The Effects of Generative AI on Critical Thinking in Educational Contexts: A Systematic Review

Simane Khales

School of Teacher Education, Nanjing University of Information Science and Technology

DOI: https://doi.org/10.47772/IJRISS.2026.101000091

Received: 09 October 2026; Accepted: 14 October 2026; Published: 29 October 2026

ABSTRACT

The rapid integration of generative artificial intelligence (GenAI) tools into educational contexts has intensified debate regarding their impact on higher-order cognitive outcomes, particularly critical thinking (CT). This systematic review synthesizes empirical evidence examining the effects of GenAI on measurable CT outcomes in formal educational settings. Following PRISMA 2020 guidelines, the review searched Web of Science, ScienceDirect, and Google Scholar from database inception to 20th July 2026; 23 peer-reviewed empirical studies met the inclusion criteria. A direction-of-effect synthesis revealed a conditional pattern: most studies reported positive effects when GenAI was embedded in scaffolded, inquiry-driven instructional designs, while a smaller subset reported negative effects associated with cognitive offloading and automation bias; several studies reported mixed or context-dependent outcomes. Effects were moderated by instructional design (scaffolding and prompt engineering), learner prior knowledge and AI literacy, and task complexity. These results suggest that GenAI can support CT when integrated with pedagogical safeguards (verification tasks, formative feedback, and AI-literacy training). Heterogeneity in operational definitions and short intervention durations limits generalisability; we therefore recommend longitudinal RCTs employing validated, performance-based CT measures. This review contributes a mechanism-based interpretation of GenAI's effects and identifies practical and methodological priorities for research and practice.

Keywords: generative artificial intelligence; critical thinking; higher-order thinking; large language models; cognitive offloading; systematic review

INTRODUCTION

Critical thinking—the disciplined analysis, evaluation, and synthesis of information to form reasoned judgments—is a foundational aim of higher education and a core 21st-century competency [1], [2]. International policy frameworks emphasize their centrality for adaptive expertise in complex, digitally mediated societies [3], [4], [5]. The recent, rapid diffusion of generative artificial intelligence (GenAI) (e.g., large language models) into teaching, assessment, and research practices has therefore raised urgent questions about how these tools influence students' development of evaluative judgment. GenAI can produce fluent explanations, propose arguments, and structure solutions at scale—capabilities that offer clear pedagogical opportunities but also risk enabling superficial reliance on machine outputs if not integrated with appropriate pedagogical support [3], [6].

Empirical research to date paints a conditional portrait: GenAI can both support and undermine higher-order cognition depending on design features, learner characteristics, and measurement choices. Several syntheses report overall positive effects of GenAI on higher-order thinking (HOT), especially problem solving, while noting smaller and more variable effects for narrowly defined critical thinking outcomes [7], [8]. Classroom and quasi-experimental studies show that when GenAI is used as a scaffold—through guided prompts, Socratic questioning, or iterative feedback loops—students display improved analytic performance, richer argumentative structure, and greater task engagement [9], [10], [11]. These studies suggest that AI can function as a

"knowledgeable other" that prompts reflection and idea refinement when learners remain engaged and when tasks require evaluation rather than rote reproduction.

At the same time, a set of robust findings signals caution. Field surveys and controlled investigations report patterns of automation bias and cognitive offloading: learners sometimes accept fluent AI outputs without sufficient verification, use generated text as ready-made solutions, and reduce monitoring efforts—behaviors associated with lower independent reasoning on performance tests [12], [13]. Qualitative and mixed-methods work likewise documents "metacognitive laziness"—declines in monitoring and source critique—when students rely on AI to do the cognitive heavy lifting [14], [15]. Taken together, these results indicate heterogeneous effects: GenAI amplifies HOT when instruction activates evaluative engagement, but it attenuates CT when learners outsource verification and reflective work.

A clear theoretical account helps reconcile these divergent findings. First, from a dual-process perspective, GenAI's fluency tends to activate heuristic, fast responses (System 1) unless instructional design deliberately triggers slow, analytic processing (System 2) through prompts that require justification and verification [16]. Second, cognitive load theory explains a dual potential: AI can reduce extraneous load by organizing information, thereby freeing capacity for germane processing; conversely, excessive offloading can erode the effortful processing necessary for schema construction and transferable reasoning [17]. Third, metacognitive regulation—students' monitoring, evaluation, and revision of outputs—operates as the proximate moderator: learners who habitually verify and revise AI suggestions treat GenAI as a scaffold and show HOT gains, whereas passive users fall prey to automation bias and shallow processing [18], [19]. Empirical moderator analyses support these mechanisms: interventions that incorporate verification prompts, require explicated reasoning, or combine AI tasks with SRL (self-regulated learning) training yield larger and more persistent HOT effects [7], [20].

A second major impediment to cumulative understanding is conceptual and measurement heterogeneity. Critical thinking (CT) and higher-order thinking (HOT) are related but distinct constructs: CT refers specifically to evaluative reasoning—interpretation, analysis, evaluation, and inference—whereas HOT is a broader category that includes problem solving and synthesis [1]. Many GenAI studies conflate these constructs or fail to specify which dimension they measure, which complicates synthesis. Equally problematic is instrument variance: research employs self-reports, course grades, bespoke tasks, and a few validated CT instruments inconsistently, making it difficult to determine whether reported gains reflect deeper evaluative skill or surface improvements in task performance [7], [8]. Notably, studies that use validated, performance-based CT measures or rubric-scored argumentative tasks tend to provide more conservative estimates of GenAI's benefit than those relying on perceptions or convenience metrics [18], [21].

Methodological shortcomings amplify these conceptual problems. Much of the literature comprises short-term classroom pilots, small samples, or quasi-experimental designs without active controls or blinded scoring, and longitudinal evidence on sustained effects is scarce [22], [23]. Moreover, contextual moderators—disciplinary practices, assessment regimes, cultural dispositions toward authority and technology—are rarely examined despite their plausible influence on whether AI becomes a scaffold or a crutch. As a result, policy recommendations are often based on fragmented or non-generalizable evidence, while universities continue to redesign assessments and academic-integrity policies in response to GenAI's rapid adoption [3], [5].

These observations point to four concise, actionable gaps that this review addresses. Despite rapid growth in GenAI research, (1) definitions and operationalizations of CT are inconsistent; (2) many studies rely on short-term, perception-based, or non-validated outcome measures rather than standardized CT assessments; (3) there is limited experimental work that manipulates pedagogical scaffolds (e.g., verification prompts, required justification) to test mechanism-level hypotheses about offloading versus scaffolding; and (4) longitudinal and cross-contextual evidence about the durability and generalizability of effects is scarce. These gaps obscure the conditions under which GenAI supports evaluative judgment and limit the field's ability to offer evidence-based pedagogical guidance.

In light of these findings, the present systematic review adopts an approach that is both mechanism-informed and measurement-sensitive. Specifically, this review integrates existing studies that examine the effects of GenAI on HOTs and CTs, with particular focus on performance-based measures or studies that report relevant psychometric/scoring information. Studies were coded for instructional features (e.g., instructional scaffolding, verification demands) and learner variables (e.g., SRL, domain knowledge); in addition, studies addressing proposed cognitive mechanisms (e.g., offloading, activation of heuristics, and metacognitive regulation) were examined. By distinguishing between CTs and HOTs, as well as focusing on measurement-sensitive studies, this review aims to clarify the circumstances in which GenAI acts as a cognitive amplifier versus a cognitive substitute.

In short, this review seeks to offer educators and policymakers not only an answer to the question of whether GenAI influences learning, but how, for whom, and under what instructional conditions it promotes disciplined evaluative reasoning. This focus on mechanism and measurement is essential for designing AI-integrated curricula and assessments that preserve and strengthen students' capacity for critical thinking in an era of powerful generative technologies.

METHODS

This systematic review was conducted to synthesize empirical evidence examining the effects of generative artificial intelligence (GenAI) on critical thinking in educational contexts. The review followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 guidelines [24] to ensure methodological transparency and reproducibility. The review process comprised a structured sequence of database searching, title and abstract screening, full-text screening, data extraction, data synthesis, and reporting of results.

2.1 Database Search Strategy

Database searches were conducted up to 20th July 2026 using Web of Science, ScienceDirect, and Google Scholar, three databases selected for their comprehensive coverage of peer-reviewed education research, learning sciences, and educational technology. Given the emergent nature of generative AI tools in educational research, no restrictions were placed on publication year.

The search strategy was designed to identify studies located at the intersection of generative artificial intelligence and critical thinking. Searches were conducted across titles, abstracts, and author-provided keywords. To enhance readability and avoid excessive technical detail in the main text, the full list of search terms and Boolean combinations applied in each database is provided in Appendix A.

Only peer-reviewed journal articles published in English were included, reflecting feasibility considerations related to reviewer language proficiency.

2.2 Screening Strategy

All records retrieved from the database searches were imported into the Rayyan platform to facilitate duplicate removal and systematic screening. Following duplicate removal, title and abstract screening were conducted by the author according to predefined eligibility criteria. Studies that met the inclusion criteria, or for which eligibility could not be determined at the abstract level, were retained for full-text screening. Full texts were subsequently retrieved and assessed by the author based on the predefined eligibility criteria. Rayyan was used solely as an organizational and screening management tool, and no AI-assisted screening or automated study-selection procedures were employed at any stage of the review process. The study selection process and the number of records included and excluded at each stage are presented in the PRISMA 2020 flow diagram (Figure 1).

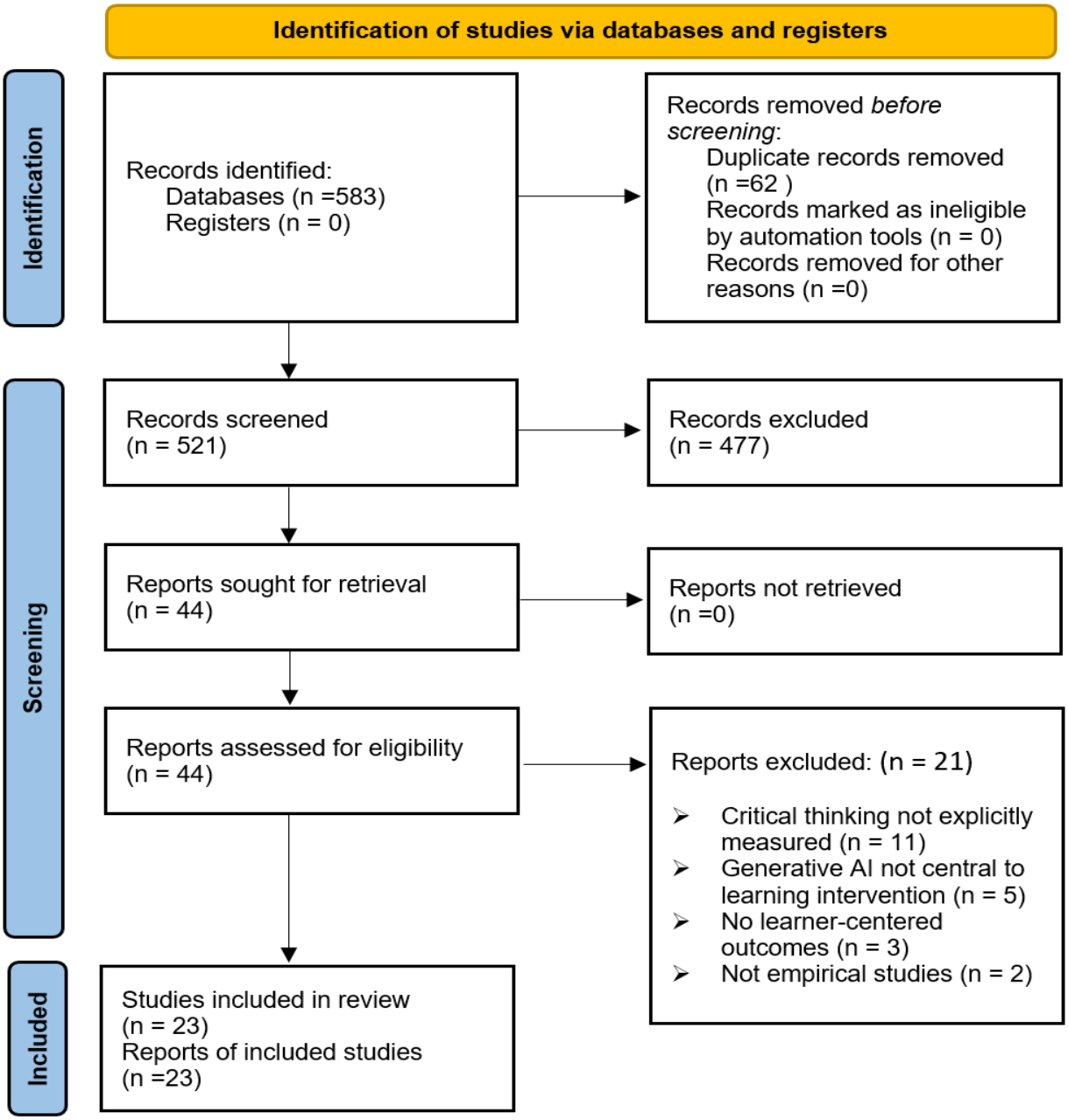


Figure 1: PRISMA 2020 Flow Diagram

2.3 Eligibility Criteria

Eligibility criteria were defined a priori to ensure the inclusion of studies providing direct empirical evidence relevant to the research question. Studies were assessed based on their topical focus, use of generative AI, educational context, population, outcome measures, study design, and type of evidence.

To be included, studies were required to:

1. Explicitly measure critical thinking as a core construct, or closely related higher-order cognitive processes such as analysis, evaluation, reflective judgment, and argumentation when these were directly linked to evaluative reasoning and judgment;

2. Examine the use of generative AI technologies within an educational or learning context;

3. Involve students, learners, or educators;

4. Report empirical findings derived from quantitative, qualitative, or mixed-methods research designs.

Studies focusing on perceptions, adoption, ethical considerations, or theoretical discussions without empirical measurement of critical thinking were excluded.

Table 1. Inclusion and exclusion criteria

Category	Inclusion criteria	Exclusion criteria
Topic focus	Critical thinking explicitly measured as a core construct, or closely related higher-order cognitive processes linked to analysis, evaluation, and reflective judgment	Critical thinking is not measured, only mentioned superficially, or limited to general creativity without critical evaluation
AI focus	Use of generative AI or ChatGPT in an educational or learning context	Studies focusing on AI ethics, adoption, perceptions, or opinions
Population	Students, learners, or educators across educational domains	Non-educational or non-learner-centered populations
Outcomes	Learning outcomes, critical thinking, or higher-order cognitive skills measured through empirical assessment	No measurable cognitive or learning outcomes
Study type	Empirical studies (quantitative, qualitative, mixed methods, RCTs, quasi-experimental, surveys)	Conceptual, theoretical, narrative, or non-empirical studies
Evidence	Use of measurable or validated assessment tools	Purely descriptive or speculative discussions

2.4 Data Extraction

Data extraction was performed by the author using a predefined data extraction framework. Extracted data included bibliographic information, educational context and participant characteristics, study design, type of generative AI tool, critical thinking framework or assessment instrument, and key findings related to critical thinking outcomes. The extracted data were organised in a spreadsheet to support systematic comparison and synthesis across studies. An overview of the characteristics of the included studies is presented in the Results section (Table 2), with a more detailed data extraction summary provided in Appendix C.

2.5 Data Analysis

Data synthesis was conducted using a direction-of-effect approach because the included studies were methodologically heterogeneous, making statistical pooling inappropriate. The studies varied considerably in design, including randomized controlled trials, quasi-experimental studies, qualitative studies, and mixed-methods research, as well as in the operationalization of critical thinking, intervention types, and educational contexts. For this reason, a meta-analysis was not considered methodologically suitable. The reported effects of generative AI on critical thinking were categorized as positive, negative, or mixed based on the direction and consistency of the findings. This approach enabled comparison across diverse studies while avoiding inappropriate statistical aggregation. Patterns in the direction of effects were then examined to identify overall trends and areas of inconsistency within the literature.

2.6 Methodological Considerations

Given the emergent nature of generative AI research in education and the methodological diversity of the included studies, no formal standardised quality appraisal tool was applied. Instead, methodological rigour was assessed descriptively during synthesis, with attention to study design, sample characteristics, and the validity of critical thinking measures. This approach was used because the included studies varied substantially in design

and outcome measurement, making a single quality scoring framework difficult to apply consistently. Nevertheless, studies with weaker internal validity were interpreted with caution in the discussion of findings.

RESULTS AND DISCUSSION

3.1 Study Demographics and Characteristics

Methodologically, the studies included in this review have shown substantial diversity, with a strong focus on empirical study designs. The evidence base is dominated by quasi-experimental, mixed-method studies, and a few experimental, correlational, and qualitative studies. This distribution reflects a concerted effort in the literature to empirically investigate the impact of generative artificial intelligence on critical thinking. Detailed information on the pattern of study designs is shown in Figure 2, with further methodological characteristics shown in Table 2 and Appendix B.

Table 2. Characteristics of the included studies

Study	Population	Discipline	CT Dimensions Assessed	Key Outcomes
[25]	Nursing Undergraduate	Geriatric Nursing	Reflective judgment, Credibility	Significant improvement in CT tendencies
[26]	Postgraduate	ICT Research	Facione's 5 dimensions	Significant gains in all 5 CT dimensions
[27]	Graduate	Science Ed	Accuracy, Synthesis	High rubric scores in synthesis and evaluation
[28]	Undergraduate	EFL	Analysis, Evaluation	Validated AI-powered inquiry tool
[29]	Graduate	Ed Tech	Sense-making, Innovation	SKI-based ChatGPT enhanced higher-order thinking
[30]	Graduate	Academic Writing	Confidence, Valuing CT	AI transformation positively predicted CT
[31]	Nursing Undergraduate	Nursing PBL	Analytical reasoning (CCTST)	LLM-PBL significantly outperformed traditional PBL
[32]	Undergraduate	Engineering	Reflective judgment, Evaluation	Large effect size (f^2=0.703) for CT
[33]	Undergraduate	STEM	Truth seeking, Analyticity	Control group outperformed AI group (Offloading)
[34]	Nursing Undergraduate	Pediatric Nursing	Clinical reasoning, Reflection	Control group outperformed AI group (Offloading)
[15]	Med Undergraduate/Post	Medical Ed	Deductive/Inductive reasoning	Significant increase in CCTT scores
[35]	High School	Text Analysis	Depth of analysis, Perspectives	Improvement dependent on prior knowledge
[36]	Undergraduate	Organic Chem	Credibility evaluation	Enhanced evidence-based reasoning
[37]	Med Undergraduate	Medical English	Argumentation, Ethical reasoning	Hybrid feedback enhanced writing & CT

[38]	Undergraduate	Animation	Creativity, Problem-solving	Positive impact on innovative thinking
[39]	Undergraduate	Accounting	Critical evaluation	Reperformance framework improved analytical skills
[40]	Undergraduate	Business Admin	Higher-order thinking	AI-aware colloquial assessment reliability
[41]	Undergraduate	General Univ	Analytical reasoning	ChatGPT interaction promoted complex CT
[42]	Undergraduate	Ed Tech	Facione's 6 dimensions	Fostered CT through multimodal composition
[43]	Vocational Undergraduate	EFL Writing	Inference, Argument evaluation	Gains limited to language construction
[44]	Undergraduate	EFL	Critical openness, Skepticism	Significant gains in CT for ChatGPT group
[45]	Undergraduate	EFL	Facione's dimensions	Positive correlation between AI writing & CT
[46]	Graduate	Ed Tech	CT self-reported	Significant improvement in CT subscales

Participant populations were mainly represented from higher education, with undergraduate students being part of 15 studies (65.22%). Postgraduate students were represented in 3 studies (13.04%) and alongside undergraduates in 2 studies (8.70%), while secondary and vocational education settings were explored in one study each. The disciplinary scope was broad, spanning nursing and medical education [15], [25], [31], EFL [28], [44], [45], STEM and engineering studies [32], [36], to business education [40]. In terms of geographical distribution, the majority of studies were conducted in Asia, with a focus on China and Taiwan, and further contributions from North America and Europe.

In the literature, ChatGPT was the most commonly investigated GenAI tool [25], [26], [30], together with other large language models and dedicated platforms such as Doubao, DeepSeek, QuillBot, and Duolingo.

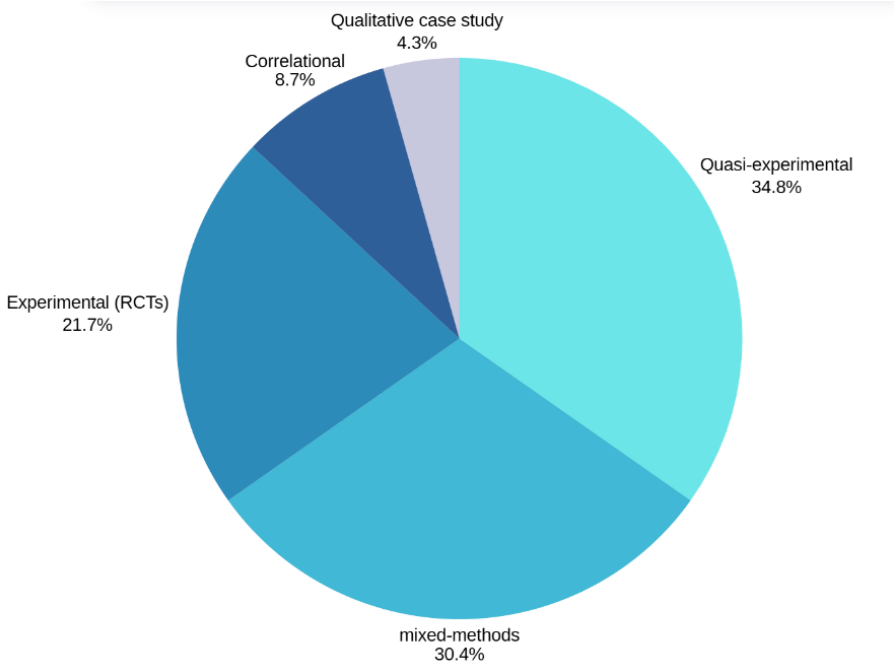


Figure 2: Distribution of Study Designs Across Included Studies (n = 23)

3.2 Studies Reporting Positive Effects of Generative AI on Critical Thinking

The synthesis of the included studies reveals that generative artificial intelligence (GenAI) appeared to support the development of critical thinking (CT) when it is embedded in a structured pedagogical framework. Rather than acting as a simple information retrieval tool, GenAI was frequently associated with enhancements in higher-order cognitive skills when used as a reasoning scaffold, a source of formative feedback, or an inquiry problem-solving catalyst. These positive outcomes tended to manifest across several core dimensions of CT, especially in analysis, evaluation, and reflective judgment.

Scaffolding Reasoning and Cognitive Engagement

A primary pathway through which GenAI supported CT was by providing scaffolding that reduced extraneous cognitive load, thereby allowing learners to focus on more complex analytical tasks. For instance, Hwang et al. (2025) observed that a GenAI-guided prompt-based learning approach appeared to enhance students' CT tendencies by providing a structured environment for designing nursing activities. By using well-defined prompts to guide the AI, students were able to engage in deeper sense-making and innovative thinking. This mechanism was further elucidated by Lee and Wu (2025), who found that scaffolded ChatGPT interactions were associated with higher levels of cognitive engagement in online discussions. In these contexts, the AI served as a "cognitive partner," prompting students to clarify their reasoning and consider alternative perspectives, which in turn appeared to foster more rigorous evaluative processes.

These findings suggest that GenAI can function as a scaffold rather than only as an information source. This view is in line with research that stresses the importance of well-structured implementation of AI. It also extends Cognitive Load Theory and scaffolding theory, as these findings show that GenAI can reduce cognitive load and free working memory to engage in more in-depth thinking [17], [25], [47], [48].

Supporting Critical Inquiry and Question Generation

GenAI tools also demonstrated potential in enhancing students' inquiry skills, which are widely regarded as a fundamental gateway to CT. Several studies indicated that AI-powered platforms could effectively guide students in formulating higher-order questions. Lai et al. (2025) utilized an AI-supported tool to provide real-time feedback on students' inquiries based on Bloom's Taxonomy, which was associated with a significant improvement in the cognitive complexity of their questions. Similarly, the framework proposed by Yusuf et al. (2024) for synthesizing AI-generated texts encouraged students to move beyond surface-level acceptance toward a more skeptical and inquisitive stance. By requiring students to interrogate AI outputs through a multi-phase process of conceptualizing and inquiring, these instructional designs appeared to cultivate a habit of critical evaluation and evidentiary verification.

The role of GenAI in facilitating inquiry and metacognitive thinking aligns with inquiry-based learning and metacognitive regulation. The role of GenAI in encouraging inquiring thinking and metacognitive evaluation of the content produced by GenAI can facilitate metacognitive thinking [26], [28], [49]. These discussions point to GenAI as a learning technology whose effectiveness depends on critical integration.

Enhancing Evaluative Judgment through Feedback and Synthesis

The capacity of GenAI to provide immediate, formative feedback was another key feature associated with positive CT outcomes. In language learning and academic writing contexts, tools like Doubao and QuillBot were often used to support the synthesis of complex information. Pan et al. (2025) found that GenAI-supported feedback in medical English courses appeared to enhance higher-order reasoning dimensions, such as thesis support and ethical engagement, by allowing students to iterate on their arguments more frequently. Furthermore, the study by Jin et al. (2025) suggested that using AI for text transformation, where students actively restructure and refine AI-generated content, was a positive predictor of CT skills. This indicates that the active, critical manipulation of AI outputs, rather than passive consumption, is a critical condition for skill development.

Overall, GenAI appears most effective when embedded in inquiry-driven, scaffolded learning environments that emphasize active evaluation and reflective synthesis, positioning learners as the primary agents of their own reasoning processes. These observations are consistent with research that stresses the importance of well-structured implementation of AI and with the view that GenAI can reduce cognitive load and support deeper thinking [25], [47], [50].

Instructional Conditions for Success

The review indicates that the positive effects of generative artificial intelligence (GenAI) on critical thinking (CT) are highly contingent upon specific instructional design features. Structured scaffolding emerged as a key factor, as studies integrating clear prompt-engineering guidance or established pedagogical models such as Problem-Based Learning reported stronger CT gains [31], [51]. Prior domain knowledge also functioned as an important moderator, with GenAI proving more effective for learners capable of evaluating the accuracy and relevance of AI-generated outputs [35]. In addition, the promotion of AI literacy, particularly the ability to critically verify AI responses, was consistently associated with positive outcomes [32], [41].

These findings extend existing debates showing that GenAI, when well integrated, can facilitate cognitive augmentation. The emphasis on verification and critique points to the significance of AI literacy as a core competence [52], [53]. GenAI appears most effective when its use requires active evaluation, reflective synthesis, and verification rather than passive acceptance of machine-generated text.

3.3 Studies Reporting Negative Effects of Generative AI on Critical Thinking

Although the literature on the topic tends to emphasize the potential benefits of employing generative artificial intelligence (GenAI) in various contexts, a small group of studies revealed a negative relationship between the application of GenAI and critical thinking (CT) skills. This was explained as a function of cognitive offloading, where the convenience of AI-generated outputs appeared to bypass the effortful reasoning processes required for deep analysis and evaluation.

A prominent mechanism observed was the reduction in active reasoning when students utilized GenAI without sufficient instructional constraints. For instance, Ji et al. (2025) found that a control group engaged in traditional collaborative learning outperformed a ChatGPT-assisted group in post-test CT scores. The author suggested that the immediate availability of AI-generated solutions may have led to intellectual laziness, where students accepted AI outputs without rigorous verification or internal deliberation. This phenomenon was similarly noted by Shin et al. (2024) in a nursing education context, where students using ChatGPT for case-based learning achieved significantly lower scores in critical thinking and clinical reasoning compared to those using traditional textbooks. In this instance, the AI appeared to function as a shortcut that hindered the development of the evidence-based knowledge integration essential for professional judgment.

These negative effects were particularly evident in tasks requiring high levels of originality and procedural rigor. When GenAI was used as a primary source of information rather than a scaffold, it was associated with a decrease in the depth of inquiry. The tendency for students to exhibit blind trust in AI outputs, often failing to scrutinize hallucinations or biases, was identified as a critical barrier to CT [33]. Collectively, these studies suggest that in the absence of explicit pedagogical interventions that mandate verification and critique, GenAI use may inadvertently discourage the very analytical habits it is intended to support, particularly among learners who lack the metacognitive skills to manage AI interactions effectively.

Further, engagement with the literature on cognitive offloading and automation bias suggests that the risk is not inherent but one of unregulated integration. The difference lies in whether AI replaces thought or enables intellectual engagement [33].

3.4 Studies Reporting Mixed or Conditional Effects

A substantial proportion of the reviewed studies reported mixed or context-dependent effects, indicating that the influence of generative artificial intelligence (GenAI) on critical thinking (CT) is not uniform but mediated by

several moderating factors. These studies highlight the importance of learner prior knowledge, task complexity, and instructional design in shaping CT outcomes.

A notable conditional pattern aligns with the "Matthew Effect" in AI-supported learning, whereby higher-performing learners benefit more from GenAI than their lower-performing peers. Li et al. (2025) found that a GenAI-assisted prompt framework significantly enhanced CT only among students with sufficient domain knowledge, while novice learners showed no significant gains, suggesting that GenAI functions as a knowledge amplifier rather than a foundational instructional tool. Task-related factors further moderated outcomes. Gao et al. (2025) reported no overall CT differences between GenAI-supported and control groups in a vocational writing context, and Xu and Liu (2025) similarly found no significant CT differences between ChatGPT and Duolingo interventions.

These findings support the idea that the impact of GenAI depends on learners' cognitive schema. Knowledgeable learners are better able to assess AI outcomes, while novices are more likely to become dependent [35]. GenAI could be seen as a tool that acts as a knowledge amplifier. Without prior knowledge, critical inquiry is likely to turn into dependency. The impact of GenAI is highly dependent on learners' cognitive schema.

The evidence does not suggest that there is a particular discipline that is more likely to produce positive outcomes. Rather, it is the pedagogy that seems to influence outcomes. Discipline is not as important as how GenAI is used within a particular task. These results also support the expertise reversal effect and schema theory, which help explain why the same tool may produce different effects depending on learner readiness and instructional design [54], [55].

3.5 Overall Synthesis, Limitations, and Implications

The synthesis of 23 studies reveals that there is a conditional relationship between Generative Artificial Intelligence (GenAI) and critical thinking (CT). Three findings can be highlighted. First, most studies suggest that the impact of GenAI can enhance critical thinking in educational settings. The enhancement of critical thinking was found in studies in which GenAI was used in structured pedagogical conditions with active student engagement. Second, the impact of GenAI in facilitating critical thinking is highly dependent on active learner engagement. The impact is more evident in conditions in which students go beyond passive consumption of information generated by GenAI. Such active scrutiny is a vital component of critical thinking development. Third, a small number of studies point to potential negative effects associated with cognitive offloading and excessive reliance on AI outcomes. Overall, this body of research supports the conditional nature of GenAI outcomes on critical thinking. The present systematic review offers an empirically informed synthesis of the impact of Generative Artificial Intelligence (GenAI) on critical thinking under different pedagogical conditions, clarifying the conditional nature of the effect of GenAI on CT.

The impact of GenAI on critical thinking appears conditional on factors other than academic discipline, such as instruction, learners, and task factors. GenAI is more likely to be effective on critical thinking when it offers cognitive scaffolding and reasoning support under supportive instruction. Neutral or negative results are obtained when GenAI is used to perform routine or low-cognitive-demand tasks that automate processes without challenging higher-order cognition [43]. Another factor that influences GenAI is prior knowledge, supported by the "Matthew Effect," expertise reversal effect, and schema theory. The difference lies in whether AI replaces thought or enables intellectual engagement.

Several limitations of the review need to be considered. The majority of the included studies were cross-sectional or short-term in nature, and the long-term impact of GenAI on cognitive development still needs to be examined. The geographical distribution of the included studies was predominantly Asian. In addition, screening and data extraction were conducted by a single reviewer, which may increase the possibility of selection and extraction bias. The review was also limited by the absence of a formal risk-of-bias assessment and by the inclusion of Google Scholar, which may reduce the exact replicability of search results. The heterogeneity of the included studies meant that a meta-analysis was not possible, so the direction-of-effect synthesis was used to address this limitation.

This review provides several contributions to scholarship on GenAI and CT. The focus on empirical outcome provides an evidence-based synthesis across diverse educational contexts. The direction-of-effect method enabled nuanced interpretation of heterogeneous results, clarifying the conditional effects of GenAI. The current review makes a theoretical contribution by providing a mechanism-based interpretation grounded in cognitive scaffolding and metacognitive regulation and specifying the instructional conditions, student characteristics, and task demands under which GenAI may facilitate CT.

Future studies should investigate cognitive development using longitudinal approaches to examine long-term effects of sustained GenAI engagement. Experimental studies can manipulate scaffolding levels and assess effects on CT while controlling for cognitive offloading. The moderation effects of prior knowledge should be investigated to clarify differences found in prior studies. Studies on AI literacy should investigate effects of explicit training on CT. Comparative studies should clarify whether differences are due to discipline or instructional approach.

For educators and policymakers, GenAI incorporation would benefit from a focus on AI literacy and prompting for critical inquiry. Teacher education in AI literacy and incorporation models may aid implementation. Policy considerations should promote responsible incorporation aligned with pedagogical goals. The design of transformation-focused activities that require critical inquiry can enhance engagement.

Generative Artificial Intelligence holds considerable potential in shaping critical thinking. Its success relies not on the presence of GenAI technology but on its pedagogical incorporation. Thoughtful incorporation supported by prior knowledge and critical evaluation skills can aid critical thinking development. Uncritical reliance on GenAI can lead to cognitive offloading. GenAI as an inquiry tool rather than a source promotes critical engagement with complex informational environments.

CONCLUSION

This systematic review synthesised empirical evidence on the effects of generative artificial intelligence (GenAI) on critical thinking across 23 peer-reviewed studies. The overall pattern is conditional: GenAI may facilitate higher-order reasoning when pedagogically integrated as a scaffold that prompts learners to critique, verify, and iteratively refine ideas; conversely, GenAI can attenuate analytical engagement when used as a shortcut that substitutes for cognitive effort. We interpret these findings through a mechanism-based lens — GenAI reduces extraneous cognitive load and enables iterative feedback (cognitive augmentation), but also creates opportunities for automation bias and mechanised convergence when verification and metacognitive oversight are absent.

For practice, educators should design tasks that require active interrogation of AI outputs (verification prompts, counter-argumentation, and text-transformation activities), embed prompt-engineering and formative feedback to position AI as a cognitive partner, and teach AI literacy to promote epistemic vigilance. Policymakers and institutions should support teacher professional development and assessment frameworks that preserve human agency and assessment validity.

Methodologically, future research must prioritise longitudinal and randomized controlled studies using validated, performance-based critical thinking instruments, consistently report effect sizes and moderator analyses (e.g., prior knowledge, task complexity, intervention duration), and investigate the persistence of effects beyond immediate post-tests. Addressing these priorities will clarify when and how GenAI can be harnessed to augment, rather than erode, critical thinking. In sum, GenAI holds conditional promise for enhancing critical thinking — a promise that depends on intentional design, verification, and pedagogical alignment.

Author Contributions

Simane Khales conducted the study, performed screening and data extraction, carried out the analysis, and wrote the manuscript.

Funding

This research received no external funding.

Data Availability Statement

No new data were created or analyzed in this study.

Conflicts of Interest

The author declares no conflict of interest.

Declaration on Generative AI

Generative AI tools were used for language refinement and structural editing. All content was reviewed and approved by the author.

REFERENCES

1. P. Facione, Critical Thinking: What It Is and Why It Counts. Insight assessment, 2011.
2. R. Paul and L. Elder, The miniature guide to critical thinking concepts and tools, 8th edition. in Thinker's Guide Library. Lanham, Md: Rowman & Littlefield, 2020.
3. OECD, OECD Digital Education Outlook 2026: Exploring Effective Uses of Generative AI in Education. in OECD Digital Education Outlook. OECD Publishing, 2026. doi: 10.1787/062a7394-en.
4. Guidance for generative AI in education and research. UNESCO, 2023. doi: 10.54675/EWZM9535.
5. European Commission. Directorate General for Education, Youth, Sport and Culture., Digital education action plan 2021-2027: improving the provision of digital skills in education and training. LU: Publications Office, 2023. doi: 10.2766/149764.
6. C. Costa and M. Murphy, "Generative artificial intelligence in education: (what) are we thinking?," Learn. Media Technol., pp. 1–12, Jun. 2025, doi: 10.1080/17439884.2025.2518258.
7. Y. Zhao, Y. Yue, Z. Sun, Q. Jiang, and G. Li, "Does Generative Artificial Intelligence Improve Students' Higher-Order Thinking? A Meta-Analysis Based on 29 Experiments and Quasi-Experiments," J. Intell., vol. 13, no. 12, p. 160, Dec. 2025, doi: 10.3390/jintelligence13120160.
8. X. Liu and B. Zhong, "Integrating generative Artificial Intelligence into student learning: A systematic review from a TPACK perspective," Educ. Res. Rev., vol. 49, p. 100741, Nov. 2025, doi: 10.1016/j.edurev.2025.100741.
9. H. B. Essel, D. Vlachopoulos, A. B. Essuman, and J. O. Amankwa, "ChatGPT effects on cognitive skills of undergraduate students: Receiving instant responses from AI-based conversational large language models (LLMs)," Comput. Educ. Artif. Intell., vol. 6, p. 100198, Jun. 2024, doi: 10.1016/j.caeai.2023.100198.
10. J. Lee et al., "Socratic Mind: Impact of a Novel GenAI-Powered Assessment Tool on Student Learning and Higher-Order Thinking," 2025, arXiv. doi: 10.48550/ARXIV.2509.16262.
11. D. Foung, L. Lin, and J. Chen, "Reinventing assessments with ChatGPT and other online tools: Opportunities for GenAI-empowered assessment practices," Comput. Educ. Artif. Intell., vol. 6, p. 100250, Jun. 2024, doi: 10.1016/j.caeai.2024.100250.
12. M. Gerlich, "AI Tools in Society: Impacts on Cognitive Offloading and the Future of Critical Thinking," Societies, vol. 15, no. 1, p. 6, Jan. 2025, doi: 10.3390/soc15010006.
13. J. Tian and R. Zhang, "Learners' AI dependence and critical thinking: The psychological mechanism of fatigue and the social buffering role of AI literacy," Acta Psychol. (Amst.), vol. 260, p. 105725, Oct. 2025, doi: 10.1016/j.actpsy.2025.105725.
14. E. E. Oliveira and T. Pereira, "A New Generation? A Discussion on Deep Generative Models in Supply Chains," in Advances in Production Management Systems. Production Management Systems for Responsible Manufacturing, Service, and Logistics Futures, vol. 689, E. Alfnes, A. Romsdal, J. O. Strandhagen, G. Von Cieminski, and D. Romero, Eds., in IFIP Advances in Information and Communication Technology, vol. 689. , Cham: Springer Nature Switzerland, 2023, pp. 444–457. doi: 10.1007/978-3-031-43662-8_32.
15. J. Wang and W. Fan, "The effect of ChatGPT on students' learning performance, learning perception, and higher-order thinking: insights from a meta-analysis," Humanit. Soc. Sci. Commun., vol. 12, no. 1, p. 621, May 2025, doi: 10.1057/s41599-025-04787-y.

16. K. E. Stanovich and R. F. West, "Advancing the rationality debate," Behav. Brain Sci., vol. 23, no. 5, pp. 701–717, Oct. 2000, doi: 10.1017/S0140525X00623439.
17. J. Sweller and P. Chandler, "Why Some Material Is Difficult to Learn," Cogn. Instr., vol. 12, no. 3, pp. 185–233, Sep. 1994, doi: 10.1207/s1532690xci1203_1.
18. G. R. Lau, W. Y. Low, L. Tay, Y. Guevarra, D. Gašević, and A. Hartanto, "Understanding Critical Thinking in Generative Artificial Intelligence Use: Development, Validation, and Correlates of the Critical Thinking in AI Use Scale," Dec. 13, 2025, arXiv: arXiv:2512.12413. doi: 10.48550/arXiv.2512.12413.
19. Y. Guo, L. Huang, C. Zhang, Q. Li, and M. Chen, "ChatGPT in education: A systematic review of its impact on critical thinking skills and dispositions," Think. Ski. Creat., vol. 60, p. 102106, Jun. 2026, doi: 10.1016/j.tsc.2025.102106.
20. H.-P. (Hank) Lee et al., "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, Yokohama Japan: ACM, Apr. 2025, pp. 1–22. doi: 10.1145/3706598.3713778.
21. N. R. Nasr, C.-H. Tu, J. Werner, T. Bauer, C.-J. Yen, and L. Sujo-Montes, "Exploring the Impact of Generative AI ChatGPT on Critical Thinking in Higher Education: Passive AI-Directed Use or Human–AI Supported Collaboration?," Educ. Sci., vol. 15, no. 9, p. 1198, Sep. 2025, doi: 10.3390/educsci15091198.
22. W. K. Monib, A. Qazi, R. A. Apong, M. T. Azizan, L. De Silva, and H. Yassin, "Generative AI and future education: a review, theoretical validation, and authors' perspective on challenges and solutions," PeerJ Comput. Sci., vol. 10, p. e2105, Dec. 2024, doi: 10.7717/peerj-cs.2105.
23. P. P. Premkumar, M. R. K. N. Yatigammana, and S. Kannangara, "Impact of Generative AI on Critical Thinking Skills in Undergraduates: A Systematic Review," J. Desk Res. Rev. Anal., vol. 2, no. 1, pp. 199–215, Dec. 2024, doi: 10.4038/jdrra.v2i1.55.
24. M. J. Page et al., "The PRISMA 2020 statement: an updated guideline for reporting systematic reviews," BMJ, p. n71, Mar. 2021, doi: 10.1136/bmj.n71.
25. G.-J. Hwang, P.-Y. Cheng, and C.-Y. Chang, "Facilitating students' critical thinking, metacognition and problem-solving tendencies in geriatric nursing class: A mixed-method study," Nurse Educ. Pract., vol. 83, p. 104266, Feb. 2025, doi: 10.1016/j.nepr.2025.104266.
26. A. Yusuf, S. Bello, N. Pervin, and A. K. Tukur, "Implementing a proposed framework for enhancing critical thinking skills in synthesizing AI-generated texts," Think. Ski. Creat., vol. 53, p. 101619, Sep. 2024, doi: 10.1016/j.tsc.2024.101619.
27. R. Ganjoo, J. Rankin, B. Lee, and L. Schwartz, "Beyond boundaries: exploring a generative artificial intelligence assignment in graduate, online science courses," J. Microbiol. Biol. Educ., vol. 25, no. 3, pp. e00127-24, Dec. 2024, doi: 10.1128/jmbe.00127-24.
28. W. Y. W. Lai, P. Kim, and J. S. Lee, "Designing and validating an AI-supported tool for enhancing critical inquiry in EFL education," Res. Methods Appl. Linguist., vol. 4, no. 3, p. 100266, Dec. 2025, doi: 10.1016/j.rmal.2025.100266.
29. H.-Y. Lee and T.-T. Wu, "Enhancing blended learning discussions with a Scaffolded Knowledge Integration–Based ChatGPT mobile instant messaging system," Comput. Educ., vol. 237, p. 105375, Nov. 2025, doi: 10.1016/j.compedu.2025.105375.
30. F. Jin, C.-H. Lin, and C. Lai, "Modeling AI-assisted writing: How self-regulated learning influences writing outcomes," Comput. Hum. Behav., vol. 165, p. 108538, Apr. 2025, doi: 10.1016/j.chb.2024.108538.
31. S. Kejingyun and R. Mingjun, "Randomized Controlled Study on the Impact of Problem-Based Learning Combined With Large Language Models on Critical Thinking Skills in Nursing Students," Nurse Educ., vol. 50, no. 4, pp. 216–220, Jul. 2025, doi: 10.1097/NNE.0000000000001879.
32. Z. Zhou, H. Guo, F. Ma, C. Yang, and Y. Gao, "The chain mediating role of critical thinking and AI self-efficacy in GenAI usage competence and engineering students' creativity," Sci. Rep., vol. 15, no. 1, p. 35945, Oct. 2025, doi: 10.1038/s41598-025-21132-0.
33. Y. Ji, Z. Zhan, T. Li, X. Zou, and S. Lyu, "Human–Machine Cocreation: The Effects of ChatGPT on Students' Learning Performance, AI Awareness, Critical Thinking, and Cognitive Load in a STEM

Course Toward Entrepreneurship," IEEE Trans. Learn. Technol., vol. 18, pp. 402–415, 2025, doi: 10.1109/TLT.2025.3554584.
34. H. Shin, J. C. De Gagne, S. S. Kim, and M. Hong, "The Impact of Artificial Intelligence-Assisted Learning on Nursing Students' Ethical Decision-making and Clinical Reasoning in Pediatric Care: A Quasi-Experimental Study," CIN Comput. Inform. Nurs., vol. 42, no. 10, pp. 704–711, Oct. 2024, doi: 10.1097/CIN.0000000000001177.
35. X. Li et al., "Navigating the Textual Maze: Enhancing Textual Analytical Skills Through an Innovative GAI Prompt Framework," IEEE Trans. Learn. Technol., vol. 18, pp. 206–215, 2025, doi: 10.1109/TLT.2025.3539104.
36. J. Kim, "Integrating Artificial Intelligence (AI) Chatbots and Green Chemistry Principles in the Synthesis of Cyclohexene," J. Chem. Educ., vol. 102, no. 7, pp. 3058–3064, Jul. 2025, doi: 10.1021/acs.jchemed.5c00212.
37. Y. Pan, "Leveraging generative AI powered rubric-indexed feedback as a formative assessment strategy for enhancing medical English education," Discov. Comput., vol. 28, no. 1, p. 284, Nov. 2025, doi: 10.1007/s10791-025-09830-9.
38. X. Yao, Y. Zhong, and W. Cao, "The analysis of generative artificial intelligence technology for innovative thinking and strategies in animation teaching," Sci. Rep., vol. 15, no. 1, p. 18618, May 2025, doi: 10.1038/s41598-025-03805-y.
39. D. Wiseman and D. Foster, "Bridging AI with cost accounting education: A pedagogical framework using reperformance and critical error review," J. Account. Educ., vol. 72, p. 100988, Sep. 2025, doi: 10.1016/j.jaccedu.2025.100988.
40. C. Kooli, N. Yusuf, and M. Y. Sarhan, "Colloquial engagement theory with AI awareness (CET-AIA): A new creative pedagogical framework for ethical assessment in the age of artificial intelligence," Think. Ski. Creat., vol. 60, p. 102051, Jun. 2026, doi: 10.1016/j.tsc.2025.102051.
41. R. Suriano, A. Plebe, A. Acciai, and R. A. Fabio, "Student interaction with ChatGPT can promote complex critical thinking skills," Learn. Instr., vol. 95, p. 102011, Feb. 2025, doi: 10.1016/j.learninstruc.2024.102011.
42. Y. Shen and Y. Han, "Fostering undergraduates' critical thinking in digital multimodal composition with generative artificial intelligence," Think. Ski. Creat., vol. 59, p. 102045, Mar. 2026, doi: 10.1016/j.tsc.2025.102045.
43. J. Gao, J. Zhang, and Y. Li, "Do AI chatbot-integrated writing tasks influence writing self-efficacy and critical thinking ability? An exploratory study," Comput. Educ. Artif. Intell., vol. 9, p. 100472, Dec. 2025, doi: 10.1016/j.caeai.2025.100472.
44. J. Xu and Q. Liu, "Uncurtaining windows of motivation, enjoyment, critical thinking, and autonomy in AI-integrated education: Duolingo Vs. ChatGPT," Learn. Motiv., vol. 89, p. 102100, Feb. 2025, doi: 10.1016/j.lmot.2025.102100.
45. X. Shen and M. F. Teng, "Three-wave cross-lagged model on the correlations between critical thinking skills, self-directed learning competency and AI-assisted writing," Think. Ski. Creat., vol. 52, p. 101524, Jun. 2024, doi: 10.1016/j.tsc.2024.101524.
46. Y.-M. Huang, P.-H. Chen, H.-Y. Lee, F. E. Sandnes, and T.-T. Wu, "ChatGPT-enhanced mobile instant messaging in online learning: Effects on student outcomes and perceptions," Comput. Hum. Behav., vol. 168, p. 108659, Jul. 2025, doi: 10.1016/j.chb.2025.108659.
47. J. Sweller, "Cognitive Load During Problem Solving: Effects on Learning," Cogn. Sci., vol. 12, no. 2, pp. 257–285, Apr. 1988, doi: 10.1207/s15516709cog1202_4.
48. Vygotsky, L. S and Cole, M, Mind in society: Development of higher psychological processes. Harvard university press. 1978.
49. J. H. Flavell, "Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry.," Am. Psychol., vol. 34, no. 10, pp. 906–911, Oct. 1979, doi: 10.1037/0003-066X.34.10.906.
50. C. K. Y. Chan and W. Hu, "Students' voices on generative AI: perceptions, benefits, and challenges in higher education," Int. J. Educ. Technol. High. Educ., vol. 20, no. 1, p. 43, Jul. 2023, doi: 10.1186/s41239-023-00411-8.
51. W. Shalong et al., "Enhancing self-directed learning with custom GPT AI facilitation among medical students: A randomized controlled trial," Med. Teach., vol. 47, no. 7, pp. 1126–1133, Jul. 2025, doi: 10.1080/0142159X.2024.2413023.

52. P. A. Kirschner and P. De Bruyckere, "The myths of the digital native and the multitasker," Teach. Teach. Educ., vol. 67, pp. 135–142, Oct. 2017, doi: 10.1016/j.tate.2017.06.001.
53. D. T. K. Ng, C. W. Tan, and J. K. L. Leung, "Empowering student self-regulated learning and science education through CHATGPT : A pioneering pilot study," Br. J. Educ. Technol., vol. 55, no. 4, pp. 1328–1353, Jul. 2024, doi: 10.1111/bjet.13454.
54. S. Kalyuga, P. Ayres, P. Chandler, and J. Sweller, "The Expertise Reversal Effect," Educ. Psychol., vol. 38, no. 1, pp. 23–31, Jan. 2003, doi: 10.1207/S15326985EP3801_4.
55. Rumelhart, D. E., & Ortony, A, The Representation of Knowledge in Memory 1. 2017. [Online]. Available: https://www.taylorfrancis.com/chapters/edit/10.4324/9781315271644-10/representation-knowledge-memory-1-david-rumelhart-andrew-ortony

Appendix A

List of terms used in Database search

Search Strategy and Terms

Sequence	Search terms
S1	"artificial intelligence" OR AI OR "generative artificial intelligence" OR "generative AI" OR GenAI OR "foundation models" OR "large generative models" OR "large language model" OR "large language models" OR LLM OR LLMs OR "autoregressive language models" OR ChatGPT OR "AI-generated text" OR "text generation" OR "content generation" OR "natural language generation"
S2	"critical thinking" OR "analytical thinking" OR "higher-order thinking" OR "critical reasoning" OR "problem solving"
S3	S1 AND S2
S4	Limit search S3: English language (due to feasibility considerations related to reviewer language proficiency); Publication type: peer-reviewed journal articles

Appendix B

Table of Articles Included in the Review

Study	**Country**	**Participants and Context**	**Design and Sample**	**AI Tool and CT Assessment**	**Key Findings Related to Critical Thinking**
[25]	Taiwan	First-year nursing undergraduates; Geriatric Nursing	Mixed-methods (N=56)	ChatGPT; 6-item Likert scale (CT tendency)	GenAI-guided prompt-based learning significantly improved critical thinking tendency scores compared to conventional technology-based learning.
Yusuf et al. (2024)[26]	Nigeria	Postgraduates; ICT in Educational Research	Mixed-methods experimental (N=304)	ChatGPT; Performance rubrics & CT Self-Assessment Scale (CTSAS)	The proposed framework significantly enhanced interpretation, analysis, and evaluation skills for synthesizing AI-generated texts through progressive training.

Ganjoo et al. (2024)[27]	USA	Science education graduates; Health Professions	Descriptive mixed-methods (N=49)	ChatGPT; Rubric-based performance assessment	Students successfully used AI for research support and demonstrated high levels of critical evaluation by identifying inaccuracies in AI-generated content.
Lai et al. (2025)[28]	Hong Kong	Year 3/4 undergraduates; EFL Teacher Education	Quasi-experimental (N=48)	Ask.SMILE (via ChatGPT API); Bloom's Taxonomy-based rubric	The AI-supported tool significantly improved student inquiry levels and demonstrated high reliability in evaluating higher-order cognitive questioning.
Lee et al. (2025)[29]	Taiwan	Masters students; Educational Technology	Randomized Controlled Trial (N=87)	ChatGPT (GPT-4o-mini); Cognitive engagement coding scheme	The scaffolded ChatGPT-MIM group showed significantly higher sense-making and innovation-level contributions compared to traditional messaging groups.
Jin et al. (2025)[30]	UK / International	Masters and PhD students; Academic Writing	Correlational SEM study (N=1073)	ChatGPT; Critical Thinking Toolkit (CriTT)	AI use for text transformation and development positively predicted critical thinking, whereas use for technical support showed no significant correlation.
Shi et al. (2025) [31]	China	First-year nursing undergraduates; Nursing PBL	Randomized Controlled Trial (N=100)	ChatGPT; California CT Skills Test (CCTST)	The LLM-integrated PBL group showed significantly greater overall improvement in critical thinking, particularly in inductive reasoning, than traditional PBL.
Zhou et al. (2025)[32]	China	Engineering undergraduates; Multi-disciplinary	Correlational PLS-SEM study (N=804)	General GenAI (e.g., ChatGPT); 6-item CT Scale	GenAI usage competence was found to be a strong predictor of critical thinking, which in turn significantly mediated the development of student creativity.
Ji et al. (2025)[33]	China	Sophomore undergraduates; STEM Entrepreneurship	Quasi-experimental (N=36)	ChatGPT (WiseGPT); Facione-based CT disposition scale	The control group outperformed the ChatGPT group in critical thinking, suggesting that overreliance on AI may reduce active reasoning and evaluation.
Shin et al.	S. Korea	Third-year nursing	Quasi-experimental (N=99)	ChatGPT v3.5; Performance-	Traditional textbook-based learning led to significantly higher critical thinking and clinical reasoning scores

(2024)[34]		undergraduates; Pediatric Nursing		based analytic rubric	than AI-assisted learning in this context.
Wang et al. (2024)[15]	China	Medical undergraduates/postgraduates; PBL	Randomized Controlled Trial (N=103)	Custom GPT (LearnGuide); Cornell CT Test (CCTT)	The AI-facilitated group showed significantly higher critical thinking scores after 12 weeks, with gains persisting through the 14-week follow-up period.
Li et al. (2025)[35]	Canada	Grade 12 high school students; Text Analysis	Controlled experimental (N=15)	ChatGPT (GPT-3.5); Expert-rated analytic scoring rubric	The prompt framework significantly improved critical thinking for students with prior background knowledge, but showed no statistical gain for novices.
Kim et al. (2025)[36]	USA	Sophomore chemistry undergraduates; Green Chemistry	Descriptive instructional (N=9)	ChatGPT, Gemini, Copilot; Self-report Likert survey	Most students reported that AI tools effectively supported their development of skills in critically evaluating the accuracy and feasibility of chemical approaches.
Pan et al. (2025) [37]	China	Medical undergraduates	Mixed-methods quasi-experimental (N=116)	Dou Bao (ByteDance); Writing rubrics & reflections	GenAI-supported feedback led to significant improvements in higher-order reasoning, including thesis support and ethical engagement in medical writing.
Yao et al. (2025)[38]	China	Undergraduates; Animation & Creative Arts	Mixed-methods quasi-experimental (N=120)	Customized GAI platform; Behavioral & task analysis	The GAI-integrated environment significantly enhanced problem-solving, creativity, and teamwork skills compared to traditional animation teaching methods.
Wiseman et al. (2025) [39]	USA	Undergraduates; Cost Accounting	Descriptive pedagogical (N=39)	ChatGPT, Copilot, Claude; Perceptual surveys	Students perceived an improved ability to evaluate AI outputs and perform error analysis, though objective critical thinking outcomes were not measured.
Kooli et al. (2026)[40]	Canada/KSA	Undergraduates; Business Administration	Quasi-experimental (N=229)	ChatGPT-3.5; Quiz scores & engagement proxies	Colloquial assessments using AI showed higher reliability and discrimination, helping to maintain academic integrity while fostering independent judgment.

Suriano et al. (2025)[41]	Italy	Undergraduates; General Higher Education	Correlational path analysis (N=181)	ChatGPT; CT Performance Test & Attitude Scale	Active engagement with ChatGPT significantly predicted critical thinking performance, with engagement serving as a stronger predictor than prior knowledge.
Shen et al. (2026)[42]	China	Sophomore undergraduates; Educational Technology	Qualitative case study (N=13)	DeepSeek, Doubao; Reflective writing & interviews	Qualitative analysis revealed increased scrutiny of AI content and growth across all six Facione CT dimensions through multimodal composition tasks.
Gao et al. (2025)[43]	China	Vocational first-year students; College English	Quasi-experimental (N=80)	Doubao (ByteDance); CT Ability Questionnaire	No significant difference was found between groups for overall critical thinking, though the AI group showed significant gains in the language construction dimension.
Xu et al. (2025)[44]	China	Young adult EFL learners; Language Institutes	True-experimental (N=244)	ChatGPT, Duolingo; Erawan CT Scale (CTS)	Both Duolingo and ChatGPT groups showed significant post-test gains in critical thinking compared to the control group, with no significant difference between tools.
Shen et al. (2024)[45]	China	Sophomore undergraduates; EFL Writing	Longitudinal cross-lagged study (N=204)	QuillBot; California CT Skills Test (CCTST)	Longitudinal analysis confirmed significant positive correlations between critical thinking skills, self-directed learning, and AI-assisted writing competence.
Huang et al. (2025)[46]	Taiwan	Masters students; Online Learning	Randomized Controlled Trial (N=63)	ChatGPT (GPT-3.5-turbo); Higher-Order Thinking Scale	The ChatGPT-enhanced mobile messaging group showed statistically significant improvements in critical thinking subscales compared to traditional messaging.