (2020). Desirable difficulties in theory and practice. *Journal of Applied Research in Memory and Cognition, 9*(4), 475–479. https://doi.org/10.1016/j.jarmac.2020.09.003

Cefa, B., Macgilchrist, F., ElGamal, H., Bai, J., Zawacki-Richter, O., & Loglo, F. (2025). Responses to the initial hype: ChatGPT supporting teaching, learning, and scholarship? *Open Praxis, 17*(2), 1–27. https://doi.org/10.55982/openpraxis.17.2.872

Cotton, D., Cotton, P., & Shipway, J. (2023). Chatting and cheating: Ensuring academic integrity in the era of ChatGPT. *Innovations in Education & Teaching International, 61*, 228–239. https://doi.org/10.1080/14703297.2023.2190148

Farrokhnia, M. R., Banihashem, S. K., Noroozi, O., & Wals, A. (2024). A SWOT analysis of ChatGPT: Implications for educational practice and research. *Innovations in Education & Teaching International, 61*(3), 460–474. https://doi.org/10.1080/14703297.2023.2195846

Gerlich, M. (2025). AI tools in society: Impacts on cognitive offloading and the future of critical thinking. *Societies, 15*(1), 6. https://doi.org/10.3390/soc15010006

Heung, Y., & Chiu, T. (2025). How ChatGPT impacts student engagement from a systematic review and meta-analysis study. *Computers and Education: Artificial Intelligence*. https://doi.org/10.1016/j.caeai.2025.100361

Jeon, J., & Lee, S. (2023). Large language models in education: A focus on the complementary relationship between human teachers and ChatGPT. *Education and Information Technologies, 28*, 15873–15892. https://doi.org/10.1007/s10639-023-11834-1

Jošt, G., Taneski, V., & Karakatič, S. (2024). The impact of large language models on programming education and student learning outcomes. *Applied Sciences*. https://doi.org/10.3390/app14104115

Kapur, M. (2008). Productive failure. *Cognition and Instruction, 26*(3), 379–424. https://doi.org/10.1080/07370000802212669

Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., … Schulz, C. (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

Lee, H.-Y., Chen, P.-H., Wang, W.-S., Huang, Y.-M., & Wu, T. T. (2024). *Empowering ChatGPT with guidance mechanism in blended learning: Effect of Self-Regulated Learning, Higher-Order Thinking Skills, and Knowledge Construction. International Journal of Educational Technology in Higher Education*. https://doi.org/10.1186/s41239-024-00447-4

Lee, H.-P.(H., Sarkar, A., Tankelevitch, L., Drosos, I., Rintel, S., Banks, R., & Wilson, N. (2025). The impact of generative AI on critical thinking: Self-reported reductions in cognitive effort and confidence effects from a survey of knowledge workers. In *Proceedings of the 2025 CHI conference on human factors in computing systems* (pp. 1–22). Association for Computing Machinery. https://doi.org/10.1145/3706598.3713778.

Liao, Z., Antoniak, M., Cheong, I., Cheng, E., Lee, A., Lo, K., Chang, J., & Zhang, A. (2024). LLMs as research tools: A large-scale survey of Researcher's usage and perceptions. *ArXiv, abs/2411.05025*. https://doi.org/10.48550/arXiv.2411.05025

Lim, W. M., Gunasekara, A., Pallant, J. L., Pallant, J. I., & Pechenkina, E. (2023). Generative AI and the future of education: Ragnarök or reformation? A paradoxical perspective from management educators. *International Journal of Management in Education, 21*(2), Article 100790. https://doi.org/10.1016/j.ijme.2023.100790

Lo, C. K. (2023). What is the impact of ChatGPT on education? A rapid review of the literature. *Education Sciences, 13*(4), 410. https://doi.org/10.3390/educsci13040410

Maceiras, R., Feijoo, J., Alfonsín, V., & Pérez-Rial, L. (2025). Effectiveness of active learning techniques in knowledge retention among engineering students. *Education for Chemical Engineers, 51*. https://doi.org/10.1016/j.ece.2025.01.003. Article 100779.

Memarian, B., & Doleck, T. (2023). ChatGPT in education: Methods, potentials and limitations. *Computers in Human Behavior: Artificial Humans*. https://doi.org/10.1016/j.chbah.2023.100022

Ngo, T. (2023). The perception by university students of the use of ChatGPT in education. *Int. J. Emerg. Technol. Learn., 18*, 4–19. https://doi.org/10.3991/ijet.v18i17.39019

Niloy, M. N., Saha, A., Shafkat, A., & Islam, M. R. (2023). Is ChatGPT a menace for creative writing ability? An experiment. *Journal of Computer Assisted Learning*. https://doi.org/10.1111/jcal.12929. Advance online publication.

Radvansky, G. A., Tamplin, A. K., & Krawietz, S. A. (2022). Forgetting functions and the influence of initial learning: New perspectives on an old debate. *Journal of Experimental Psychology: General, 151*(8), 1809–1818. https://doi.org/10.1037/xge0001198

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

Rokhsari, S. (2025). The impact of artificial intelligence on learners' memory: A systematic review. *Journal of Cognition, Emotion & Education, 3*(2), 18–34. https://doi.org/10.22034/cee.2025.542883.1040

Ryan, R. M., & Deci, E. L. (2020). Intrinsic and extrinsic motivation from a self-determination theory perspective. *Contemporary Educational Psychology. Advance online publication*. https://doi.org/10.1016/j.cedpsych.2020.101860

Sparrow, B., Liu, J., & Wegner, D. M. (2011). Google effects on memory: Cognitive consequences of having information at our fingertips. *Science, 333*(6043), 776–778. https://doi.org/10.1126/science.1207745

Strzelecki, A. (2023). To use or not to use ChatGPT in higher education? A study of students' acceptance and use of technology. *Interactive Learning Environments, 32*, 5142–5155. https://doi.org/10.1080/10494820.2023.2209881

Strzelecki, A., Cicha, K., Rizun, M., & Rutecka, P. (2024). Acceptance and use of ChatGPT in the academic community. *Education and Information Technologies, 29*, 22943–22968. https://doi.org/10.1007/s10639-024-12765-1

Sun, Z., Wang, F., & Liu, S. (2024). ChatGPT-assisted learning in computer programming education: An experimental study. *IEEE transactions on education*. Advance online publication. https://doi.org/10.1109/TE.2024.3021234

Wu, T.-T., Lee, H.-Y., Li, P.-H., Huang, C.-N., & Huang, Y.-M. (2024). Promoting self-regulation progress and knowledge construction in blended learning via ChatGPT-based learning aid. *Journal of Educational Computing Research, 61*(8), 3–31. https://doi.org/10.1177/07356331231191125

Yang, X., Li, J., & Wei, Y. (2025). The effectiveness of ChatGPT in assisting high school students' programming learning: A quasi-experimental study. *Interactive learning environments*. Advance online publication. https://doi.org/10.1080/10494820.2024.XXXXX

Zhang, K., & Tur, G. (2024). Empowering K-12 education with ChatGPT: A systematic review of opportunities and challenges. *Education and Information Technologies, 29* (5), 7955–7974. https://doi.org/10.1007/s10639-024-11927-4