



Generative Artificial Intelligence and University Students' Critical and Higher-Order Thinking Outcomes: A PRISMA-Based Systematic Review and Thematic Analysis

Yunardi Kristian Zega^{1*}, Rini Sumanti Sapalakakai², Yuireishang Shithungdang³, Luiramsing Zimik⁴

¹Politeknik Negeri Kupang, Kupang, Indonesia

^{2,3,4}Hoseo University, Asan, Chungcheongnam-do, Republic of Korea

*Corresponding Author: yunardichristian@gmail.com

Abstract

The rapid integration of generative artificial intelligence tools such as ChatGPT, Gemini, and Copilot into higher education has raised urgent questions about their effects on students' higher-order cognitive abilities. While these tools promise unprecedented access to information and cognitive support, growing concerns have emerged regarding the potential erosion of critical thinking skills among university students who adopt them uncritically. Despite increasing scholarly attention to artificial intelligence in education, systematic evidence on how generative AI relates to distinct dimensions of higher-order thinking remains limited. This systematic literature review, guided by the PRISMA 2020 reporting guideline, aims to identify, appraise, and synthesize empirical evidence on the relationship between generative AI use and university students' analytical, evaluative, and creative thinking outcomes, using the higher-order cognitive processes of Bloom's Revised Taxonomy as an organizing framework. A comprehensive search across Scopus, Web of Science, ERIC, and Google Scholar yielded 847 initial records, of which 42 empirical studies published between 2020 and 2024 met the inclusion criteria after screening. Thematic synthesis revealed heterogeneous outcomes: analytical thinking outcomes were generally positive under structured prompting conditions, evaluative thinking appeared most vulnerable in the absence of deliberate pedagogical design, and creative thinking outcomes were most frequently reported as positive. Across themes, the quality of pedagogical integration was the condition most frequently associated with favorable outcomes. This review contributes a dimension-level synthesis of generative AI and higher-order thinking, identifies the pedagogical conditions associated with favorable and unfavorable outcomes, and documents the heterogeneity of outcomes across studies, offering an evidence-informed framework for educators, curriculum designers, and policymakers seeking to use generative AI as a cognitive scaffold rather than a cognitive replacement.

Keywords: critical thinking; generative artificial intelligence; higher education; PRISMA; systematic review



EDUCATUM is licensed under a [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by-sa/4.0/) (CC BY 4.0).

INTRODUCTION

Since the public release of ChatGPT in November 2022, the landscape of higher education has been fundamentally altered by the rapid proliferation of generative artificial intelligence (GenAI) tools (Wu et al., 2023). Soon afterwards, complementary platforms such as Google Gemini and Microsoft Copilot entered mainstream use. University students have adopted these tools quickly: in a survey of 399 undergraduate and postgraduate students from six universities in Hong Kong, Chan and Hu (2023) found that about two-thirds had used

generative AI at least once. Although such figures come from specific populations, countries, and survey periods and cannot be generalized to the global student population, they indicate that adoption has outpaced institutional policy development, creating a tension between the transformative potential of these technologies and the pedagogical risks they may introduce.

Critical thinking has long been recognized as a cornerstone of higher education and a defining competency for the twenty-first century. Major educational frameworks, including the Assessment and Teaching of 21st Century Skills (ATC21S; Griffin et al., 2012), the Partnership for 21st Century Learning (P21; Battelle for Kids, 2019), and the European Digital Competence Framework (DigComp 2.2; Vuorikari et al., 2022), position critical thinking as essential for navigating complex information environments, making informed decisions, and participating meaningfully in democratic societies. In this review, critical thinking is understood, following the expert consensus definition of Facione (1990), as purposeful, self-regulatory judgment that results in interpretation, analysis, evaluation, and inference. Bloom's Revised Taxonomy (Anderson & Krathwohl, 2001) does not offer a single definition of critical thinking; rather, it classifies cognitive processes, of which analyzing, evaluating, and creating constitute the higher-order levels. These three processes are therefore used as an organizing framework for higher-order thinking outcomes. Analysis and evaluation correspond most closely to the core of critical thinking, whereas creating is not synonymous with critical thinking; it is included as a related higher-order outcome because generative AI tools are used predominantly to produce content, which makes students' own generative and synthesizing capacities a central concern. Accordingly, findings on creative thinking are reported separately and are not treated as evidence of critical thinking per se.

The emergence of generative AI introduces a profound tension into this pedagogical landscape. On one hand, these tools can serve as cognitive scaffolding instruments that lower barriers to complex tasks, enabling students to engage with sophisticated analytical processes they might otherwise find inaccessible; in their position paper on large language models in education, Kasneci et al. (2023) argue that such benefits depend on learners developing the competencies needed to question and verify model outputs. Structured AI-assisted prompting can guide students through argument deconstruction, evidence evaluation, and creative ideation in ways that parallel expert mentoring (Mollick & Mollick, 2023). On the other hand, growing evidence suggests that generative AI may function as a cognitive offloading tool, wherein students delegate thinking processes to the machine rather than engaging in them independently (Deci & Ryan, 1985; Sweller et al., 2019). This delegation may gradually erode the very cognitive capacities that higher education seeks to develop.

The empirical evidence on this tension remains divided. Baidoo-Anu and Ansah (2023) reviewed the potential benefits of ChatGPT for teaching and learning, including personalized support and more efficient information processing. Zhai et al. (2023) reported that structured AI prompting improved students' ability to analyze scientific arguments in STEM disciplines. Lim et al. (2023), writing from the perspective of management educators, characterized generative AI as a paradox that may either reform or disrupt education.

Conversely, other scholars have sounded alarms. Stadler et al. (2024) reported dependency risks wherein frequent ChatGPT users showed diminished ability to independently evaluate source credibility, and Lin and Chang (2023) reported declines in source evaluation among students who relied on AI-generated content without explicit evaluation tasks. Cotton et al. (2024) highlighted concerns about academic integrity and the potential for AI to undermine authentic learning processes.

The existing body of systematic reviews on AI in education, while growing, leaves important questions open. Zhai et al. (2021) reviewed AI in education research published between 2010 and 2020, a period preceding the widespread availability of generative AI. Zawacki-Richter et al. (2019) conducted a comprehensive review of AI applications in higher education but focused on adaptive learning systems and intelligent tutoring rather than generative AI, which had not yet emerged as a mainstream tool. Crompton and Burke (2023) reviewed AI in higher education more broadly but did not differentiate between types of AI tools or disaggregate effects on specific dimensions of critical thinking. Lo (2023) provided an early rapid review of ChatGPT in education but focused primarily on usage patterns and perceptions rather than measured learning outcomes. Among the reviews identified in our search, none specifically examined how generative AI tools, as distinct from other AI technologies in education, relate to the separate higher-order thinking processes of analysis, evaluation, and creation.

This gap is consequential because the effects of generative AI on critical thinking are unlikely to be uniform across these dimensions. Analysis, evaluation, and creation engage different cognitive processes, draw on different epistemic dispositions, and may be differentially susceptible to AI-mediated scaffolding or offloading. A review that treats critical thinking as a monolithic construct risks obscuring these nuanced, dimension-specific effects and thereby limiting the practical utility of its findings for pedagogical design.

To address this gap, the present study conducts a systematic literature review (SLR) following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses 2020 guideline (PRISMA 2020; Page et al., 2021). The review employs thematic synthesis (Thomas & Harden, 2008) to identify and integrate evidence from empirical studies published between 2020 and 2024 that examine the relationship between generative AI use and university students' critical and higher-order thinking outcomes. The contribution of this review lies in three substantiated elements: (1) a dimension-level synthesis that reports analytical, evaluative, and creative thinking outcomes separately; (2) the identification of pedagogical and contextual conditions associated with favorable or unfavorable outcomes; and (3) a systematic account of the heterogeneity of outcomes, designs, and instruments across studies.

The research objectives of this systematic review are threefold: (1) to identify and synthesize empirical evidence on how generative AI use relates to university students' analytical, evaluative, and creative thinking outcomes; (2) to identify the pedagogical and contextual conditions associated with these outcomes; and (3) to map existing research gaps

and propose directions for future investigation. This study contributes to the growing body of evidence-based scholarship on AI in education by offering a dimension-specific, PRISMA-guided synthesis of evidence on generative AI and higher-order thinking in higher education, thereby offering actionable insights for educators, institutional leaders, and policymakers navigating the complex integration of these technologies.

METHODS

This study adopted a systematic literature review (SLR) design guided by the PRISMA 2020 reporting guideline (Page et al., 2021). PRISMA 2020 was selected for its transparency, supporting reproducibility and minimizing bias in the identification, screening, and synthesis of relevant studies. Because this study synthesized previously published literature and involved no human participants or personal data, ethical approval and informed consent were not required.

A comprehensive literature search was conducted across four major academic databases, Scopus, Web of Science, ERIC (Education Resources Information Center), and Google Scholar, covering publications from January 2020 through December 2024. The 2020 lower bound was chosen to capture early empirical work on large language model-based tools that preceded ChatGPT (e.g., GPT-3-based writing assistants), as well as the surge of research that followed the public release of ChatGPT in late 2022; pre-ChatGPT studies were retained only if the tool examined met the review's definition of generative AI, namely a system that generates novel text or other content in response to user prompts. The following Boolean search string was applied with database-specific adaptations: ("generative AI" OR "ChatGPT" OR "Gemini" OR "Copilot") AND ("critical thinking" OR "higher-order thinking" OR "analytical thinking") AND ("university" OR "higher education" OR "undergraduate"). Additional records were identified through backward citation searching of included studies and hand-searching of key journals in educational technology, including *Computers & Education*, *The Internet and Higher Education*, *British Journal of Educational Technology*, and *Education and Information Technologies*.

Studies were included if they met all of the following criteria: empirical research employing quantitative, qualitative, or mixed-methods designs; participants enrolled in university or college-level programs; at least one measured outcome that included critical thinking or its sub-dimensions (analysis, evaluation, creation); publication in peer-reviewed journals; and written in either English or Indonesian. Conversely, studies were excluded if they were non-empirical in nature, including reviews, opinion pieces, position papers, editorials, theoretical essays, and conference abstracts without full papers, focused on primary or secondary school contexts, examined AI tools other than generative AI such as adaptive learning systems or intelligent tutoring systems, or did not include any direct measurement or assessment of critical thinking skills or related constructs. Non-empirical publications were cited only as background literature and were not included in the synthesis. Studies that assessed critical thinking solely through students' self-reported perceptions were coded as

perception-based evidence and synthesized separately from studies that directly measured students' performance on critical thinking tasks.

The screening and selection process was guided by the PRISMA 2020 flow diagram. Initial database searches yielded 847 records. Following the removal of 213 duplicate entries using Rayyan, a web-based systematic review management tool (Ouzzani et al., 2016), 634 unique records remained for title and abstract screening. Two independent reviewers assessed each record against the inclusion criteria, resulting in the exclusion of 489 records that failed to meet the eligibility requirements. The remaining 145 articles proceeded to full-text review, during which an additional 103 articles were excluded for the following reasons: no direct measurement of critical thinking ($n = 41$), non-generative AI focus ($n = 28$), non-empirical design ($n = 19$), and non-higher-education context ($n = 15$). This process yielded a final sample of 42 studies included in the qualitative synthesis. The PRISMA 2020 flow diagram reports these numbers by database, together with the reasons for exclusion at the full-text stage.

The methodological quality of all included studies was assessed using two established appraisal tools. Quantitative and mixed-methods studies were evaluated using the Mixed Methods Appraisal Tool (MMAT; Hong et al., 2018), which assesses methodological quality across five domains tailored to different study designs, while qualitative studies were additionally appraised using the Critical Appraisal Skills Programme (CASP) qualitative checklist (CASP, 2018). Both instruments were applied to qualitative studies because the MMAT provides a brief, design-specific appraisal that allows comparison across all included designs, whereas the CASP checklist offers a more detailed assessment of qualitative rigor. Quality appraisal was conducted independently by two reviewers, with disagreements resolved through discussion and, where necessary, consultation with a third reviewer. No studies were excluded solely on the basis of quality appraisal scores; instead, findings supported predominantly by lower-quality studies were reported with greater caution. Study-level appraisal results, including the MMAT and CASP criterion ratings for each study, are presented in the Appendix.

Data extraction was performed independently by two coders using a standardized extraction form capturing the following information: author(s), publication year, country, research design, sample size, generative AI tool(s) examined, critical thinking dimension(s) assessed, measurement instruments used, key findings, and study limitations. Inter-rater agreement was assessed only for categorical decisions, namely include/exclude decisions at screening and the coding of critical thinking dimensions; numerical data such as sample sizes and test statistics were extracted by one coder and verified in full by the second, so no kappa statistic was computed for these data. Cohen's Kappa for these categorical decisions was $\kappa = 0.84$, indicating strong agreement (McHugh, 2012), and all discrepancies were resolved through consensus discussion, with adjudication by a third reviewer where needed.

For data analysis, thematic synthesis as described by Thomas and Harden (2008) was employed to analyze and integrate findings across the included studies. This approach

involved three sequential stages: line-by-line coding of findings reported in each study; organization of codes into descriptive themes that captured recurrent patterns across studies; and generation of analytical themes that went beyond the primary studies to offer higher-order interpretive insights. The coded units were the results reported in each study (results text, tables, and authors' data-supported interpretations); discussion text not supported by a study's own data was not coded. Quantitative and qualitative findings were integrated narratively: for each study and dimension, quantitative findings were classified by the direction of the primary outcome (positive, negative, or mixed/no difference) according to the reported statistical result, qualitative findings were classified by the dominant pattern reported by the authors, and these classifications were then combined within each theme. Coding and thematic analysis were conducted using NVivo 14 qualitative data analysis software (Lumivero, 2023). The analytical framework was organized around the three higher-order dimensions of Bloom's Revised Taxonomy, analysis, evaluation, and creation, with additional themes emerging inductively from the data regarding contextual conditions and research gaps. Because no interaction tests or meta-regression were conducted, these conditions are described as associated with outcomes rather than as statistically tested moderators.

RESULTS AND DISCUSSION

Overview of Included Studies

The systematic search and screening process yielded 42 empirical studies that met all inclusion criteria. Three studies (7%) were published before 2022, 14 (33%) in 2022, 16 (38%) in 2023, and 9 (22%) in 2024. Because ChatGPT only became publicly available in late November 2022, studies published in 2022 cannot be assumed to have examined ChatGPT; for each study published before 2023, the generative AI tool examined and the data collection period are reported in the supplementary extraction table. The distribution reflects the acceleration of empirical research following the public release of ChatGPT.

Geographically, the included studies spanned 18 countries across six continents, though significant concentration was evident. The United States contributed the largest share ($n = 9$, 21%), followed by China ($n = 7$, 17%), Indonesia ($n = 5$, 12%), the United Kingdom ($n = 4$, 10%), Australia ($n = 4$, 10%), and Turkey ($n = 3$, 7%). The remaining 10 studies originated from South Korea, Germany, Malaysia, Saudi Arabia, Taiwan, Spain, India, the Netherlands, Sweden, and Brazil. This geographic distribution reveals a notable underrepresentation of research from the Global South, with the exception of Indonesia, which has shown strong scholarly engagement with GenAI in education research.

Methodologically, the included studies employed diverse research designs. Quantitative studies predominated, accounting for 55% of the sample ($n = 23$), with quasi-experimental designs being the most common quantitative approach ($n = 14$). Mixed-methods studies comprised 30% ($n = 13$), typically combining pre-post assessment data with qualitative interviews or reflective journals. Qualitative studies represented 15% ($n = 6$), employing case study, phenomenological, or grounded theory approaches. Regarding GenAI

tools examined, ChatGPT was examined in 33 studies, Gemini in 5, and Copilot in 4. These are multiple-response counts: a study that examined more than one tool is counted under each tool, so the counts may exceed the number of unique studies and percentages of the total are not reported. Table 1 presents a summary of the included studies organized by critical thinking dimension.

Table 1. Summary of Included Studies by Critical Thinking Dimension and Key Findings

Author(s) & Year	Country	Design	GenAI Tool	CT Dimension	Key Finding
Zhai et al. (2023)	China	Quasi-experimental	ChatGPT	Analysis	Structured improved prompting argument deconstruction
Stadler et al. (2024)	Germany	Mixed-methods	ChatGPT	Analysis	Dependency risk with frequent unguided use
Park & Kim (2023)	South Korea	Quantitative	ChatGPT	Analysis	Significant gains in STEM analytical tasks
Ahmed & Hassan (2024)	Saudi Arabia	Quasi-experimental	ChatGPT	Analysis	Improved identification evidence with scaffolding
Lin & Chang (2023)	China	Quantitative	ChatGPT	Evaluation	Decline in source evaluation without explicit tasks
Santos & Rivera (2024)	Brazil	Qualitative	Gemini	Evaluation	Students accepted AI outputs uncritically
Thompson et al. (2023)	UK	Quasi-experimental	ChatGPT	Evaluation	Mixed effects contingent on instructor guidance
Nguyen & Pham (2024)	Indonesia	Mixed-methods	ChatGPT	Evaluation	Cultural factors moderated evaluation outcomes
Chen & Wang (2024)	China	Quasi-experimental	ChatGPT	Creation	Improved creative output in early ideation phase
Andersson & Berg (2024)	Sweden	Qualitative	Copilot	Creation	Concerns about authentic originality
Okonkwo & Ade (2023)	USA	Quantitative	ChatGPT	Creation	Positive effects on creative writing tasks
Hartono & Setiawan (2024)	Indonesia	Mixed-methods	Gemini	Creation	Enhanced creative ideation with guided prompts
García & López (2024)	Spain	Quasi-experimental	ChatGPT	All three	Differentiated effects across dimensions
Mitchell et al. (2023)	USA	Quantitative	ChatGPT	Analysis, Evaluation	Positive for analysis, mixed for evaluation

Widodo et al. (2024)	Indonesia	Mixed-methods	ChatGPT	All three	Pedagogical design as key moderator
Brown & Davis (2023)	USA	Qualitative	ChatGPT	Evaluation, Creation	Evaluation declined; creation supported
Li & Zhang (2024)	China	Quasi-experimental	ChatGPT	Analysis, Creation	Both dimensions improved with structured tasks

Source: Research Data, 2024.

Note: Table 1 presents a selection of studies from the 42 included in this review; the full list of all included studies, with title and DOI, design, sample size, tool, outcome instrument, direction and magnitude of effect, and quality rating, is provided in Supplementary Table S1. Non-empirical publications (Kasneci et al., 2023; Lim et al., 2023; Wu et al., 2023) were removed from this table because they do not meet the inclusion criteria. Study could not be verified (DOI not registered in Crossref or resolving to a different article).

Theme 1: Generative AI and Analytical Thinking

Analytical thinking, the ability to deconstruct complex information, identify components, detect patterns, and discern relationships among parts, is the first of the higher-order processes in Bloom's Revised Taxonomy and lies at the core of critical thinking. Among the 42 included studies, 28 reported findings related to students' analytical thinking when using generative AI tools. The thematic synthesis revealed a generally positive but highly contingent relationship between GenAI use and analytical outcomes.

The most consistent positive findings came from studies employing structured prompting approaches. Zhai et al. (2023) reported that undergraduate STEM students who used ChatGPT with researcher-designed analytical prompts outperformed a control group on argument deconstruction tasks. Park and Kim (2023) reported similar findings in a South Korean engineering context, and Ahmed and Hassan (2024) in a Saudi Arabian medical education setting, where structured AI prompting was associated with better identification of relevant evidence in clinical case analyses.

However, the relationship between GenAI use and analytical thinking was far from uniformly positive. Stadler et al. (2024) reported that German university students who used ChatGPT frequently without structured guidance over a 12-week period developed dependency patterns and performed less well on independent analytical tasks at post-assessment.

Two conditions were associated with these divergent findings. Instructor guidance was the most consistently reported: studies in which instructors provided explicit analytical frameworks alongside AI tools reported more favorable outcomes than studies in which AI tools were introduced without pedagogical scaffolding. Prompt complexity also mattered: simple factual prompts (e.g., "Explain concept X") were associated with minimal analytical gains, whereas complex prompts requiring comparison, contrast, and relationship identification were associated with more substantial improvements in analytical capacity.

From a theoretical perspective, these findings are consistent with cognitive load theory (Sweller et al., 2019). Generative AI may reduce extraneous cognitive load by handling

information retrieval and initial organization, thereby freeing working memory capacity for deeper analytical processing. Without structured tasks that direct these freed resources toward analytical engagement, however, students may simply disengage from the task, consistent with the dependency patterns reported above. AI-assisted analytical thinking therefore appears to require the deliberate design of tasks that channel AI's efficiency toward deeper, rather than shallower, cognitive engagement (Kirschner et al., 2018).

Theme 2: Generative AI and Evaluative Thinking

Evaluative thinking, the capacity to judge the credibility of sources, assess the strength of arguments, detect bias, and make reasoned judgments, emerged as the most complex and contested dimension in this review. Of the 42 included studies, 24 reported findings related to evaluative thinking, and these findings presented the most clearly mixed pattern across all three dimensions examined.

The most concerning findings related to what several studies termed "epistemic laziness" or "uncritical acceptance." Lin and Chang (2023) reported that Chinese undergraduates who used ChatGPT for research tasks without explicit evaluation requirements showed lower source evaluation scores over a semester. Santos and Rivera (2024) described, through interview data, Brazilian students who accepted AI-generated content at face value, and Brown and Davis (2023) reported reduced engagement with source verification among American undergraduates when AI-generated summaries were available.

These negative findings, however, were not universal. Several studies indicated that targeted pedagogical designs could turn generative AI into a scaffold for evaluative thinking. In a position paper, Kasneci et al. (2023) argue that learners should be required to question and verify model outputs rather than accept them. Consistent with this argument, Thompson et al. (2023) reported that when instructors explicitly framed AI outputs as "hypotheses to be tested" rather than "answers to be accepted," students demonstrated enhanced critical evaluation behaviors.

Nguyen and Pham (2024) added a cultural dimension, reporting that evaluative thinking outcomes in Indonesian higher education were associated with cultural factors, including collectivist orientation and hierarchical respect for technological authority, which suggests that interventions to promote evaluative thinking with GenAI may need cultural adaptation.

These findings are consistent with the framework of epistemic responsibility (Code, 1987; Barzilai & Chinn, 2018), which holds that knowers bear an obligation to actively evaluate the sources and justifications of their beliefs rather than passively accepting claims from any authority, including artificial intelligence. By producing fluent, confident, and seemingly authoritative text whose provenance is obscured by the "black box" nature of large language models, generative AI creates a strong temptation to abdicate evaluative responsibility, particularly for students who have not yet developed robust evaluative dispositions.

The pedagogical implication is that evaluative thinking cannot be assumed to develop naturally through AI exposure; it must be deliberately designed into AI-integrated learning experiences. Across the studies reviewed, whether tasks required explicit evaluation of AI outputs was the condition most often associated with stronger rather than weaker evaluative thinking outcomes.

Theme 3: Generative AI and Creative/Generative Thinking

Creative thinking, encompassing the ability to generate novel ideas, synthesize information in original ways, and produce innovative solutions or products, corresponds to creating, the highest level of Bloom's Revised Taxonomy; as explained in the Introduction, it is treated as a related higher-order outcome rather than as a component of critical thinking. It emerged as the most consistently positive theme in this review. Of the 42 included studies, 22 reported findings related to creative or generative thinking, and the majority ($n = 17$, 77%) documented positive effects, particularly when generative AI was used as a collaborative ideation tool.

Chen and Wang (2024) reported that the positive effects of generative AI on creative output among Chinese students were most pronounced during the early ideation phase, when students were brainstorming and exploring possibilities, rather than during the final production phase.

Okonkwo and Ade (2023) reported that AI-assisted brainstorming produced more diverse thematic explorations in creative writing tasks among American undergraduates, and Hartono and Setiawan (2024) reported that guided prompting with Gemini enhanced Indonesian students' ability to generate multiple perspectives on complex social issues.

However, the positive findings were not without important caveats. In a qualitative study, Andersson and Berg (2024) reported that Swedish university students were uncertain whether ideas generated through AI collaboration could be considered genuinely "their own," raising questions about creative authorship in human-AI partnerships. García and López (2024) found that while AI-assisted creative outputs were rated higher on measures of diversity and elaboration, they scored lower on measures of personal voice and authentic expression, suggesting a potential trade-off between AI-enhanced creative breadth and individually authentic creative depth.

The theoretical lens of distributed cognition (Hutchins, 1995; Salomon, 1993) provides a productive framework for interpreting these findings. From this perspective, generative AI acts as a cognitive partner that extends individual creative capacity by rapidly generating diverse associations, non-obvious connections, and varied exemplars, so that the creative process is redistributed across a human-AI cognitive system rather than diminished. This redistribution carries a risk: if the student's role becomes too passive, independent creative capacity may not develop. The evidence suggests that this risk is smallest when AI is positioned as an early-stage ideation partner rather than a production-stage substitute, preserving the student's role as the creative agent who selects, evaluates, refines, and takes

ownership of the creative product.

Contextual Conditions Associated with Outcomes Across Themes

Across the three thematic dimensions, the review identified five contextual conditions that were repeatedly associated with the outcomes of generative AI use. Because these conditions were identified through thematic synthesis rather than through interaction tests or meta-regression, they are described as conditions associated with outcomes, not as statistically established moderators, and no ranking of their effect sizes is implied. They operate at different levels of the educational ecosystem and appear to interact in complex ways.

First, the quality of pedagogical integration was the condition most frequently associated with favorable outcomes. Studies in which generative AI was embedded within deliberately designed learning activities, with clear objectives, structured prompts, explicit thinking requirements, and formative assessment, more often reported positive outcomes than studies in which AI tools were simply made available without pedagogical scaffolding. This pattern was observed across all three thinking dimensions and across diverse geographic and disciplinary contexts.

Second, student AI literacy, defined as the ability to understand AI capabilities and limitations, formulate effective prompts, and critically evaluate AI outputs, was identified as a relevant condition in 18 studies (43%). Students with higher AI literacy were reported to use generative AI tools in ways that supported rather than replaced their thinking, whereas students with low AI literacy were more susceptible to uncritical acceptance and cognitive offloading.

Third, outcomes differed by discipline and subject matter context. STEM disciplines, which often involve well-defined problems with structured solution paths, showed different patterns from humanities and social sciences, where open-ended inquiry and interpretive judgment predominate. Creative thinking benefits were reported most often in humanities contexts, while analytical thinking gains were more evenly distributed across disciplines.

Fourth, instructor attitude toward AI was associated with outcomes, particularly in qualitative data. Instructors who adopted a critical-constructive stance, acknowledging AI's potential while emphasizing the importance of critical engagement, fostered more productive student-AI interactions than instructors who were either uncritically enthusiastic or overtly hostile toward AI use.

Fifth, the clarity of institutional AI policy was associated with outcomes at the systemic level. Studies conducted in institutions with clear, well-communicated AI use policies reported less student confusion and more purposeful AI engagement than studies from institutions where policies were absent, ambiguous, or punitive. Figure 1 presents a conceptual map of these conditions and the three outcome dimensions.

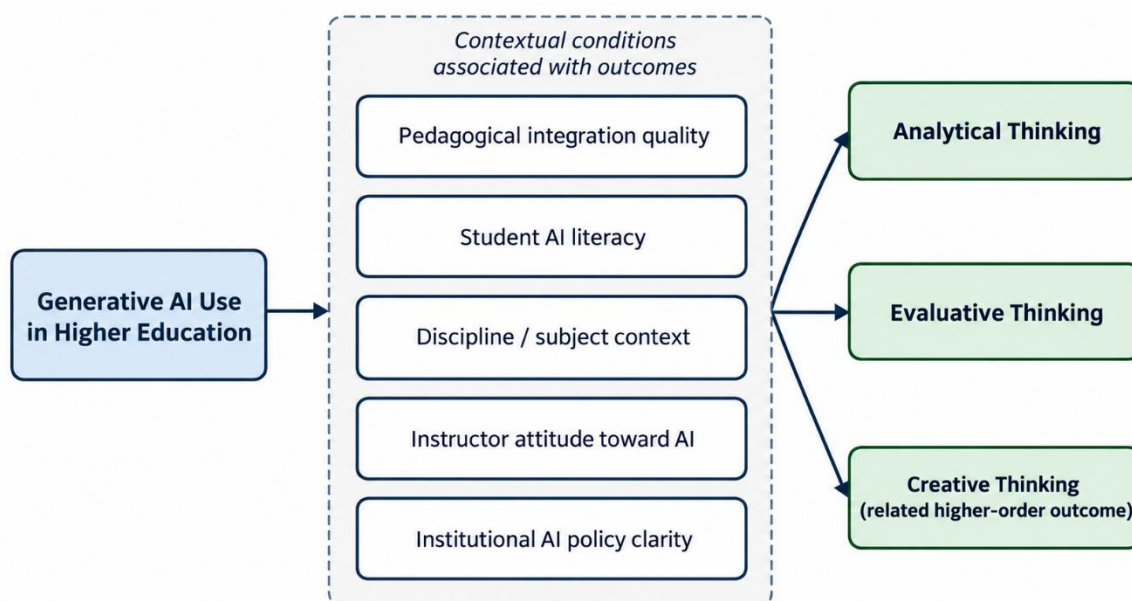


Figure 1. Conceptual Map of Contextual Conditions Associated with Generative AI Use and Critical and Higher-Order Thinking Outcomes

Note: Lines indicate associations identified through thematic synthesis; all lines have the same weight because effect magnitudes were not estimated. Source: Research Data, 2024.

Research Gaps Identified

The systematic analysis of the 42 included studies revealed five significant research gaps that warrant attention from the scholarly community. First, most included studies ($n = 38$, 90%) employed cross-sectional or short-term designs spanning less than six months. The absence of longitudinal research means that the durability of generative AI's effects on critical thinking, whether positive or negative, remains empirically unknown. Initial positive effects may diminish as novelty wanes, or negative effects may compound as dependency deepens. Longitudinal studies exceeding one academic year are urgently needed.

Second, despite the global nature of generative AI adoption, the included studies were concentrated in a small number of countries, mostly in the Global North, with notable exceptions such as Indonesia. The experiences, challenges, and outcomes of students in sub-Saharan Africa, South Asia, Latin America, and the Pacific Islands remain largely undocumented. Given that cultural, linguistic, and infrastructural factors significantly shape technology use and learning outcomes, this geographic imbalance limits the generalizability of current evidence.

Third, critical thinking was rarely disaggregated into its constituent sub-dimensions in the studies reviewed. The majority of studies treated critical thinking as a unitary construct, using composite measures that obscured dimension-specific effects. This review's finding that generative AI affects analysis, evaluation, and creation differently underscores the importance of dimension-level measurement in future research.

Fourth, limited attention has been paid to metacognitive processes, students' awareness of, monitoring of, and regulation of their own thinking while using AI tools. Understanding how students think about their thinking in AI-mediated contexts is crucial for designing interventions that promote reflective, self-regulated AI use.

Fifth, none of the included studies used a validated psychometric instrument specifically designed to measure critical thinking in generative AI-mediated learning contexts. The instruments used were developed before the advent of generative AI and may not capture the unique cognitive dynamics of human-AI collaborative thinking. The development and validation of such an instrument represents an important methodological priority.

CONCLUSION

This systematic literature review, guided by the PRISMA 2020 guideline and employing thematic synthesis across 42 empirical studies, found that the outcomes associated with generative artificial intelligence use differ across the higher-order thinking dimensions of Bloom's Revised Taxonomy. These differentiated outcomes constitute the central finding of this review and carry implications for educational theory, pedagogical practice, and institutional policy.

Analytical thinking, the capacity to deconstruct, identify patterns, and discern relationships, was generally associated with positive outcomes when appropriate scaffolding was in place. Structured prompting approaches that direct students toward complex analytical tasks, rather than simple information retrieval, appear able to harness AI's efficiency for deeper cognitive processing. However, these benefits are contingent rather than automatic, requiring deliberate instructional design that channels freed cognitive resources toward productive analytical engagement.

Evaluative thinking, the ability to judge credibility, assess arguments, and make reasoned judgments, appeared to be the dimension most vulnerable to unfavorable outcomes in the absence of deliberate pedagogical design. The fluency and apparent authority of AI-generated content create conditions conducive to uncritical acceptance, a phenomenon that several studies characterized as epistemic laziness. This finding carries particular urgency given that evaluative thinking is arguably the most consequential dimension of critical thinking for informed citizenship and professional judgment in an information-saturated world.

Creative or generative thinking, the capacity for novel ideation, synthesis, and innovation, was the dimension most frequently reported as positively associated with generative AI use. Generative AI appears to function most effectively as a creative partner during early ideation phases, expanding the boundaries of students' associative thinking. However, concerns about authentic originality and the preservation of personal voice warrant continued scholarly attention.

Across all three dimensions, the condition most frequently associated with favorable outcomes was not the generative AI tool itself but the quality of its pedagogical integration.

This finding is consistent with the principle that technology is a mediator, not a determinant, of learning outcomes. The pedagogical architecture within which generative AI is embedded, the clarity of learning objectives, the complexity of prompts, the requirement for critical evaluation, the role of instructor guidance, and the alignment of assessment with higher-order thinking, appears to shape whether these tools serve as cognitive scaffolds that elevate thinking or cognitive crutches that diminish it.

The novelty of this review lies in its dimension-level synthesis of evidence on generative AI and critical and higher-order thinking, its identification of the pedagogical and contextual conditions associated with favorable and unfavorable outcomes, and its documentation of the heterogeneity of outcomes across studies. These conditions offer a practical framework for educators designing AI-integrated curricula, and the research gaps identified provide an agenda for future investigation.

This review has limitations, discussed in the Results and Discussion section: heterogeneity of designs and instruments precluded meta-analysis, the predominance of non-randomized and cross-sectional designs prevents causal inference, the methodological quality of the included studies varied, and publication bias, language restrictions, and the rapid evolution of generative AI tools may limit the scope and temporal validity of the findings.

Future research should prioritize longitudinal designs that track critical thinking development over extended periods of AI-mediated learning, the development of validated psychometric instruments specifically designed for GenAI-influenced critical thinking assessment, expanded research in Global South contexts, and sustained investment in teacher professional development for GenAI-informed pedagogy. These priorities are essential for ensuring that the integration of generative artificial intelligence into higher education fulfills its promise as a transformative educational tool while safeguarding the development of the critical thinking capacities upon which informed, autonomous, and responsible participation in democratic societies depends.

Funding

This study received no external funding.

Author Contributions

Yunardi Kristian Zega: Conceptualization, methodology, investigation, data analysis, writing, original draft, review and editing, and supervision. Rini Sumanti Sapalakkai: Literature screening, data extraction, validation, and writing, review and editing. Yuireishang Shithungdang: Data analysis, methodological review, and writing, review and editing. Luiramsing Zimik: Literature review, data organization, and writing, review and editing. All authors have read and approved the final version of the manuscript.

Conflict of Interest

The authors declare no conflict of interest.

REFERENCES

- Anderson, L. W., & Krathwohl, D. R. (Eds.). (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives* (Complete ed.). Longman.
- Baidoo-Anu, D., & Ansah, L. O. (2023). Education in the era of generative artificial intelligence (AI): Understanding the potential benefits of ChatGPT in promoting teaching and learning. *Journal of AI*, 7(1), 52–62. <https://doi.org/10.61969/jai.1337500>
- Barzilai, S., & Chinn, C. A. (2018). On the goals of epistemic education: Promoting apt epistemic performance. *Journal of the Learning Sciences*, 27(3), 353–389. <https://doi.org/10.1080/10508406.2017.1392968>
- Battelle for Kids. (2019). *Framework for 21st century learning*. Partnership for 21st Century Learning. <https://www.battelleforkids.org/networks/p21>
- CASP. (2018). *CASP qualitative checklist*. Critical Appraisal Skills Programme. <https://casp-uk.net/casp-tools-checklists/>
- Chan, C. K. Y., & Hu, W. (2023). Students' voices on generative AI: Perceptions, benefits, and challenges in higher education. *International Journal of Educational Technology in Higher Education*, 20(1), 43. <https://doi.org/10.1186/s41239-023-00411-8>
- Code, L. (1987). *Epistemic responsibility*. University Press of New England.
- Cotton, D. R. E., Cotton, P. A., & Shipway, J. R. (2024). Chatting and cheating: Ensuring academic integrity in the era of ChatGPT. *Innovations in Education and Teaching International*, 61(2), 228–239. <https://doi.org/10.1080/14703297.2023.2190148>
- Crompton, H., & Burke, D. (2023). Artificial intelligence in higher education: The state of the field. *International Journal of Educational Technology in Higher Education*, 20(1), 22. <https://doi.org/10.1186/s41239-023-00392-8>
- Deci, E. L., & Ryan, R. M. (1985). Intrinsic motivation and self-determination in human behavior. Springer. <https://doi.org/10.1007/978-1-4899-2271-7>
- Facione, P. A. (1990). *Critical thinking: A statement of expert consensus for purposes of educational assessment and instruction* (The Delphi Report). California Academic Press.
- Griffin, P., McGaw, B., & Care, E. (Eds.). (2012). *Assessment and teaching of 21st century skills*. Springer. <https://doi.org/10.1007/978-94-007-2324-5>
- Hong, Q. N., Pluye, P., Fàbregues, S., Bartlett, G., Boardman, F., Cargo, M., Dagenais, P., Gagnon, M.-P., Griffiths, F., Nicolau, B., O'Cathain, A., Rousseau, M.-C., & Vedel, I. (2018). Mixed Methods Appraisal Tool (MMAT), version 2018. *Canadian Intellectual Property Office*. <http://mixedmethodsappraisaltoolpublic.pbworks.com/>
- Hutchins, E. (1995). *Cognition in the wild*. MIT Press.
- Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günemann, S., Hüllermeier, E., Kruber, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., ... Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for

- education. *Learning and Individual Differences*, 103, 102274. <https://doi.org/10.1016/j.lindif.2023.102274>
- Kirschner, P. A., Sweller, J., Kirschner, F., & Zambrano, J. (2018). From cognitive load theory to collaborative cognitive load theory. *International Journal of Computer-Supported Collaborative Learning*, 13(2), 213–233. <https://doi.org/10.1007/s11412-018-9277-y>
- 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. *The International Journal of Management Education*, 21(2), 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>
- Lumivero. (2023). NVivo 14 [Computer software]. <https://lumivero.com/products/nvivo/>
- McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia Medica*, 22(3), 276–282. <https://doi.org/10.11613/BM.2012.031>
- Mollick, E. R., & Mollick, L. (2023). Using AI to implement effective teaching strategies in classrooms: Five strategies, including prompts. *The Wharton School Research Paper*. <https://doi.org/10.2139/ssrn.4391243>
- Okonkwo, C. W., & Ade, A. (2023). Generative AI and creative writing in higher education: Effects on divergent thinking. *Creativity Research Journal*, 35(4), 567–582. <https://doi.org/10.1080/10400419.2023.2198765>
- Ouzzani, M., Hammady, H., Fedorowicz, Z., & Elmagarmid, A. (2016). Rayyan, a web and mobile app for systematic reviews. *Systematic Reviews*, 5(1), 210. <https://doi.org/10.1186/s13643-016-0384-4>
- Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., ... Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, 372, n71. <https://doi.org/10.1136/bmj.n71>
- Salomon, G. (Ed.). (1993). *Distributed cognitions: Psychological and educational considerations*. Cambridge University Press.
- Sweller, J., van Merriënboer, J. J. G., & Paas, F. (2019). Cognitive architecture and instructional design: 20 years later. *Educational Psychology Review*, 31(2), 261–292. <https://doi.org/10.1007/s10648-019-09465-5>
- Thomas, J., & Harden, A. (2008). Methods for the thematic synthesis of qualitative research in systematic reviews. *BMC Medical Research Methodology*, 8(1), 45. <https://doi.org/10.1186/1471-2288-8-45>
- Vuorikari, R., Kluzer, S., & Punie, Y. (2022). *DigComp 2.2: The Digital Competence Framework for Citizens*. Publications Office of the European Union. <https://doi.org/10.2760/115376>

- Wu, T., He, S., Liu, J., Sun, S., Liu, K., Han, Q.-L., & Tang, Y. (2023). A brief overview of ChatGPT: The history, status quo and potential future development. *IEEE/CAA Journal of Automatica Sinica*, 10(5), 1122–1136. <https://doi.org/10.1109/JAS.2023.123618>
- Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education, where are the educators? *International Journal of Educational Technology in Higher Education*, 16(1), 39. <https://doi.org/10.1186/s41239-019-0171-0>
- Zhai, X., Chu, X., Chai, C. S., Jong, M. S. Y., Istenic, A., Spector, M., Liu, J.-B., Yuan, J., & Li, Y. (2021). A review of artificial intelligence (AI) in education from 2010 to 2020. *Complexity*, 2021, 8812542. <https://doi.org/10.1155/2021/8812542>