education sciences

MDPI

Systematic Review

Beyond Prompting: Student–GenAI Interaction and Critical Thinking in STEM Education—A Systematic Literature Review

Rosiebel A. Abao [1], Evelyn B. Ballenas [2], Vicente A. Pitogo [3] and Junrie B. Matias [4,*]

[1] School of Graduate Studies, Caraga State University, Butuan City 8600, Philippines; rosiebel.abao2@carsu.edu.ph
[2] Department of Mathematics Education, Caraga State University, Butuan City 8600, Philippines; ebballenas@carsu.edu.ph
[3] Department of Information Systems, Caraga State University, Butuan City 8600, Philippines; vapitogo@carsu.edu.ph
[4] Department of Computer Science, Caraga State University, Butuan City 8600, Philippines
* Correspondence: jbmatias@carsu.edu.ph

Abstract

Generative artificial intelligence (GenAI) is becoming increasingly common in science, technology, engineering, and mathematics (STEM) education, but its role in students' critical thinking (CT) remains unclear. This systematic literature review examines how students interact with GenAI, the factors influencing these interactions, and their relationships with CT-related processes in STEM learning. Following PRISMA 2020 guidelines, we searched Scopus, Web of Science, and IEEE Xplore and identified 204 records, of which 17 empirical studies met the inclusion criteria. Six clusters of CT-related processes were found through narrative and thematic synthesis: strategic questioning and task framing; interpretation and analytical decomposition; comparison, evaluation, and verification; error detection and corrective modification; justification, judgment, and independent decision-making; and reflection, metacognition, and self-regulation. Across the included studies, interactions involving questioning AI-generated solutions, verifying their accuracy, examining alternatives, and revising initial interpretations were associated with CT-related processes. Conversely, reliance on GenAI to perform essential reasoning tasks or uncritical acceptance of AI-generated responses may limit opportunities for independent reasoning. These findings suggest that the nature of student–GenAI interactions, particularly the extent to which students retain cognitive responsibility, is important for understanding how GenAI relates to CT processes. Based on the synthesis, we propose an evidence-informed conceptual framework that links student–GenAI interactions, cognitive responsibility, and CT-related processes with personal, instructional, and task-related factors. The framework is derived from the reviewed evidence and requires further empirical validation. However, methodological heterogeneity and limited direct measurement of independent CT outcomes prevent definitive conclusions about whether GenAI improves students' critical-thinking ability. The review highlights the importance of designing STEM learning activities that encourage questioning, verification, reflection, and independent judgment when interacting with GenAI. Further experimental and longitudinal research is needed to validate the proposed framework and examine how student–GenAI interactions influence critical-thinking development over time.

Keywords: generative artificial intelligence; critical thinking; STEM education; student–AI interaction; cognitive responsibility

Check for updates

Academic Editors: Marta Montenegro-Rueda and José Fernández-Cerero

Received: 8 October 2026
Revised: 14 November 2026
Accepted: 18 November 2026
Published: 24 November 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

GenAI is rapidly gaining acceptance in STEM education, but the existing research base is uneven in its emphases. Most existing research has been on AI literacy (Hashemi, 2026), prompt engineering (Lee & Palmer, 2025), adoption (Isaeva et al., 2026), and learning outcomes (X. Sun et al., 2026) rather than the deeper processes of how students interact with AI during learning (Aruleba et al., 2025; Belkina et al., 2025; Lu, 2026; Qu et al., 2025; Tassoti, 2024; Walter, 2024). Recent literature on STEM education has also started to examine how GenAI may be pedagogically structured via computational-thinking activities, adaptive scaffolding, assessment design, and discipline-specific instructional frameworks in addition to prompt formulation (Feldman-Maggor et al., 2025; Hsu, 2025; Kwong et al., 2026; Paredes-Cayllahua, 2026). This pattern reflects that the field has moved more quickly in articulating what GenAI is used for than in explaining how GenAI use shapes student thinking (Belkina et al., 2025; Cong-Lem & Thuy-Dung, 2026; W. Zhang & Liu, 2025).

The most significant concern is that GenAI is often seen as a way to increase efficiency, engagement, or performance. However, the processes by which students interact with it in STEM learning are still poorly understood (Al Darayseh & Mersin, 2025; Cong-Lem & Thuy-Dung, 2026; Qu et al., 2025; Yan et al., 2025). Some studies suggest that GenAI may support critical thinking when students use it in structured, reflective, and guided ways. On the other hand, relying too much on technology or accepting AI-generated answers without challenging them could restrict opportunities for autonomous thought (Krupp et al., 2024; W. Zhang & Liu, 2025). While much of this research focuses on how students use GenAI or the results of its use, with less attention paid to the interactions and reasoning processes involved, other studies emphasize the significance of prompting strategies and AI literacy (Aruleba et al., 2025; Bhuiyan et al., 2025; Krupp et al., 2024; Lee & Palmer, 2025; Tassoti, 2024).

A key gap is the limited integration of evidence on how personal, instructional, and contextual factors shape student–GenAI interaction and critical-thinking-related processes in STEM learning (Al Darayseh & Mersin, 2025; Belkina et al., 2025; Cong-Lem & Thuy-Dung, 2026). Recent studies have begun to examine student–GenAI interaction more directly, including AI dispositions, prompting competencies, and human–AI collaboration in STEM learning (Kim et al., 2026). Existing studies have examined personal factors such as AI literacy, prior experience, confidence, and prompting skills, as well as instructional factors such as teacher guidance, scaffolding, task design, and assessment practices (Lee & Palmer, 2025; Tassoti, 2024; Yan et al., 2025). Related work in science education has also explored AI-supported assessment design, illustrating how the instructional configuration surrounding GenAI can shape the role it plays in learning and assessment (Abualrob & Ghannam, 2026). Discipline-specific implementations further illustrate how GenAI activities can be deliberately scaffolded within STEM courses rather than introduced simply as unrestricted access to an AI tool (Bell & Birceanu, 2025). Other studies have highlighted contextual conditions, including the learning environment, access to AI technologies, institutional support, and characteristics of STEM learning tasks (Al Darayseh & Mersin, 2025; Belkina et al., 2025; Isaeva et al., 2026). However, these factors are often examined separately and primarily in relation to specific outcomes rather than as interconnected influences on how students actually engage with GenAI.

Despite these findings, little is known about how students choose whether to accept, challenge, understand, assess, validate, alter, or reject information produced by AI. Furthermore, it is still unclear how these interactions affect students' critical evaluation, independent judgment, reasoning, and reflection in STEM learning, as well as the circumstances in which they might facilitate or hinder these processes (Cong-Lem & Thuy-Dung, 2026; Krupp et al., 2024; Qu et al., 2025; Válek et al., 2025; W. Zhang & Liu, 2025). GenAI may function as a

cognitive scaffold that supports deeper engagement, or it may encourage cognitive offloading, uncritical acceptance, and over-reliance (Krupp et al., 2024; W. Zhang & Liu, 2025). Therefore, the question is not simply whether GenAI enhances critical thinking; rather, it is how students' experiences and interactions with GenAI, under a variety of personal, instructional, and contextual conditions, relate to critical-thinking-related processes and outcomes in STEM learning.

Although previous studies have examined GenAI adoption, AI literacy, prompting strategies, cognitive impacts, and educational outcomes, emerging research on student–GenAI interaction processes remains fragmented (Fang & Zhou, 2025; Hashemi, 2026; Yan et al., 2025). Similarly, studies investigating the personal, instructional, and contextual conditions surrounding GenAI use have provided valuable insights, yet these findings have not been sufficiently integrated to explain how such conditions shape students' cognitive engagement with AI-generated content (Cong-Lem & Thuy-Dung, 2026; Krupp et al., 2024; Qu et al., 2025; Válek et al., 2025; W. Zhang & Liu, 2025). Consequently, an integrated understanding remains limited regarding how these conditions influence students' questioning, interpretation, reasoning, evaluation, verification, reflection, and independent judgment during GenAI interactions in STEM learning. Furthermore, the extent to which these interactions reflect observable critical-thinking-related processes, rather than demonstrated improvements in critical-thinking ability, remains insufficiently established. Therefore, a systematic synthesis of existing empirical evidence is needed to clarify student–GenAI interaction patterns, their influencing factors, associated critical-thinking processes and outcomes, and the remaining gaps in STEM education research.

In this context, this review differs from prior syntheses focused primarily on AI literacy, prompt engineering, or overall critical-thinking outcomes (Pangestu et al., 2026; Raitskaya & Tikhonova, 2025; Reicho et al., 2026) by treating student–GenAI interaction as the primary unit of analysis. It examines not only how GenAI is used, but also what cognitively and epistemically consequential work learners retain, share, or delegate during these interactions.

Accordingly, this systematic literature review examines existing evidence on how students use and experience GenAI in STEM learning, focusing on critical-thinking-related processes and the personal, instructional, and contextual factors that influence their interactions with the technology. The review examines patterns and situations that may explain how students interact with GenAI and how these interactions relate to critical-thinking processes, rather than assuming that the use of GenAI directly strengthens or worsens critical thinking. Additionally, it points out gaps in the literature that can direct further investigation and theoretical advancement. The review answers the following research questions in order to have a deeper understanding of these relationships:

1. RQ1. What patterns of GenAI use and student experiences are reported in STEM learning?
2. RQ2. What personal, instructional, and contextual factors influence student–GenAI interactions in STEM learning?
3. RQ3. What critical-thinking-related processes are reported or observed during students' interactions with GenAI in STEM learning?
4. RQ4. What overarching relationships and research gaps emerge from the evidence on GenAI use, student–GenAI interaction, and critical-thinking processes in STEM learning?

2. Materials and Methods

2.1. Review Design

This study employed a systematic literature review to identify, evaluate, and synthesize empirical evidence on students' interactions with generative artificial intelligence (GenAI)

and their relationship with critical-thinking-related processes in STEM education. The review was conducted and reported in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 statement (Page et al., 2021).

This review looked at empirical studies indexed in three complementary databases: IEEE Xplore, Web of Science (WoS), and Scopus in order to gather evidence from various domains. It concentrated on four interconnected topics: how students utilize and interact with GenAI; variables affecting student-GenAI interactions; critical thinking processes; and emerging patterns, methodological limitations, and research gaps.

The review followed a systematic process that included identifying relevant studies, removing duplicates, screening titles and abstracts, assessing full-text eligibility, extracting data, and synthesizing the findings. Pre-defined eligibility criteria were used to choose the studies, with special consideration given to their applicability to student-GenAI interactions and critical-thinking procedures in STEM education. Narrative and thematic synthesis were employed to detect common patterns, correlations, and differences across the findings because the included studies differed in terms of research design, educational setting, GenAI implementations, and outcome measures.

Throughout the analysis, we made a distinction between increases in critical-thinking skills tested separately and critical-thinking-related behaviors that students reported or displayed while using GenAI. The PRISMA 2020 flow diagram (Figure 1) shows the study-selection procedure; the search strategy, eligibility requirements, and screening outcomes are explained in more detail in the sections that follow.

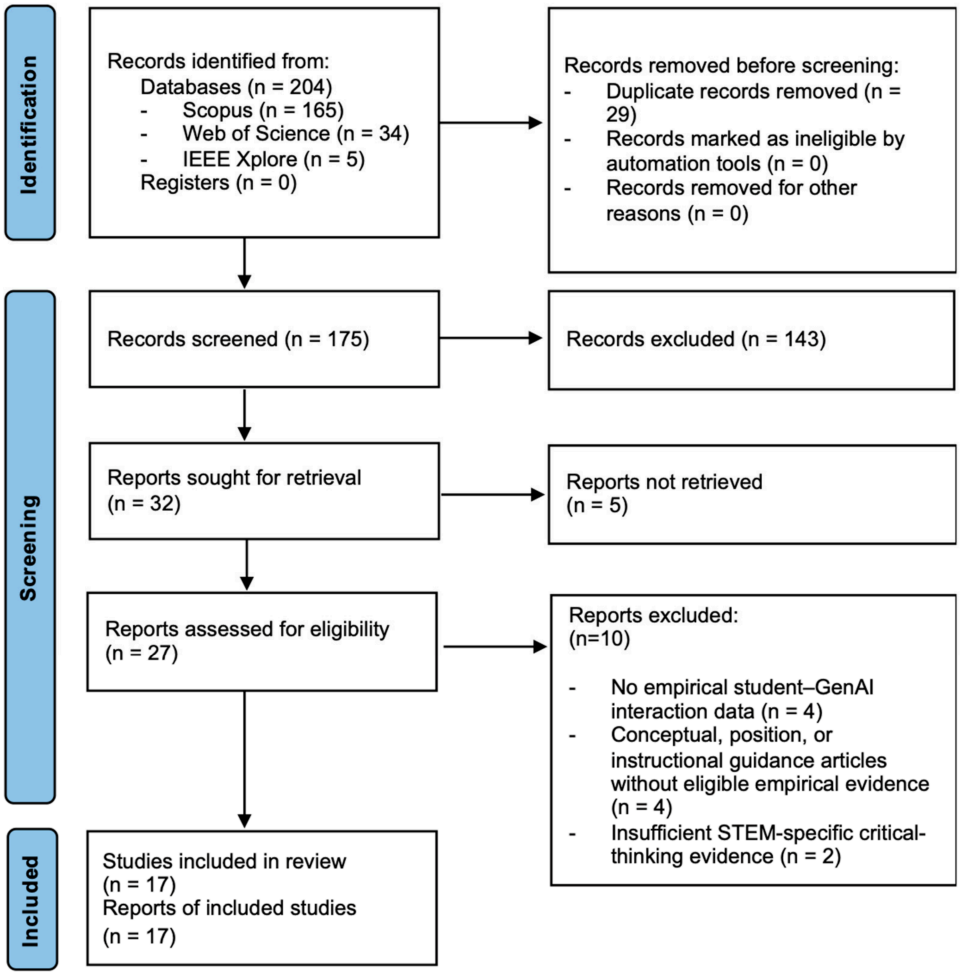


Figure 1. PRISMA study selection diagram.

2.2. Information Source and Search Strategy

Three academic databases—Scopus, Web of Science (WoS), and IEEE Xplore—were used to conduct a thorough literature search. In order to collect a wider range of research in education, science, engineering, and computing, the search was first conducted in Scopus on 5 August 2026, and then extended to WoS and IEEE Xplore on 30 October 2026. Because of their multidisciplinary coverage, Scopus and WoS were chosen, and IEEE Xplore offered more access to engineering, computer, and associated technology research. When combined, these databases provided complementary sources of information about critical thinking, STEM education, and GenAI.

Using Boolean operators (AND and OR), the search strategy concentrated on four related areas: (1) prompt engineering, which entails creating and improving prompts when interacting with GenAI; (2) GenAI technologies, such as ChatGPT, GPT, large language models (LLMs), and related systems; (3) critical thinking, which is the primary cognitive process covered in this review; and (4) STEM fields, which include science, technology, engineering, mathematics, and computing. To ensure transparency and reproducibility, the exact database-specific search strings, execution dates, search fields, publication-type filters, and corresponding record counts are provided in Supplementary Table S7. The generalized Boolean structure underlying the database-specific searches was as follows:

("prompt engineering")
AND (ChatGPT OR GPT OR LLM* OR "large language model*")
AND ("critical thinking")
AND (STEM OR mathematics OR engineering
OR science OR computing OR technology).

The search syntax was adapted to the respective database interfaces, using title, abstract, and keyword fields where supported. The total number of records from the combined searches was 204: 165 from Scopus, 34 from WoS, and 5 from IEEE Xplore. Following the removal of 29 duplicates, 175 unique records were left to be screened against pre-defined eligibility criteria (Table 1).

Table 1. Eligibility criteria for study selection.

Criterion	Inclusion Criteria	Exclusion Criteria
Generative AI Technology	Studies involving Generative AI (GenAI) technologies used in learning, including ChatGPT, Gemini, Copilot, large language models (LLMs), generative AI assistants/tutors, or comparable systems capable of generating substantive content in response to learner input.	Studies involving only conventional/non-generative AI, rule-based intelligent tutoring systems, predictive AI, automated grading, learning analytics, or other technologies without a substantive GenAI component.
STEM Focus	Studies involving STEM education or learning, including science, technology, engineering, mathematics, programming, laboratory work, modeling, inquiry, data analysis, or STEM-related problem-solving.	Studies with no substantive STEM education or learning component, including studies exclusively in non-STEM disciplines.
Population	Studies involving students or learners engaged in STEM education or STEM-related learning activities.	Studies involving only teachers, instructors, researchers, or professionals without relevant student/learner evidence.
Student–GenAI Interaction	Studies examining how students use, interact with, experience, engage with, or respond to GenAI, including prompting, questioning, interpreting, evaluating, verifying, modifying, accepting, or rejecting AI-generated information.	Studies in which GenAI is used only by instructors, researchers, or institutions, or where student–GenAI interaction/experience is not substantively examined.

Table 1. *Cont.*

Criterion	Inclusion Criteria	Exclusion Criteria
Cognitive/Process Relevance	Studies addressing critical thinking or directly related higher-order cognitive processes, including reasoning, analysis, evaluation, inference, interpretation, reflection, metacognition, judgment, problem-solving, argumentation, verification, or independent decision-making.	Studies examining only acceptance, intention to use, satisfaction, usability, engagement, academic performance, or similar outcomes without relevant cognitive or learning-process evidence.
Educational Context	Formal or informal educational settings involving students engaged in STEM learning, including secondary, higher, vocational, or professional education where applicable.	Non-educational GenAI applications or studies without connection to STEM learning.
Study Design/Evidence	Empirical qualitative, quantitative, mixed-methods, case, or educational intervention studies providing substantive evidence relevant to at least one review question.	Conceptual papers, opinion pieces, purely theoretical discussions, or technical studies without relevant empirical student evidence.
Publication Type	Peer-reviewed journal articles retrieved through the predefined searches of Scopus, Web of Science, and IEEE Xplore and meeting the review's eligibility criteria.	Reviews, editorials, letters, book chapters, conference papers/proceedings, abstracts, posters, theses/dissertations, and non-peer-reviewed publications.
Full Text	Full text available for detailed assessment.	Full text unavailable after reasonable retrieval attempts.
Language	English-language publications.	Non-English publications.

The search strategy specifically required the terms "prompt engineering" and "critical thinking" to identify studies situated at the intersection of student–GenAI interaction and critical-thinking research. "Prompt engineering" was used to prioritize studies explicitly examining how users formulate, structure, or refine their interactions with GenAI, while "critical thinking" was required to maintain a direct connection with the review's primary cognitive focus. During eligibility assessment and data extraction, related cognitive processes, including reasoning, evaluation, metacognition, verification, judgment, and problem-solving, were also considered. Nevertheless, these mandatory search terms increased specificity at the expense of sensitivity and may have excluded relevant studies describing comparable interaction practices or cognitive processes using alternative terminology, such as iterative prompting, human–AI collaboration, AI-assisted problem-solving, or evaluative reasoning. This restriction is acknowledged as a limitation of the review's search coverage. The detailed study-selection results are presented in the PRISMA 2020 flow diagram (Figure 1).

2.3. Eligibility Criteria

Eligibility criteria were established a priori to identify empirical studies relevant to student–GenAI interaction and critical-thinking processes in STEM learning. Studies were assessed according to the GenAI technology, STEM learning context, learner population, nature of student–GenAI interaction, cognitive/process relevance, educational context, study design, and publication characteristics. The criteria used for screening and full-text eligibility assessment are presented in Table 1.

The eligibility criteria applied to studies included in the systematic synthesis; additional literature cited in the Introduction and Discussion was used only to contextualize the review and was not treated as part of the included evidence base.

2.4. Study Selection

The Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA 2020) guidelines were adhered to during the study selection procedure. A single dataset was created by combining records from IEEE Xplore, Web of Science, and Scopus; duplicates were eliminated before screening.

The remaining records were screened by title and abstract to determine their relevance to the review objectives and eligibility criteria. Studies that looked at how students used generative artificial intelligence (GenAI) tools in STEM education and offered empirical data about critical-thinking procedures or results were taken into consideration for inclusion. Records that did not fit these requirements were eliminated.

After that, full-text articles were obtained and their eligibility evaluated. This evaluation concentrated on whether the studies documented actual student-GenAI interactions, were carried out in STEM-related educational environments, and offered proof of processes or results connected to critical thinking. Studies that lacked pertinent empirical data, concentrated primarily on conceptual or instructional discussions, or did not offer enough proof of critical thinking in STEM contexts were eliminated.

Two reviewers independently screened the titles and abstracts of the retrieved records against the predefined eligibility criteria. Potentially eligible studies subsequently underwent full-text assessment by the same two reviewers. Disagreements during either screening stage were resolved through discussion and consensus, with a third reviewer consulted when necessary to resolve outstanding eligibility disagreements. The final inclusion decisions were established following this process.

The PRISMA 2020 flow diagram, which shows the number of records found, duplicates eliminated, records screened, full-text reports sought and evaluated, reasons for exclusion, and studies included in the final synthesis, illustrates the selection process.

2.5. Data Extraction

A structured data-extraction framework aligned with the research questions was used to extract data from the included studies. Extracted information included study characteristics, STEM context and task, participants, GenAI platform, study design, student–GenAI interaction patterns, influencing factors, CT measures or related cognitive processes, key outcomes, and study limitations. Extracted data were independently checked by a second reviewer against the full-text articles, and discrepancies were resolved through discussion and consensus.

Particular attention was given to CT-related processes demonstrated or reported during interaction, including questioning, analysis, comparison, error detection, evaluation, verification, modification, reflection, judgment, and independent decision-making. Evidence was classified according to whether CT was directly measured, assessed through a related reasoning construct, or represented through performance-based, behavioral, process-level, or self-reported evidence.

2.6. Methodological Quality Assessment

The methodological quality of 15 included studies was assessed using the Mixed Methods Appraisal Tool (MMAT 2018), with the appraisal criteria selected according to their respective research designs (Hong et al., 2018). The remaining two studies were examined separately using the initial MMAT screening questions to assess their empirical basis: S1, whether clear research questions were presented, and S2, whether the collected data were sufficient to address those questions. These two studies were not assessed using the design-specific MMAT criteria because their reported methodologies did not permit reliable classification within the tool's established study-design categories. The initial

screening questions were used to examine the empirical basis of these studies and inform the interpretation of their findings, rather than to determine study eligibility or assign a methodological quality rating. Accordingly, the screening results for these two studies were distinguished from the design-specific MMAT assessments of the other 15 studies.

The two screening-only studies were retained in the narrative and thematic synthesis because they provided relevant descriptive evidence of student–GenAI interaction practices. However, their findings were interpreted cautiously and were not treated as independently verified evidence of improvement in critical-thinking ability. This distinction allowed the synthesis to incorporate relevant interaction patterns while avoiding stronger outcome claims than the studies' methodological reporting could support.

2.7. Data Synthesis and Analysis

A structured narrative and thematic synthesis was conducted across the 17 included empirical studies. It was deemed inappropriate to conduct a quantitative meta-analysis due to the heterogeneity across studies in terms of research design, STEM discipline, GenAI application, and critical thinking (CT) outcomes. Instead, an inductive approach was adopted to identify recurring patterns, relationships and differences across the evidence related to the four research questions.

The analysis was carried out in three phases. First, an evidence matrix was constructed to organize information on study characteristics, student–GenAI interaction patterns, influencing factors, CT-related processes, and reported outcomes. The matrix provided a common basis of comparison across studies while also allowing for differences in research context and methodological approach.

Second, concepts and cognitive processes related to CT were extracted and coded from the identified evidence. The first reviewer performed the initial coding, producing 45 codes that were subsequently independently checked by a second reviewer. During the verification process, the relevance and consistency of the codes, their assignment to the extracted evidence, and the suggested thematic groupings were taken into account. Consensus was reached in resolving disagreements about interpretation, code assignment, and thematic boundaries; a third reviewer was consulted as needed. Based on the process, the 45 initial codes were grouped into six major themes: (1) strategic questioning and task framing; (2) interpretation and analytical decomposition; (3) comparison, evaluation, and verification; (4) error detection and corrective modification; (5) justification, judgment, and independent decision-making; and (6) reflection, metacognition, and self-regulation. Supplementary Table S3 shows the primary and secondary thematic allocations, and Supplementary Table S4 shows the distribution of the first codes across the six themes. These themes were developed to organize and interpret the evidence reviewed and should not be viewed as independently validated dimensions of critical thinking.

To clarify the nature of the supporting evidence, study–code intersections were categorized as discussed (D), observed (O), measured (M), or not established (-), according to the definitions provided in Supplementary Table S3. These classifications helped distinguish processes discussed by study authors from those documented through student interactions or directly assessed. These classifications were used to describe the representation of CT-related processes across studies, rather than to demonstrate the effectiveness of GenAI or improvements in students' independent critical-thinking ability.

Third, the coded findings were compared across studies to identify recurring patterns of interaction, variations, contextual influences, and gaps in the evidence. Special attention was paid to distinguishing between students' self-reported experiences of CT, CT-related behaviors observed by the authors, related cognitive measures, and direct assessments of CT. This distinction helped avoid interpreting participation in CT-related activities as evidence of

improved independent critical-thinking ability. This synthesis also aided the interpretation of cognitive and epistemic responsibility in student–GenAI interactions without assuming causal relationships.

3. Results

3.1. Study Selection and Characteristics

The systematic search across Scopus, Web of Science (WoS), and IEEE Xplore identified 204 records, comprising 165 from Scopus, 34 from WoS, and five from IEEE Xplore. Following the removal of 29 duplicate records, 175 unique records underwent title and abstract screening, of which 143 were excluded based on the predefined eligibility criteria. The remaining 32 reports were sought for retrieval; however, five could not be obtained, leaving 27 full-text reports for eligibility assessment. Of these, ten were excluded: four lacked eligible empirical student–GenAI interaction data, four primarily presented conceptual, position, or instructional guidance without eligible empirical evidence, and two provided insufficient STEM-specific critical-thinking evidence (Supplementary Table S2). Consequently, 17 empirical studies met the inclusion criteria and were retained for qualitative synthesis, as illustrated in the PRISMA 2020 flow diagram (Figure 1).

The 17 included studies covered a range of STEM disciplines and educational contexts, including programming, chemistry, engineering mathematics, calculus, mathematical modeling, robotics, biochemistry, database and web development, life sciences, nanotechnology, and science education research. The activities investigated ranged from balancing chemical equations and solving non-routine integral problems to programming, debugging, robotic design, scientific verification, laboratory analysis, and research writing. The detailed characteristics of individual studies, including their research designs, participants, GenAI tools, and interaction-related evidence, are presented in Appendix A (Table A1).

Table 2 shows a descriptive distribution that suggests that research on student–GenAI interaction and critical thinking in STEM education has been on the rise in recent years. The highest share of publications was in 2025 (52.9%), followed by 2026 (29.4%) and 2024 (17.6%). In terms of disciplinary coverage, programming, computing, and database education and engineering and robotics were equally represented, each accounting for 29.4% of the included studies. Mathematics and mathematical modeling, chemistry, and life sciences and biochemistry each constituted 11.8%, whereas science education research comprised 5.9%. This distribution shows that the evidence is focused on computational and engineering-oriented learning contexts, but other STEM disciplines are present in the dataset.

Regarding educational level, 15 studies (88.2%) were conducted in tertiary education, whereas only two (11.8%) involved secondary-level learners. This concentration suggests that most empirical evidence comes from higher education settings, particularly undergraduate STEM learning. Hence, the generalizability of the results to secondary education is less well assured and requires further investigation.

The GenAI applications and methodological strategies of the included studies also varied. The most prominent platform was ChatGPT, followed by Kimi Intelligent Assistant, GitHub Copilot, Gemini, Perplexity AI, and tailored retrieval-augmented generation (RAG) assistants. GenAI supported a variety of activities, including concept explanation, problem-solving, code generation and debugging, scientific writing, information retrieval, engineering design, feedback, and guided inquiry. Methodologically, the studies used mixed-methods, quasi-experimental, pre–post, observational, qualitative, and case-based approaches (Appendix A, Table A1). This diversity is a result of the various approaches used to study student-GenAI interaction in STEM learning environments.

Table 2. Distribution of included studies by publication year, STEM discipline, and educational level.

Characteristic	Category	Frequency (n)	Percentage (%)
Publication year	2024	3	17.6
	2025	9	52.9
	2026	5	29.4
	Subtotal	**17**	**100.0**
STEM discipline	Programming, computing, and database education	5	29.4
	Engineering and robotics	5	29.4
	Mathematics and mathematical modeling	2	11.8
	Chemistry	2	11.8
	Life sciences and biochemistry	2	11.8
	Science education research	1	5.9
	Subtotal	**17**	**100.0**
Educational level	Secondary education	2	11.8
	Tertiary education	15	88.2
	Subtotal	**17**	**100.0**

Student–GenAI interaction likewise varied, ranging from simple prompting and answer seeking to more iterative and evaluative engagement. Reported activities included contextual and iterative prompting, task decomposition, debugging, comparison of alternatives, external verification, modification and rejection or refinement of AI-generated outputs. These interactions were analyzed using different types of evidence such as chat logs, screen recordings, code submissions, student reflections, surveys, learning artifacts, and performance data (Appendix A, Table A1).

The studies further differed in how critical thinking (CT) was operationalized and assessed. Explicit CT-related assessment was reported in a limited number of studies. Gong et al. (2025) assessed CT as a dimension of computational thinking, whereas Nathaniel et al. (2025) used a tailored HOTS assessment in which CT was one of four domains evaluated through rubric-scored programming tasks. Supriadi and Suherman (2026) directly assessed the related construct of mathematical logical reasoning. Most other studies provided process-level or self-reported CT-related evidence, including questioning, analysis, error detection, evaluation, verification, modification, reflection, and judgment. In Cruz-Angeles et al. (2026), for example, CT-related processes were self-reported in Phase I, while Phase II measured academic performance rather than CT.

The included studies represent an emerging but methodologically heterogeneous evidence base. Accordingly, the synthesis distinguishes direct critical thinking (CT) measurement from related reasoning measures, self-reported perceptions, and observed interaction behaviors. The methodological assessment covered 15 of the 17 studies using applicable MMAT 2018 criteria, revealing strengths in examining student–GenAI interactions alongside variations in sampling, outcome measurement, and consideration of potential confounding factors. The remaining two studies (Chen et al., 2024; Štuikys et al., 2025) were evaluated using only the initial MMAT screening questions because the reported methodological information did not permit confident classification under the MMAT design-specific categories. Their findings were retained as descriptive evidence of student–GenAI interaction practices but were not interpreted as independently verified evidence of improvement in critical-thinking ability. These methodological characteristics informed the interpretation of the synthesized findings, particularly the strength of evidence supporting CT-related outcomes. Criterion-level results are presented in Supplementary Table S1, with addi-

tional considerations for the two studies evaluated using only the initial MMAT screening questions documented in the Supplementary Materials.

For example, Chen et al. (2024) described student–GenAI interactions involving prompt formulation, code generation, verification, and iterative optimization during intelligent vehicle competition training. Although the study presented experiences from three student groups, it provided limited information on participant selection, systematic data collection and analysis, and the measurement of critical-thinking outcomes. Because the reported methodological information did not permit confident classification under a standard MMAT study-design category, the study was assessed using only the initial screening questions rather than the design-specific appraisal criteria. Consequently, its findings were interpreted as descriptive evidence of student–GenAI interaction practices and associated reasoning processes rather than as evidence of improvement in independently assessed critical-thinking ability. This example illustrates how methodological limitations informed the interpretation of findings and the strength of conclusions drawn from the synthesis.

3.2. Patterns of GenAI Use and Student–GenAI Interaction (RQ1)

Across the included studies, GenAI served multiple roles in STEM learning, ranging from answer and content generation to interactive support for problem solving, inquiry, verification, and refinement. Student engagement likewise varied considerably, revealing a spectrum of retained cognitive responsibility.

3.2.1. Roles of GenAI in STEM Learning

GenAI functioned as an information and explanation resource, problem-solving and modeling assistant, coding and debugging tool, and instructional scaffold. For example, students used GenAI to solve non-routine integral problems (Supriadi & Suherman, 2026), support Fermi modeling (Gallart et al., 2026), design and troubleshoot autonomous robots (Štuikys et al., 2025), support programming and engineering problem-solving (Chen et al., 2024; Dominguez et al., 2026; Gong et al., 2025; Nathaniel et al., 2025), and generate, explain, and debug code in web-development tasks (Buendía-García & Piris-Ruano, 2025). In research-oriented activities, GenAI supported concept clarification, methodology development, data interpretation, and academic writing (Picardal, 2025). More structured instructional roles were evident in LLM-CAS inquiry and cross-verification in engineering mathematics (Matzakos & Moundridou, 2025), guided inquiry for balancing chemical equations (Lin et al., 2026), prompting frameworks in chemistry (Tassoti, 2024), independent-before-AI comparison in database learning (X. Zhang, 2025), and customized RAG-based assistance in life sciences and nanotechnology (Cruz-Angeles et al., 2026) (Appendix A, Table A1).

These cases show that GenAI was used not only to generate solutions but also to support iterative learning processes involving explanation, experimentation, feedback, and verification. Its role therefore varied according to the STEM task and the instructional structure surrounding its use.

3.2.2. Spectrum of Retained Cognitive Responsibility

Student–GenAI interactions varied across the included studies in the nature and extent of cognitive engagement demonstrated or reported by learners. To interpret these variations, the present synthesis proposes a spectrum of retained cognitive responsibility (Figure 2), ranging from simple prompting and answer seeking to contextual and iterative prompting, problem decomposition, evaluation and verification, and the modification, challenge, or rejection of AI-generated outputs. This spectrum represents an interpretive organization

of recurring interaction patterns rather than a directly measured or empirically validated continuum of cognitive responsibility.

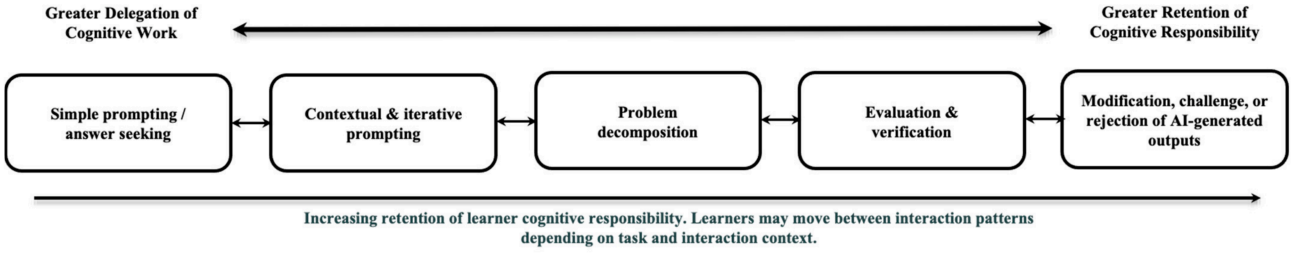


Figure 2. Proposed spectrum of student–GenAI interaction according to retained cognitive responsibility. The figure conceptually organizes recurring interaction patterns identified across the included STEM studies, ranging from greater delegation of cognitive work through simple prompting and answer seeking to more active learner involvement through evaluation, verification, modification, challenge, or rejection of AI-generated outputs. The spectrum is an interpretive framework rather than a validated measurement scale or developmental sequence.

For the purposes of this synthesis, these interaction patterns were characterized according to the nature of the cognitive work described or documented in the included studies. Simple prompting involved direct answer seeking with limited scrutiny. Contextual and iterative prompting involved progressively refining prompts in response to AI-generated outputs. Problem decomposition involved breaking complex tasks into components and selectively using GenAI for specific subtasks. Evaluation and verification involved assessing AI-generated outputs against disciplinary knowledge, independent solutions, external sources, computational tools, or empirical evidence. Finally, modification, challenge, or rejection involved actively correcting, contesting, refining, or discarding AI-generated outputs. These categories are neither mutually exclusive nor sequential, as students may engage in multiple interaction patterns depending on the task, learning context, and purpose of GenAI use. Their organization within the proposed spectrum is interpretive and does not imply distinct, empirically established levels of cognitive responsibility.

At the greater-delegation end of the proposed spectrum, some studies documented interaction patterns involving simple prompts, requests for complete solutions, or acceptance of AI-generated outputs with limited scrutiny. Examples included the guided/acceptance-oriented pattern (Gong et al., 2025), the naive copy-paste pattern (Supriadi & Suherman, 2026), and the high-delegation PS-expert profile (Gallart et al., 2026). Similar tendencies appeared in (Tassoti, 2024), where students frequently initiated interaction by transferring existing chemistry questions directly into ChatGPT, and in Phase I of the study by Cruz-Angeles et al. (2026), where students largely remained close to the instructor-provided example prompt. Findings from Dominguez et al. (2026) further indicated that prompting frequency should not automatically be interpreted as deeper cognitive engagement, as more frequent querying was associated with poorer performance. Within the proposed framework, these interaction patterns were interpreted as reflecting relatively greater delegation of cognitive work to GenAI.

At the greater-retention end of the proposed spectrum, several studies documented interactions in which students took a more active role in examining AI-generated outputs. These included questioning responses, identifying errors, comparing alternatives, verifying information, testing solutions, and modifying or rejecting AI recommendations. Such behaviors were reflected in the debating and summarizing interaction patterns (Gong et al., 2025), iterative debugging and error identification (Supriadi & Suherman, 2026), independent programming followed by critical comparison and refinement of ChatGPT-generated solutions described (Aruleba et al., 2025), external verification of scientific information

(Reddy et al., 2024), LLM-CAS cross-checking (Matzakos & Moundridou, 2025), and empirical testing and refinement of AI-supported engineering and robotics solutions in (Chen et al., 2024; Štuikys et al., 2025). These behaviors were interpreted as indicating greater learner involvement in cognitive processes, although the studies did not uniformly measure cognitive responsibility as a distinct construct.

The studies suggest that student engagement with GenAI cannot be adequately characterized by prompt complexity or frequency alone. Interactions involving questioning, evaluating, verifying, refining, and judging AI-generated responses reflected more active learner participation in cognitive work, whereas accepting AI-generated solutions with limited scrutiny suggested greater delegation of reasoning and decision-making. These interaction patterns, along with representative supporting evidence from the included studies, are summarized in Supplementary Table S5. However, these distinctions should not be interpreted as evidence that particular interaction patterns necessarily improve students' independent critical-thinking proficiency. Rather, they provide a basis for understanding how students engage with GenAI and for examining the personal, instructional, and contextual factors associated with these interactions in RQ2.

3.3. Factors Influencing Student–GenAI Interaction (RQ2)

The included studies identified personal, instructional, and contextual or task-related factors associated with student–GenAI interaction. These factors were related to how students engaged with AI-generated responses, ranging from reliance on GenAI for cognitive tasks to more active questioning, verification, refinement, and independent judgment. The specific factors, supporting studies, and their reported relationships with student–GenAI interaction are summarized in Supplementary Table S6. As the studies differed in design and evidence, these relationships are interpreted as associations and contextual influences rather than established causal effects.

3.3.1. Personal Factors

Several learner characteristics were associated with differences in student–GenAI interaction, particularly AI literacy, prompt literacy and quality, prior experience, and disciplinary knowledge. Supriadi and Suherman (2026) reported that students using more structured and iterative prompts demonstrated stronger logical reasoning and greater engagement in identifying and correcting inaccurate AI responses. Similarly, Gallart et al. (2026) highlighted the role of prior problem-solving experience in how students approached mathematical modeling with ChatGPT. Although AI literacy and disciplinary knowledge may help students recognize the limitations of AI-generated information, the available evidence does not consistently establish these characteristics as independent predictors of interaction quality (Supplementary Table S6).

Other personal factors included disciplinary knowledge, metacognitive competence, confidence, and epistemic trust. Disciplinary knowledge provided a basis for recognizing questionable AI-generated outputs and determining when further verification was necessary (Reddy et al., 2024; Supriadi & Suherman, 2026), whereas comprehension difficulties, academic-writing confidence, and the need for clarification or validation influenced how students engaged with GenAI (Picardal, 2025). Behavioral evidence from Dominguez et al. (2026) further indicated that interaction effectiveness depended on the quality rather than quantity of prompting. AI-driven problem solving showed the strongest positive association with assignment performance ($\rho = 0.77$, $p < 0.001$), while query efficiency and prompt-refinement depth were also positively associated with performance. In contrast, query frequency was negatively associated with performance, indicating that frequent prompting alone was not a marker of more effective GenAI engagement.

Although these findings highlight the potential relevance of personal factors in student–GenAI interaction, the strength and nature of the supporting evidence varied across studies. Prompting practices, prior experience, and some learner characteristics were explored through reported behaviors, experiences, or performance-related measures. AI literacy, metacognitive competence, and epistemic trust were not consistently investigated as independent influencing factors. Thus, some of the relations were not directly established via empirical analysis but were deduced from the findings reported. Therefore, these factors should be considered as potentially associated with CT-related engagement rather than as established predictors of the quality of student interaction or critical thinking outcomes.

3.3.2. Instructional Factors

Instructional scaffolding shaped student–GenAI interaction through different mechanisms. Lin et al. (2026) embedded pedagogical structure within a 5E-guided GenAI coach that used Socratic questioning, hints, and corrective feedback to support chemistry problem-solving. In contrast, Buendía-García and Piris-Ruano (2025) structured learners' interactions through prompt scaffolds and incremental prompting during web-development tasks. Other studies similarly supported sustained engagement with AI-generated responses through progressive prompting (Gong et al., 2025) and contextual and iterative prompting (Supriadi & Suherman, 2026). Collectively, these findings suggest that scaffolding can operate either through instructional guidance embedded within the GenAI system or through external structures that guide how learners formulate, refine, and respond to their interactions with GenAI.

The influence of explicit prompt training was particularly evident (Garg et al., 2025). Engineering students who received structured prompt training, including the CLEAR framework, few-shot prompting, and chain prompting, achieved higher overall post-test performance than students who used GenAI without prompt training and those without access to GenAI. Compared with unguided GenAI use, the advantage of prompt training was particularly evident at the higher Bloom levels of application and analysis, whereas no significant difference was observed at the understanding level. Structured prompt guidance was also evident (Tassoti, 2024), where a prompting framework encouraged students to provide greater context, specificity, structure, and feedback when interacting with ChatGPT.

Instructional approaches also encouraged students to evaluate and verify AI-generated information. Reddy et al. (2024) required students to check AI-generated scientific information against external sources, while Matzakos and Moundridou (2025) required LLM results or hypotheses to be checked with Maxima. Similarly, Štuikys et al. (2025) and Chen et al. (2024) tested AI-supported engineering recommendations against physical or task-based evidence. Cruz-Angeles et al. (2026) further demonstrated different forms of instructional support across two phases. Phase I explicitly required students to verify AI-generated information against external sources, whereas Phase II used customized retrieval-augmented generation (RAG) assistants designed to provide Socratic guidance. The instructional factors identified across the studies, including structured prompt training, guided inquiry, scaffolding, feedback, and verification requirements, are further detailed in Supplementary Table S6. These approaches illustrate how instructional design may encourage students to remain actively involved in evaluating and refining AI-generated responses rather than relying on GenAI primarily for completed answers.

3.3.3. Contextual and Task-Related Factors

Student–GenAI interaction also varied according to task complexity, disciplinary demands, workload, resource availability, and AI access. Task complexity was particularly evident (Gallart et al., 2026). Across the two Fermi problems, autonomous use of ChatGPT

as a problem-solving support decreased from 45.8% to 25.6% as task complexity increased, while greater delegation to ChatGPT became more frequent. The relationship between task complexity and problem-solving profile was statistically significant, $\chi^2(2,266) = 8.81$, $p < 0.05$, V = 0.18, indicating that greater task complexity was associated with increased delegation of modeling decisions to GenAI. The nature of the STEM task also shaped interaction patterns: programming and web-development activities (Buendía-García & Piris-Ruano, 2025; Gong et al., 2025; Nathaniel et al., 2025) frequently involved iterative prompting, coding, debugging, and refinement, whereas robotics and engineering-design tasks (Chen et al., 2024; Štuikys et al., 2025) enabled students to evaluate AI recommendations through physical testing, troubleshooting, and observable system performance.

Practical learning conditions also influenced how students engaged with GenAI. Comprehension difficulties, time pressure and workload, uncertainty about research structure, confidence in grammar and academic writing, resource limitations, validation needs, and preferences for personalized assistance shaped GenAI use among students in the study by Picardal (2025). Technological conditions also formed part of the learning context, with Cruz-Angeles et al. (2026) highlighting differences in access to GenAI tools and platform capabilities that affected the forms of support available to students.

Comparisons between standard and premium customized GenAI conditions showed significant differences, especially for larger synthesis tasks. These tasks involved X-ray diffraction and combined differential scanning calorimetry and Fourier-Transform Infrared Spectroscopy (DSC/FTIR) reports and comprehensive final reports. However, several shorter or more discrete tasks showed no statistically significant differences. These data indicate that the importance of access to and capability in using GenAI may vary by the requirements of the STEM work.

The findings suggest that student–GenAI interaction was shaped by the interplay of learner capabilities, instructional support, and task demands. While personal factors influenced how students approached and evaluated AI-generated responses, instructional guidance provided opportunities for more structured engagement. Contextual and task-related conditions further influenced how students used GenAI to address learning challenges and whether they remained actively involved in reasoning and decision-making. These factors and their reported relationships with student–GenAI interaction are summarized in Supplementary Table S6. Consequently, the findings provide a basis for examining the critical-thinking processes identified in RQ3.

3.4. Critical-Thinking-Related Processes in Student–GenAI Interaction (RQ3)

Across the included studies, critical-thinking-related engagement with GenAI was represented through a set of recurring cognitive and interactional processes. The processes were synthesized into six higher-order analytic groups: (1) strategic questioning and task framing; (2) interpretation and analytical decomposition; (3) comparison, evaluation, and verification; (4) error detection and corrective modification; (5) justification, judgment, and independent decision-making; and (6) reflection, metacognition, and self-regulation. These groups summarize observed, reported, or self-reported manifestations of critical engagement. These processes centered on evaluating and verifying AI-generated information, detecting and correcting errors, and exercising reflection and judgment over whether AI-generated outputs should be accepted, modified, or rejected. Although direct CT measurement was limited to a small number of studies, the interaction behaviors summarized in Table 3 provide process-level evidence of CT-related engagement during interaction.

Table 3. Critical-thinking-related processes identified across the included studies.

CT-Related Process	Keywords	Manifestation in Student–GenAI Interaction	Illustrative Studies
Strategic questioning and task framing	Strategic questioning; task framing; prompt formulation and refinement; inquiry; follow-up questioning; hypothesis generation; information seeking; problem definition	Defining the goal, relevant information, constraints, expected output, or sequence of solution steps before or during prompting; formulating and refining prompts; asking follow-up questions, requesting explanations, and probing or challenging AI responses.	(Dominguez et al., 2026; Gong et al., 2025; Nathaniel et al., 2025; Picardal, 2025; Supriadi & Suherman, 2026; Tassoti, 2024)
Interpretation and analytical decomposition	Analytical thinking; analytical interpretation; logical reasoning; mathematical reasoning; problem decomposition; conceptual understanding; data analysis; scientific reasoning	Making sense of AI explanations, code, calculations, scientific information, or recommendations; breaking complex problems into components; and examining the relationships, assumptions, or reasoning underlying AI-generated responses or solutions.	(Buendía-García & Piris-Ruano, 2025; Gallart et al., 2026; Nathaniel et al., 2025; Picardal, 2025; Supriadi & Suherman, 2026; X. Zhang, 2025)
Comparison, evaluation, and verification	Critical evaluation; comparison of alternatives; evidence-based reasoning; verification; cross-checking; information reliability; accuracy assessment; independent judgment	Comparing AI outputs with independent solutions, literature, alternative tools, prior knowledge, or other evidence; assessing their accuracy, relevance, credibility, feasibility, or quality; and checking them against disciplinary knowledge, computational tools, external sources, or empirical observations.	(Chen et al., 2024; Cruz-Angeles et al., 2026; Gong et al., 2025; Matzakos & Moundridou, 2025; Štuikys et al., 2025; Supriadi & Suherman, 2026; X. Zhang, 2025)
Error detection and corrective modification	Error detection; error correction; debugging; code optimization; iterative refinement; problem-solving; solution improvement; hallucination detection	Recognizing hallucinations, faulty reasoning, inaccurate information, incorrect calculations, or problematic code; and revising prompts, code, calculations, designs, text, or solutions in response to identified errors or limitations.	(Buendía-García & Piris-Ruano, 2025; Chen et al., 2024; Gong et al., 2025; Nathaniel et al., 2025; Štuikys et al., 2025; Supriadi & Suherman, 2026; Tassoti, 2024)
Justification, judgment, and independent decision-making	Justification; evidence-based reasoning; logical reasoning; critical judgment; independent decision-making; problem-solving; conceptual reasoning; synthesis; autonomous learning	Providing reasons for accepting, modifying, or rejecting AI-generated solutions or recommendations; determining whether outputs are appropriate, credible, or suitable for use; and retaining responsibility for the final solution, interpretation, design, or decision.	(Chen et al., 2024; Cruz-Angeles et al., 2026; Gallart et al., 2026; Gong et al., 2025; Nathaniel et al., 2025; Reddy et al., 2024; Štuikys et al., 2025; Supriadi & Suherman, 2026)
Reflection, metacognition, and self-regulation	Reflection; reflective thinking; metacognition; metacognitive awareness; metacognitive planning and monitoring; self-regulation; autonomous learning; iterative improvement	Reflecting on one's reasoning, prompting strategies, AI credibility, or dependence on AI; monitoring the adequacy of AI responses and one's own interaction strategy; and deciding whether to revise, elaborate, seek additional evidence, or change strategy.	(Cruz-Angeles et al., 2026; Dominguez et al., 2026; Gong et al., 2025; Nathaniel et al., 2025; Picardal, 2025; Supriadi & Suherman, 2026; Tassoti, 2024)

Note. The processes represent recurring CT-related behaviors or cognitive processes reported across the included studies and should not be interpreted as equivalent to direct measurement of critical-thinking ability. Individual studies may contribute evidence to more than one process.

3.4.1. Strategic Questioning, Interpretation, and Analytical Decomposition

These processes concern how students define and cognitively organize a task before treating GenAI output as usable evidence. Strategic questioning is more than the production of a technically elaborate prompt; it involves specifying the problem, identifying constraints, requesting the form of explanation needed, and revising the request when the initial response is inadequate. In the 16-week engineering study (Dominguez et al., 2026), students' prompting behavior and refinement were examined in relation to multi-step case analysis, design, problem solving, and data-analysis tasks. Questioning also became critical when students moved beyond clarification to probe, challenge, or object to AI responses (Gong et al., 2025; Supriadi & Suherman, 2026; Tassoti, 2024).

On the other hand, interpretation and analytical decomposition become visible when the learner must decide how an AI response maps onto the structure of the disciplinary problem. These processes followed when students made sense of AI explanations, code, calculations, or recommendations and broke complex problems or generated reasoning into manageable components (Nathaniel et al., 2025; X. Zhang, 2025). In mathematical modeling (Gallart et al., 2026), students displayed different cognitive-delegation profiles while solving Fermi problems, and greater task complexity was associated with greater delegation of modeling decisions to ChatGPT. In web-development tasks (Buendía-García & Piris-Ruano, 2025), interaction logs documented incremental prompting, code explanation, troubleshooting, and debugging, providing process-level evidence that students repeatedly interpreted generated code and reorganized the task around emerging technical problems. Qualitative interaction artifacts and reflection logs in science-education research (Picardal, 2025) similarly showed contextual, iterative, file-based, and role-based prompting across conceptualization, methodology, data interpretation, and academic writing. These behaviors indicate a shift from merely soliciting an output toward structuring the problem and actively interpreting what the AI contributes.

3.4.2. Comparison, Evaluation, Verification, and Corrective Modification

There was consistent evidence of critical engagement, with students treating the output of the AI as a provisional claim to be checked independently. Students compared the results generated by the LLM to Maxima within the framework of engineering mathematics (Matzakos & Moundridou, 2025). This approach made the verification explicitly cross-tool and computational. Students in a biochemistry class were tasked with evaluating AI-generated scientific writing and verifying the information and citations against external sources (Reddy et al., 2024). Verification also occurred through the observed performance of systems and artifacts. Students applied the generated suggestions for coding, troubleshooting, physical testing, and refinement in their work on robotics and design projects (Štuikys et al., 2025). Thus, the appropriateness of an AI recommendation was constrained by how well the robot actually performed as intended. Similarly, engineers who created intelligent vehicles (Chen et al., 2024) utilized code validation, iterative feedback, proportional-integral-derivative (PID) controller tuning, and real-world testing. If the AI recommendations did not meet either physical or technical evidence, they could be ignored or modified. Thus, verification was either epistemic, computational, or empirical, depending on the STEM task being performed. Error recognition, however, did not always lead to proper correction. In the chemistry prompting study (Tassoti, 2024), students progressed from relatively simple copy-and-paste prompting toward more structured Five-S prompting, accompanied by recognition of inaccuracies and refinement of their prompts. For this review, corrective modification represents a distinct form of critical engagement, as learners move beyond detecting discrepancies to revising, correcting, or redirecting their interaction with GenAI.

3.4.3. Justification, Judgment, Independent Decision-Making, and Metacognitive Self-Regulation

Students demonstrated greater critical engagement when they moved beyond simply checking AI-generated answers to evaluating how the information could be used and justifying their decisions. In the programming context, interaction patterns included high-demand questioning, objection, critique, and summarization (Gong et al., 2025). Additionally, critical thinking was assessed as a dimension within a broader computational-thinking measure, providing direct self-reported CT evidence alongside coded interaction behaviors. In integral calculus, Supriadi and Suherman (2026) also reported a relationship between prompt quality and performance in mathematical logical reasoning. The interaction records showed iterative and debugging patterns involving error identification, error verification, corrective prompting, and justification.

Programming evidence also showed that judgment can be embedded in performance rather than reported only through perceptions. In C++ problem-solving tasks (Nathaniel et al., 2025), students engaged in prompting, response modification, scaffolded feedback, iterative refinement, and coding, while critical thinking was assessed within a tailored performance-based HOTS framework. In database education (X. Zhang, 2025), students first developed independent problem-solving solutions and then used AI-assisted comparison, creating an instructional structure in which the learner retained an initial basis for judgment before consulting the system. Such designs are especially relevant to critical-thinking research because they make it possible to distinguish AI-supported reconsideration from AI-first substitution of the learner's own reasoning.

Reflection, metacognition, and self-regulation extend judgment across time by requiring students to monitor both the AI and their own interaction strategy. In life sciences and nanotechnology (Cruz-Angeles et al., 2026), students reported credibility questioning and external verification, while customized retrieval-augmented assistants in a later phase used Socratic guidance. The first phase provided self-reported CT-related processes, whereas the second phase assessed academic performance rather than critical thinking directly. These findings demonstrated that reflective or evaluative behaviors may be consistent with critical thinking, but they should not be treated as proof of improved CT ability unless supported by appropriate direct or performance-based measures. This process-based pattern provides the basis for examining the relationship between GenAI interaction and critical thinking in RQ4.

Across the reviewed studies, justification, judgment, independent decision-making and self-regulation reflected learners' retention of responsibility for the final epistemic outcome of the interaction. In these forms of interaction, critical engagement transcends merely determining the accuracy of an AI-generated response. It entails articulating justifications for accepting, modifying, or rejecting the response, evaluating the necessity for additional evidence or validation, and adapting future prompting, reasoning, or problem-solving approaches accordingly. The reviewed evidence most consistently fits with critical engagement when students frame the task, interpret and decompose AI-supported reasoning, compare and verify outputs against independent evidence, correct identified limitations, justify acceptance or rejection, and regulate their subsequent strategy. These patterns indicate noticeable, observable processes by which student-GenAI interaction can facilitate, limit, or delegate the cognitive tasks associated with critical thinking.

3.5. Relationships Between GenAI Interaction and Critical-Thinking-Related Processes (RQ4)

The synthesis revealed two consistent patterns of critical engagement across the examined studies. Interaction was more consistent with cognitive scaffolding when students remained involved in reasoning, evaluation, verification, correction, and final judgment,

whereas it was more consistent with cognitive offloading when substantial cognitive work was delegated to GenAI with limited scrutiny. These patterns represent evaluative interpretations instead of distinct causal pathways.

3.5.1. GenAI as Cognitive Scaffold

Scaffolding-consistent patterns were evident when GenAI outputs became material for further student reasoning. Gong et al. (2025) found that higher-CT students exhibited debating and summarizing interaction patterns characterized by higher-demand questioning, objection, critique, and synthesis. Supriadi and Suherman (2026) reported an association between prompt quality and mathematical logical reasoning, while iterative and debugging-oriented interactions involved error identification, corrective prompting, verification, and justification. Nathaniel et al. (2025) further provided performance-based CT evidence within a broader higher-order thinking skills (HOTS) programming framework.

Similar patterns were evident across studies in instructional approaches that required students to move beyond output generation. 5E-oriented Socratic questioning (Lin et al., 2026), LLM–CAS comparison and computational cross-checking (Matzakos & Moundridou, 2025), explicit prompt training (Garg et al., 2025), external verification (Cruz-Angeles et al., 2026; Reddy et al., 2024), physical testing of engineering and robotics recommendations (Chen et al., 2024; Štuikys et al., 2025), and iterative code interpretation and debugging (Buendía-García & Piris-Ruano, 2025) required students to examine, test, and refine AI-generated material. These findings suggest that GenAI is most consistent with a scaffolding role when the learner retains responsibility for determining whether generated information is accurate, relevant, and valid.

3.5.2. GenAI as Cognitive Substitute/Cognitive Offloading

A contrasting pattern emerged when students delegated substantial cognitive work to GenAI. Gong et al. (2025) found that lower-level interaction was characterized by low-demand questions, direct guidance, and frequent acceptance of AI responses. Similarly, Supriadi and Suherman (2026) identified a naive copy-paste pattern associated with blind acceptance and weaker logical reasoning, while Gallart et al. (2026) showed that some students adopted a high-delegation PS-expert profile, transferring much of the mathematical modeling process to ChatGPT, particularly as task complexity increased. Tassoti (2024) likewise found that students initially relied on simple copy-paste prompting before developing more structured and corrective forms of interaction.

Dominguez et al. (2026) further showed that greater query frequency was negatively associated with academic performance, whereas query efficiency and refinement were more positively associated with performance. Thus, these findings suggest that interaction was more consistent with cognitive offloading when learners relied on generated solutions with limited evaluation or transferred substantial reasoning and decision-making to GenAI. In contrast, interaction was more consistent with critical engagement when learners continued to question, verify, correct, and make the final judgment themselves.

Across the reviewed evidence, cognitive responsibility provides an evidence-informed interpretive lens for distinguishing scaffolding-consistent from offloading-consistent interaction. Critical engagement was more evident when learners retained responsibility for questioning, verification, correction, reasoning, and final judgment.

3.6. Integrated Synthesis

Across RQ1–RQ4, the findings indicate that GenAI use alone did not distinguish more from less critical engagement. Instead, the reviewed studies differed in the quality of student–GenAI interaction and in the extent to which learners remained involved in

cognitively consequential work. More critically engaged interaction generally involved questioning, interpretation, decomposition, comparison, verification, correction, justification, and judgment, whereas less critically engaged interaction involved greater reliance on AI-generated responses with limited scrutiny.

Integrated CT-Related Processes Structure

The six CT-related process clusters identified in Section 3.4 were examined collectively to clarify their functional relationships during student–GenAI interaction (Figure 3). Rather than representing a fixed sequence, the clusters describe potentially overlapping and iterative cognitive activities. Students may move between task framing, interpretation, evaluation, correction, justification, and reflection as they examine and respond to AI-generated content. Reflection and self-regulation may operate across these activities, influencing whether learners continue to question, revise, or accept AI-generated responses. This integrated representation is interpretive and does not imply that the clusters constitute empirically validated stages of critical-thinking development.

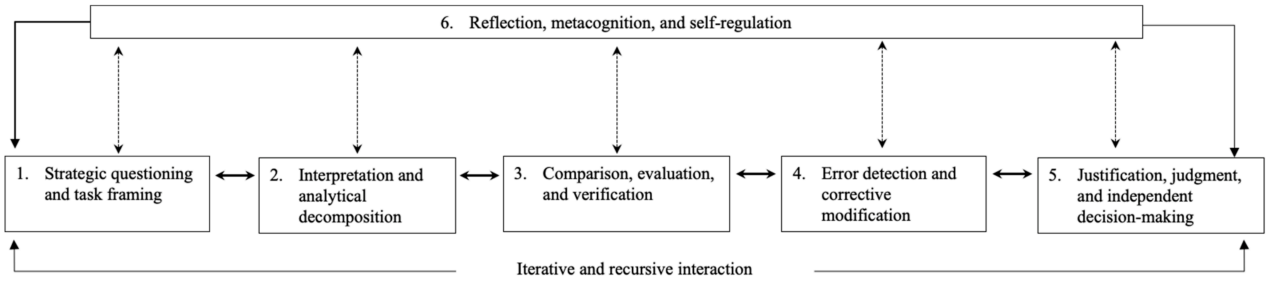


Figure 3. Integrated critical-thinking-related processes during student–GenAI interaction. The figure summarizes six recurring clusters identified across the included studies: strategic questioning and task framing; interpretation and analytical decomposition; comparison, evaluation, and verification; error detection and corrective modification; justification, judgment, and independent decision-making; and reflection, metacognition, and self-regulation. The connections indicate the iterative and interrelated nature of these processes, with learners potentially moving among processes as they evaluate and refine AI-generated outputs. **Note.** Bidirectional arrows indicate iterative movement between related processes rather than fixed sequential stages. Reflection, metacognition, and self-regulation operate across the other processes and may redirect interaction among them as needed. The outer feedback connection represents the recursive nature of student–GenAI interaction rather than a prescribed pathway.

A consistent descriptive distinction concerned the degree of learner involvement after GenAI generated an initial response. More critically engaged interaction involved continued learner participation through questioning, verification, correction, justification, and final judgment. Less critically engaged interaction involved greater delegation of these activities to GenAI.

Variations in interaction were also linked to contextual/task-related, instructional, and personal factors. Students' interactions with GenAI seemed to be influenced by prior knowledge, prompting competence, metacognition, instructional scaffolding, verification requirements, task complexity, workload, and access conditions. Nevertheless, these factors did not establish a common explanatory pathway because they were typically investigated independently across studies. Therefore, rather than explaining causal effects on critical-thinking ability, the synthesis offers stronger evidence for identifying recurrent interaction patterns, CT-related processes, and conditions associated with critical engagement. The evidence gaps discussed in the next section are informed by this distinction.

Building on the descriptive patterns found in Section 3, Figure 4 offers an evidence-based framework for understanding how student-GenAI interaction, the distribution of cognitive responsibility, and CT-related processes may be influenced by personal, instructional, and contextual/task-related conditions. Rather than a proven causal model, the framework depicts relationships that the synthesis suggests.

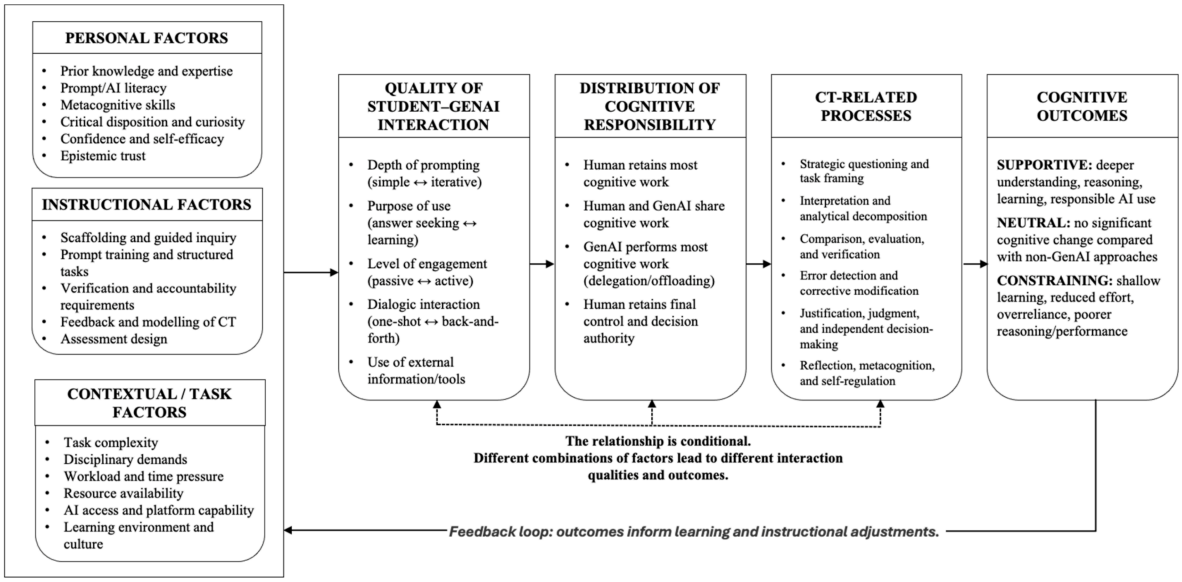


Figure 4. Integrated framework of factors, student–GenAI interaction, cognitive responsibility, and critical-thinking-related outcomes in STEM learning. The framework illustrates how personal, instructional, and contextual/task-related factors may shape the quality of student–GenAI interaction and the distribution of cognitive responsibility. These conditions may influence the enactment of CT-related processes and may be associated with supportive, neutral, or constraining cognitive outcomes. The relationships are conditional and iterative rather than strictly linear, with outcomes potentially informing subsequent learning and instructional adjustments.

3.7. Evidence Gaps

Despite growing empirical attention to GenAI in STEM learning, the synthesis revealed several important gaps that limit claims about its relationship with critical thinking. These gaps concern construct measurement, independence of CT outcomes, causal and developmental designs, interaction-level process evidence, integration of influencing factors, the range of STEM learning contexts examined, and the extent to which the findings apply to other STEM contexts and GenAI systems.

First, critical thinking was rarely measured directly and consistently. Among the included studies, Gong et al. (2025) directly assessed CT as one dimension within computational thinking, while Nathaniel et al. (2025) incorporated CT within a tailored performance-based HOTS assessment. Supriadi and Suherman (2026) directly measured the closely related construct of mathematical logical reasoning. Most other studies inferred CT-related engagement from behaviors, artifacts, or self-reports such as questioning, comparison, error detection, verification, reflection, and judgment. These processes provide meaningful evidence of critical engagement, but their occurrence is not equivalent to validated measures of CT ability. Future studies should therefore combine direct or performance-based CT measures with interaction-level evidence to determine whether particular forms of GenAI engagement correspond to measurable changes in critical-thinking performance.

A closely related gap concerns outcome independence. In many AI-assisted tasks, the same GenAI interaction that is being studied also contributes to the product used to infer learning. This makes it difficult to separate student capability from AI-supported task

success. Future studies should therefore include assessments completed without AI, such as follow-up problem-solving tasks, delayed tests, oral explanations, or transfer tasks, to determine whether students can apply critical thinking independently.

Second, the evidence base provided limited support for causal and developmental conclusions. Many studies used small or convenience samples, single courses or institutions, short interventions, observational or case-study designs, or lacked comparison groups. Even where significant relationships were reported, such as between prompt quality and logical reasoning (Supriadi & Suherman, 2026) or between interaction characteristics and performance (Dominguez et al., 2026), these findings do not establish that particular GenAI interaction patterns caused the observed cognitive outcomes. Longitudinal and repeated-measures designs are also needed to determine whether students become more strategic and evaluative with experience or increasingly transfer reasoning to GenAI over time.

Third, student–GenAI interaction was not examined at the same depth across studies. Some studies analyzed chat logs, screen recordings, prompts, responses, code, or other interaction artifacts (Buendía-García & Piris-Ruano, 2025; Dominguez et al., 2026; Gallart et al., 2026; Gong et al., 2025; Matzakos & Moundridou, 2025; Nathaniel et al., 2025; Picardal, 2025; Supriadi & Suherman, 2026; Tassoti, 2024), whereas others relied more heavily on surveys, performance outcomes, or instructional conditions. Consequently, the evidence was stronger for identifying broad patterns of GenAI use than for explaining the moment-to-moment processes through which students moved from receiving an AI response to questioning, verifying, modifying, or rejecting it. Future research should capture complete interaction sequences and connect specific student actions with subsequent reasoning and learning outcomes.

Fourth, very few studies used a single analytical framework to look at personal, instructional, contextual, interactional, and cognitive factors. Prompt literacy, disciplinary knowledge, prior experience, confidence, metacognition, epistemic trust, instructional scaffolding, verification requirements, task complexity, workload, resource availability, and access to GenAI tools are some of the factors that this review found to be potentially relevant. Nevertheless, the majority of research only looked at a portion of these variables, making it challenging to ascertain how they interact or which are most closely linked to significant student-GenAI engagement. To fully comprehend their individual and collective contributions to the quality of these interactions, more research is required.

Lastly, there was an uneven distribution of the evidence across learning tasks, GenAI systems, STEM fields, and educational levels. Programming, engineering, mathematics, and related problem-solving activities received greater attention, while other STEM disciplines and younger learners were less frequently represented. Additionally, the studies varied in terms of instructional strategies, access conditions, and GenAI tools and versions. Results from one learning environment or system might not necessarily apply to others as GenAI technologies develop further. Therefore, similar interaction patterns and critical-thinking processes should be investigated in future research across various STEM fields, educational levels, learning activities, and GenAI environments.

All things considered, these gaps indicate that the existing literature provides a better understanding of how students engage with GenAI and the circumstances surrounding critical engagement than it does of whether these interactions result in long-term gains in critical-thinking (CT) skills. Future studies ought to go beyond merely contrasting students who use GenAI with those who don't. More focus should be placed on how students challenge, validate, improve, and assess responses produced by AI; how students and GenAI share cognitive responsibility; and how learning environments affect these processes over time. It would be possible to determine whether these interaction patterns lead to quantifiable learning gains through longitudinal and intervention-based studies that use

direct measures of CT. Such evidence would also support further testing and refinement of the integrated framework presented in Figure 4.

4. Discussion

This systematic literature review synthesized 17 empirical studies examining student–GenAI interaction and critical-thinking-related processes across STEM learning contexts. The discussion differentiates findings reported by the primary studies from cross-study interpretations and the conceptual contributions of this review. Although the included studies documented various forms of evaluative, reflective, and problem-solving engagement, direct evidence of changes in independent critical-thinking (CT) ability remains limited. Accordingly, the discussion considers how interaction patterns, learner responsibility, and instructional conditions may inform the interpretation of CT-related engagement without assuming causal relationships.

4.1. From GenAI Use to Quality of Student–GenAI Interaction

A central finding of this review is the distinction between GenAI use and critical engagement. This distinction also marks an important contribution relative to existing syntheses. Rather than treating GenAI use, prompting competence, or AI-supported performance as proxies for critical thinking, the present review examines the cognitive processes enacted during student–GenAI interaction and the extent to which learners remain responsible for consequential reasoning and judgment.

The contribution of this review therefore extends beyond identifying potentially beneficial or problematic forms of GenAI use. By integrating interaction patterns, factors associated with those interactions, and CT-related processes, the synthesis identifies the distribution of cognitive responsibility as an evidence-informed interpretive lens for understanding variation in student–GenAI interaction and critical engagement in STEM learning. From this perspective, attention shifts from how much or how sophisticatedly students use GenAI to the cognitively consequential work they retain, share, or delegate during the interaction.

Dominguez et al. (2026) found that higher query frequency was negatively associated with performance, whereas greater query efficiency and refinement were positively associated with performance. Similarly, Supriadi and Suherman (2026) identified distinct interaction patterns, including naïve copy-paste and iterative/debugging, while Gong et al. (2025) distinguished acceptance-oriented interactions from more evaluative forms of engagement. Taken together, these findings suggest that the frequency or technical form of prompting may be less informative than how students engage with, evaluate, and act on AI-generated responses.

Evidence beyond the included studies provides further support for this distinction. Farrokhnia et al. (2026) found that although chain-of-thought prompting produced higher-quality GenAI feedback, this advantage did not translate into significantly greater improvement in students' subsequent revisions. This finding suggests that the quality of AI-generated output alone may not explain learning outcomes; how learners interpret, evaluate, and act on that output also matters.

Similarly, Ding et al. (2026) found that engagement with LLMs could remain limited when interactions stopped at initial or surface-level questions, whereas greater cognitive engagement in prompting was associated with larger learning gains. These findings further distinguish the act of generating prompts from sustained cognitive engagement with AI-generated content through follow-up interaction and critical evaluation.

This perspective highlights the need for a stronger emphasis on prompting skills and AI literacy in the wider literature. Prompting can help learners clarify goals, establish

constraints, request explanations, and obtain meaningful feedback. However, prompt sophistication should not be treated as equivalent to critical thinking. A technically sophisticated prompt may still delegate substantial reasoning to GenAI, whereas a relatively simple prompt may initiate a highly evaluative interaction if the learner subsequently questions, verifies, corrects, and judges the response. The educationally relevant issue, therefore, is not prompt form alone but the quality of cognitive engagement maintained throughout the interaction.

The findings converge with previous research emphasizing AI literacy, prompting competence, and cognitive engagement as important considerations in GenAI-supported learning (Pangestu et al., 2026; Raitskaya & Tikhonova, 2025; Reicho et al., 2026). However, the present review extends this discussion by distinguishing the formulation of prompts from the cognitive activities that follow AI-generated responses. The evidence indicates that learners may use sophisticated prompts without necessarily retaining responsibility for evaluating or justifying the resulting outputs. Conversely, relatively simple prompting may initiate critical engagement when followed by questioning, comparison, verification, and correction. Thus, the review contributes a process-oriented perspective in which the quality of student–GenAI interaction is examined through the cognitive and epistemic responsibilities retained by learners, rather than inferred from the frequency or sophistication of AI use alone.

4.2. Cognitive Responsibility as an Evidence-Informed Interpretive Lens

The synthesis identifies the distribution of cognitive and epistemic responsibility as a useful evidence-informed interpretive lens for explaining variation in student–GenAI interaction. Cognitive responsibility refers to the extent to which students remain engaged in cognitively consequential activities, such as defining problems, decomposing tasks, analyzing information, evaluating alternatives, correcting errors, and developing solutions. Epistemic responsibility refers to the extent to which students retain responsibility for determining whether information is reliable, sufficient, relevant, and defensible, and for making the final judgment or decision.

The modeling profiles identified by Gallart et al. (2026) illustrate the importance of this distinction. Students varied in the extent to which they delegated the modeling process to ChatGPT, with greater delegation associated with more difficult tasks. Across the reviewed studies, responsibility did not appear to be a fixed characteristic of the learner but varied with factors such as task complexity, prior knowledge, time pressure, confidence, instructional structure, and the affordances of the AI system. Importantly, cognitive offloading is not necessarily detrimental. Delegating routine, peripheral, or already-mastered processes may allow learners to direct greater attention toward higher-order thinking. The pedagogical concern arises when GenAI performs the reasoning, evaluation, or judgment that the learning activity is intended to develop or assess, particularly when learners have limited independent basis for evaluating the resulting output.

The distinction between scaffolding and substitution is therefore better understood as a range of interaction states rather than as a fixed characteristic of GenAI. Interaction is more consistent with scaffolding when learners retain responsibility for interpreting, verifying, correcting, and making decisions. In contrast, interaction is more consistent with substitution when these functions are delegated to GenAI with limited independent scrutiny. Although this pattern recurred across the reviewed studies, the evidence does not establish cognitive and epistemic responsibility as empirically validated mechanisms. Rather, the synthesis supports these constructs as evidence-informed interpretive concepts. Direct empirical testing is needed before either can be treated as a mediator, moderator, or causal mechanism in the relationship between GenAI interaction and critical-thinking outcomes.

A further contribution of this synthesis is to provide a more differentiated account of learner responsibility during GenAI interaction. Whereas existing literature has raised concerns about cognitive offloading, over-reliance, and the preservation of human judgment, the present review distinguishes cognitive responsibility for performing reasoning-related work from epistemic responsibility for evaluating whether AI-generated information is reliable, sufficient, relevant, and defensible. This distinction helps explain why GenAI cannot be characterized as inherently scaffolding or substitutive; rather, its role depends partly on which cognitively and epistemically consequential activities remain under learner control during a particular interaction.

The proposed perspective also intersects with self-regulated learning (SRL), particularly learners' planning, monitoring, evaluation, and regulation of their learning activities. However, SRL encompasses broader motivational, behavioral, and metacognitive processes across learning contexts, whereas the present synthesis focuses more specifically on the distribution of reasoning-related and epistemic responsibilities during student–GenAI interaction. Cognitive responsibility therefore should not be considered a replacement for SRL, cognitive offloading, or scaffolding. Instead, it offers a complementary analytical perspective for examining which cognitive activities learners perform themselves, which they delegate to GenAI, and how they evaluate and exercise judgment over AI-generated outputs.

4.3. Scaffolding-Consistent Interaction

The findings suggest that GenAI is more likely to serve as a learning scaffold when students use AI-generated content as a starting point for further reasoning rather than accepting it as a final answer. In terms of education, it doesn't matter how well students create prompts or how sophisticated the AI tool is; what matters is how much control they have over analyzing data, creating justifications, and coming to conclusions. Opportunities to compare, assess, and improve AI-generated solutions can keep students actively engaged in reasoning and judgment, as demonstrated by Aruleba et al. (2025). These actions, however, do not always signify that students' capacity for independent critical thinking (CT) has improved.

This interpretation is consistent with Q. Sun and Yin (2026), who described critical engagement with GenAI as a process involving questioning, interpretation, critique, refinement, and selective use of AI-generated feedback. In practical terms, this entails creating educational exercises that motivate students to challenge AI answers, verify their accuracy, and determine which recommendations are suitable for application. Similar to this, Pramanik and Husin (2025) suggested the Retrieval-Augmented Pedagogical Assistant (RAPA) and Critical AI Analysis Assignment as teaching strategies that integrate structured evaluation tasks with source-grounded AI responses. These methods show how students can continue to actively participate in evaluating AI-generated content instead of depending solely on it.

These viewpoints emphasize the significance of teaching strategies that uphold students' accountability for independent judgment, reasoning, and verification. The evidence that is currently available does not show consistent improvements in general CT ability, even though such approaches may promote CT-related engagement. Therefore, their main educational value is found in giving students the chance to think critically, assess information, and make well-informed decisions during GenAI-supported learning.

4.4. Risks of Cognitive Substitution and Uncritical Reliance

The findings highlight the potential educational risks of delegating cognitively demanding tasks to GenAI without sufficient evaluation or independent reasoning. As suggested by (Cruz-Angeles et al., 2026; Gallart et al., 2026; Gong et al., 2025; Supriadi & Suherman, 2026),

interaction with GenAI does not necessarily indicate meaningful cognitive engagement. The central concern is not the use of AI assistance itself, but the extent to which students transfer responsibility for problem formulation, evaluation, and decision-making to the system.

From a learning perspective, cognitive substitution may reduce opportunities for students to independently construct explanations, assess evidence, identify errors, and justify conclusions. This distinction is particularly relevant in STEM education, where producing a correct answer does not necessarily demonstrate understanding of the underlying concepts or reasoning processes. Consequently, interaction frequency, prompt quantity, and task completion should not be treated as sufficient indicators of critical-thinking engagement or learning quality.

These findings underscore the importance of distinguishing productive cognitive offloading from uncritical reliance. While delegating routine tasks may support efficiency, excessive delegation of reasoning and verification may limit opportunities to develop and demonstrate independent judgment. However, the reviewed evidence does not establish that GenAI reliance directly causes deterioration in critical-thinking ability. Rather, it highlights the need for instructional and assessment practices that make students' reasoning, verification, and decision-making visible, even when GenAI assistance is available.

4.5. Personal, Instructional, and Contextual Conditions Shaping Critical Engagement

The findings indicate that the quality of student–GenAI interaction and the distribution of cognitive responsibility are shaped by the conditions in which learning occurs. Personal, instructional, and contextual/task-related factors may influence whether students engage critically with AI-generated information or delegate substantial cognitive work to the system. These conditions are particularly relevant because similar forms of GenAI use may involve different levels of student reasoning, verification, and independent judgment.

At the personal level, prompt literacy, disciplinary knowledge, prior experience, competence, confidence, and epistemic trust emerged as relevant influences on student–GenAI interaction. Disciplinary knowledge is particularly important because evaluating AI-generated responses requires sufficient understanding to recognize inaccuracies, omissions, or questionable assumptions. This presents an educational challenge: students who require greater assistance may also have less developed knowledge for independently assessing the reliability of AI-generated information.

Instructional design represents a more directly modifiable influence. Learning activities that require independent problem formulation, verification, justification, and empirical testing may help preserve students' responsibility for cognitively demanding tasks. Complementary evidence beyond the 17 included studies provides additional support for this instructional interpretation. Bai et al. (2026) reported that theory-informed prompt-engineering scaffolding was associated with improved prompting proficiency and academic performance, alongside qualitative evidence of metacognitive awareness and critical evaluation. Their incorporation of gradually reduced instructional support further illustrates how scaffolding may encourage increasing learner independence rather than continued dependence on external guidance. However, these findings should not be interpreted as direct evidence of improvements in independently assessed critical thinking.

Contextual and task-related conditions also influence how students engage with GenAI. Gallart et al. (2026) associated greater task complexity with increased cognitive delegation, while Picardal (2025) identified workload, comprehension difficulties, confidence, resource limitations, and validation needs as relevant factors. Cruz-Angeles et al. (2026) additionally highlighted differences in access to GenAI tools and platform capabilities. These findings suggest that cognitive responsibility is not necessarily a stable characteristic of individual learners but may vary across tasks, instructional settings, and technological environments.

These findings reinforce the conceptual contribution of the evidence-informed framework presented in Figure 4. Rather than treating GenAI use as inherently supportive or detrimental to critical thinking, the framework emphasizes how learning conditions may influence the quality of student engagement and the cognitive responsibilities retained or delegated during interaction. This perspective highlights the importance of designing instructional environments that support verification, independent reasoning, and reflective judgment while recognizing that interaction patterns are context-dependent. The proposed relationships remain interpretive and require empirical validation, particularly through studies examining how these conditions influence independently assessed critical-thinking outcomes.

4.6. Implications for STEM Teaching and Assessment

The findings have several implications for STEM teaching and assessment. GenAI-supported activities may be designed around cognitively consequential interactions rather than focusing primarily on output generation. Students may, for example, be required to identify limitations in AI-generated responses, compare them with independent solutions or authoritative evidence, document corrections, justify decisions to accept or reject particular outputs, and reflect on how GenAI influenced their reasoning.

Emerging instructional designs illustrate how these principles can be operationalized. Schapiro et al. (2026), for example, structured GenAI-supported learning around explicit reflection, iterative prompting, evaluation, revision, and accountability. Their REACT framework required students to review AI-generated material, evaluate its adequacy, make adjustments, formulate clarifying or follow-up prompts, and determine whether subsequent responses represented an improvement. Such designs illustrate how GenAI-supported assignments can make students' evaluative reasoning and revision decisions visible rather than assessing only the AI-assisted final product.

Another advantage of prompt instruction is that it can go beyond techniques for getting accurate or detailed answers. Artificial intelligence (AI) and prompt literacy may include epistemic practices such as when to challenge an answer, what counts as enough evidence for verification, how to handle uncertainty or inconsistency, and when independent disciplinary judgment should take precedence over an AI assessment.

Assessment practices could include weighing the final product and the reasoning process between students and GenAI. Some of the evidence that could be used to determine whether or not students retained responsibility for important cognitive decisions includes chat histories, verification records, revision trails, debugging logs, comparison matrices, reflective explanations, and oral defenses. It may be especially useful to employ such strategies in situations where polished AI-assisted outputs make it difficult to infer student comprehension based solely on the final artifact.

Most importantly, educators can align the level and type of support provided by GenAI with the learning objective. If the desired outcomes include mathematical modeling, scientific reasoning, programming logic, or engineering judgment, then GenAI should consistently allow students to demonstrate their independent understanding of the cognitive processes being evaluated. AI, on the other hand, may be used to automate peripheral operations in an appropriate manner when doing so enables greater attention to be paid to the higher-order process that is being monitored.

4.7. Implications for Future Research

Given the evidence gaps identified in Section 3.7, future research should move beyond general comparisons of people who use GenAI and people who do not. Instead, research should explore the relationship between specific patterns of interaction with GenAI and

independently measured critical thinking (CT) outcomes. By combining logs of interactions and direct CT assessments, we can investigate if certain behaviors (such as evaluation, revision and questioning of AI-generated responses) relate to better learning outcomes. Sequential analysis, process tracing, and mixed-method approaches may be especially useful in determining when students move from receiving responses from AI to evaluating or extending those responses.

Future studies will need to differentiate between AI-assisted performance and independent learning. Delayed no-AI tasks, transfer problems, oral explanations, or parallel assessments can be used, respectively, to find out if the CT processes that were shown during the interaction with GenAI still exist after the AI support was taken away. These approaches may be useful in testing whether observed improvements are attributable to long-term learning and transferable critical thinking skills, or short-term gains using GenAI.

Longitudinal, controlled, and multi-institutional studies are needed to determine whether repeated GenAI use promotes increasingly strategic and evaluative interaction or greater dependence on AI. Future research should test moderators suggested by the synthesis, including disciplinary expertise, AI/prompt literacy, metacognition, epistemic trust, task complexity, time pressure, instructional scaffolding, verification requirements, and access conditions. Because GenAI systems change rapidly, studies should also report model and version, access level, relevant system configuration, task instructions, prompting conditions, and the role of retrieval or other auxiliary tools. The evidence-informed framework in Figure 4 can then be treated as a set of testable relationships: whether these conditions predict interaction quality, whether interaction quality is associated with the distribution of cognitive responsibility, and whether retained responsibility predicts independently measured CT-related outcomes.

4.8. Limitations of the Review

This review has several limitations. First, although the literature search covered Scopus, Web of Science, and IEEE Xplore, the specificity of the Boolean search strategy may have limited the comprehensiveness of the evidence base. In particular, the mandatory inclusion of the terms "prompt engineering" and "critical thinking" may have excluded relevant studies describing comparable student–GenAI interaction practices or cognitive processes using alternative terminology, including iterative dialogue, human–AI collaboration, AI-assisted reasoning, metacognition, and evaluation or verification of AI-generated outputs. Consequently, the findings should be interpreted within the scope of the applied search strategy rather than as an exhaustive representation of student–GenAI interaction research in STEM education. The language, publication-type, and accessibility restrictions may also have introduced selection and publication biases.

Second, only 17 studies met the final eligibility criteria, and the evidence base was highly heterogeneous in STEM discipline, educational level, task, GenAI system, study design, sample size, intervention duration, and outcome measurement. This heterogeneity supported the use of narrative and thematic synthesis rather than meta-analysis and limits the extent to which findings can be generalized across STEM contexts and learner populations.

Third, critical thinking was directly assessed in only a small proportion of the included studies. Much of the evidence consisted instead of related reasoning measures, self-reports, performance evidence, or behavioral and process indicators. Consequently, the review provides stronger evidence for identifying patterns of CT-related engagement during student–GenAI interaction than for establishing changes in independent CT ability. The findings should therefore not be interpreted as demonstrating that particular forms of GenAI interaction improve or diminish critical-thinking capacity.

Fourth, the six CT-related process clusters and the constructs of cognitive and epistemic responsibility are review-level analytical interpretations derived from patterns across the included studies rather than constructs independently validated within a common empirical model. Although they provide an evidence-informed framework for organizing the reviewed findings, their relationships with interaction quality and CT outcomes require direct empirical testing and independent validation.

Lastly, the rapid evolution of GenAI sets another boundary for the temporal generalizability of the findings. As model capabilities, versions, access levels, retrieval mechanisms, and platform functionalities evolve, so may how students interact with these systems and what kinds of cognitive work can be delegated to them. Thus, results related to specific GenAI configurations should be interpreted according to their technology and instructional context, and may not generalize directly to future systems.

5. Conclusions

This systematic literature review synthesized 17 empirical studies examining student–GenAI interaction and critical-thinking (CT)-related processes in STEM education. The findings revealed recurring interaction patterns, ranging from basic prompting and question answering to more iterative forms of engagement involving problem decomposition, evaluation, verification, and the refinement or rejection of AI-generated responses. Across the reviewed studies, six clusters of CT-related processes were identified: strategic questioning and task framing; interpretation and analytical decomposition; comparison, evaluation, and verification; error detection and corrective modification; justification, judgment, and independent decision-making; and reflection, metacognition, and self-regulation. These processes were reported, observed, or inferred to varying degrees across the studies and should not be interpreted as direct evidence of improvements in students' independent CT ability.

This synthesis suggests that the educational value of GenAI use depends not only on whether or how frequently students interact with AI, but also on what cognitive activities they engage in while interacting with AI. In particular, challenging, contrasting, validating, correcting, and justifying AI-generated responses represent forms of critical engagement identified across the reviewed studies. The review also proposes cognitive and epistemic responsibility as complementary interpretive concepts for understanding how learners retain, share, or assign reasoning and evaluative decisions in GenAI-supported learning.

The main conceptual contribution of this review is an evidence-informed framework that integrates student–GenAI interaction, the distribution of cognitive responsibility, and critical-thinking (CT)-related processes with personal, instructional, and contextual or task-related factors. The framework brings together relationships identified across the reviewed studies and provides a basis for future empirical investigation. However, these relationships remain conceptual and require further empirical testing before causal conclusions can be drawn.

Importantly, the findings do not establish that GenAI-supported interaction consistently improves or diminishes independent critical-thinking ability. There was little direct CT assessment, and the majority of the evidence came from performance indicators, self-reports, related reasoning tests, and observed or recorded interaction behaviors. As a result, conclusions concerning how critical thinking-related engagement appears during student–GenAI interaction are more strongly supported by the review than conclusions regarding changes in students' independent CT capabilities.

Future research should examine these relationships through longitudinal and controlled studies that combine interaction data with validated CT assessments, delayed no-AI tasks, and transfer measures. Such approaches are necessary to determine whether critical-

thinking-related behaviors demonstrated during GenAI-supported activities translate into sustained and independently demonstrated critical-thinking abilities. For STEM education, the findings underscore the importance of designing learning activities in which GenAI supports students' reasoning while preserving their responsibility for evaluation, justification, and independent judgment.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/educsci16122013/s1, Table S1. Methodological Quality Assessment of Included Studies Using MMAT 2018; Table S2. Studies Excluded After Full-Text Assessment; Table S3. Study-Level Evidence and Thematic Mapping of CT-Related Cognitive Processes; Table S4. Consolidated Distribution by Primary Theme; Table S5. Summary of Student–GenAI Interaction Patterns; Table S6. Factors Associated with Student–GenAI Interaction Across the Reviewed Studies; and Table S7. Database-Specific Search Strategies, Execution Dates, and Applied Filters. Refs. (Malhotra et al., 2025; Shan et al., 2026; Zhao et al., 2026) are cited in Supplementary Materials.

Author Contributions: Conceptualization, R.A.A., E.B.B., and J.B.M.; methodology, E.B.B. and J.B.M.; validation, R.A.A., E.B.B. and V.A.P.; formal analysis, R.A.A., E.B.B., and J.B.M.; investigation, R.A.A., E.B.B., and J.B.M.; resources, V.A.P. and J.B.M.; data curation, V.A.P. and J.B.M.; writing—original draft preparation, R.A.A., E.B.B., and J.B.M.; writing—review and editing, R.A.A., E.B.B., V.A.P. and J.B.M.; visualization, V.A.P.; supervision, J.B.M.; funding acquisition, V.A.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Office of the Vice President for Research, Development, Innovation, and Extension of Caraga State University through the Regular Agency Fund.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

Acknowledgments: The authors gratefully acknowledge the Office of the Vice President for Research, Development, Innovation, and Extension and the School of Graduate Studies of Caraga State University for providing institutional support for this research.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
CAS	Computer Algebra System
CLEAR	Concise, Logical, Explicit, Adaptive, Reflective
CT	Critical Thinking
DSC	Differential Scanning Calorimetry
FTIR	Fourier-Transform Infrared Spectroscopy
GenAI	Generative Artificial Intelligence
HOTS	Higher-Order Thinking Skills
LLM	Large Language Model
MMAT	Mixed Methods Appraisal Tool
PID	Proportional–Integral–Derivative
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RAG	Retrieval-Augmented Generation
RAPA	Retrieval-Augmented Pedagogical Assistant
RQ	Research Question
SQL	Structured Query Language
STEM	Science, Technology, Engineering, and Mathematics
UX	User Experience

Appendix A. Detailed Evidence Matrix of Included Studies

Table A1. Characteristics of the included studies.

Study	STEM Context/Task	GenAI Tool	Design/Participants	Student–GenAI Interaction and CT-Related Evidence
Gong et al. (2025)	Programming; Python adventure-game development	Kimi Intelligent Assistant	One-group pre–post mixed learning analytics study; $n = 44$ undergraduates	Progressive prompting, questioning, objection, critique, and summarization; direct self-reported CT evidence within a computational-thinking measure plus coded interaction behaviors
Lin et al. (2026)	Grade 8 chemistry; balancing chemical equations	5E-structured GenAI coach	Two-class quasi-experiment; $n = 60$ eighth-grade students	5E instructional model (Engage, Explore, Explain, Elaborate, and Evaluate); 5E-guided GenAI interaction incorporating Socratic questioning, hints, and corrective feedback; log-sequence evidence of interaction with the intended 5E structure; no direct CT measure.
Matzakos and Moundridou (2025)	Engineering mathematics; matrices and civil-engineering problems	LLM + Maxima	Mixed-methods mathematics-lab intervention; population = 33 civil engineering students; pre-survey n = 23, post-survey n = 21	LLM–CAS comparison and cross-verification with iterative, verification-oriented problem-solving interaction; process-level comparison, error checking, evaluation, and verification
Supriadi and Suherman (2026)	Integral calculus; non-routine/HOTS problems	Text-based GenAI/LLM	Explanatory sequential mixed methods; $n = 85$ first-year Mathematics Education students	Contextual and iterative/debugging prompting; direct mathematical logical-reasoning measure plus interaction evidence
Dominguez et al. (2026)	Engineering; case analysis, design, multistep problem solving, and data analysis	ChatGPT-4o via API	16-week stratified randomized AI-access design; n = 128 fourth-year engineering students	Prompting behavior, refinement, and AI-driven problem solving, query efficiency; process-level problem-solving/metacognitive evidence
Gallart et al. (2026)	Mathematical modeling; Fermi problems	ChatGPT 3.5	Mixed qualitative–quantitative empirical study; $n = 133$ pre-service teachers	Cognitive-delegation profiles and task complexity; process-level decomposition, modeling, validation, and learner autonomy
Garg et al. (2025)	Engineering; Python-based data aggregation and merging analysis	GenAI (ChatGPT)	Three-group quasi-experimental instructional study; $n = 157$ first-year engineering students	Structured prompt training using CLEAR (Concise, Logical, Explicit, Adaptive, Reflective), few-shot, and chain prompting; performance assessed through Python data-analysis tasks and Bloom-aligned understanding, application, and analysis measures; no direct CT measure
Nathaniel et al. (2025)	Programming; C++ problem-solving tasks	ChatGPT/GitHub Copilot	One-group pretest–posttest explanatory sequential mixed methods intervention; n = 25 second-year computer science students	Prompting, response modification, scaffolded feedback interaction, iterative refinement, and coding; direct performance-based CT evidence within a tailored HOTS assessment

Table A1. *Cont.*

Study	STEM Context/Task	GenAI Tool	Design/Participants	Student–GenAI Interaction and CT-Related Evidence
Štuikys et al. (2025)	Design-based STEM/robotics; autonomous folkrace robot	ChatGPT-4o + Gemini 2.0 Flash	Design-based case evaluation; n = 48 Grade 9 students	Coding, design exploration, troubleshooting, physical testing, and refinement; process-level problem solving, evaluation, and testing
Reddy et al. (2024)	Biochemistry; scientific writing and evaluation of AI-generated information	ChatGPT and U-M GPT	Structured GenAI-assisted assignment; survey n = 33 respondents	Creating, evaluating, and iteratively refining GenAI essays, and reference-checking; process-level evaluation, verification, and correction
X. Zhang (2025)	Database Education; Database design, Structured query language (SQL), and problem solving	AI tools including Copilot	Mixed-method teaching-tip case; course n = 35; survey plus email interviews n = 3	Independent problem-solving solution followed by AI-assisted comparison; self-reported/process-level comparison, reflection, and evaluation illustrated by selected artifacts.
Tassoti (2024)	Chemistry; problem solving and prompt engineering	ChatGPT 3.5 via Discord bot	Empirical educational study; Work-log content analysis + pre/post perception survey; n = 23 preservice chemistry teachers	Initial copy-paste prompting shifted to more structured Five S prompting, recognition of inaccuracies, and prompt refinement; process-level evaluation and reflection
Chen et al. (2024)	Engineering/robotics; intelligent-car design, PID tuning, coding and testing	ChatGPT	Inquiry- and project-based empirical study; three student groups; 3–4 students per group	Platform-specific prompting, code validation, repeated feedback, PID tuning, and real-world testing, AI recommendations, physical testing, rejection, and refinement; process-level evaluation, judgment, correction, and verification
Buendía-García and Piris-Ruano (2025)	UX-oriented Web development; PHP, HTML/CSS, SQL/database programming	GitHub Copilot; preliminary ChatGPT-4	Case study; n = 21	Incremental prompting, code explanation, troubleshooting, and debugging documented through interaction logs; process-level evidence of task organization, interpretation, and refinement; no direct CT measure
Cruz-Angeles et al. (2026)	Life sciences and nanotechnology; research/laboratory tasks	ChatGPT 3.5 + customized RAG GPT assistants	Phase I exploratory observational study, n = 94 undergraduate students; Phase II non-randomized quasi-experimental, n = 70 fourth-semester undergraduate students	Guided prompting, external verification, and Socratic/RAG support; self-reported CT processes in Phase I; academic performance assessed in Phase II
Picardal (2025)	Science education research; conceptualization, methodology, data interpretation, and academic writing	ChatGPT, Perplexity AI, Gemini, other tools	Qualitative descriptive; n = 65 interaction artifacts/reflection logs and n = 15 interviews	Contextual, iterative, file-based, and role-based prompting; process-level questioning, interpretation, refinement, validation, and reflection, cognitive/metacognitive scaffolding

Table A1. *Cont.*

Study	STEM Context/Task	GenAI Tool	Design/Participants	Student–GenAI Interaction and CT-Related Evidence
Aruleba et al. (2025)	Undergraduate computer science students (n = 21); Java programming	ChatGPT	Four-week longitudinal qualitative study; programming tasks, reflective journals, and interviews	Qualitative evidence of CT-related processes derived from student reflections, interviews, and programming artifacts; no direct standardized or self-reported CT ability assessment.

Note. CT = critical thinking; HOTS = higher-order thinking skills; LLM = large language model; RAG = retrieval-augmented generation. Direct CT evidence refers to explicit measurement of CT; performance-based CT evidence refers to CT assessed within a broader task-based framework; related reasoning evidence refers to direct measurement of a closely related construct; and process-level evidence refers to reported, observed, or self-reported CT-related cognitive or interaction processes rather than direct measurement of CT ability.

References

Abualrob, M., & Ghannam, D. (2026). AI-supported authentic assessment in science education: Overcoming logistical barriers and enhancing outcomes. *International Journal of Learning, Teaching and Educational Research*, *25*(5), 227–251. [CrossRef]

Al Darayseh, A., & Mersin, N. (2025). Integrating generative AI into STEM education: Insights from science and mathematics teachers. *International Electronic Journal of Mathematics Education*, *20*(3), em0832. [CrossRef]

Aruleba, K., Sanusi, I. T., Obaido, G., Ogbuokiri, B., & Mienye, I. D. (2025). Beyond the prompt: Student strategies, ethical reflections, and learning with ChatGPT in computer science. *Technology, Knowledge and Learning*. [CrossRef]

Bai, S., Yeung, S. S., & Lo, C. K. (2026). Enhancing the effect of AI-assisted learning: The use of scaffolding strategies to develop students' prompt engineering skills. *Interactive Learning Environments*, 1–22. [CrossRef]

Belkina, M., Daniel, S., Nikolic, S., Haque, R., Lyden, S., Neal, P., Grundy, S., & Hassan, G. M. (2025). Implementing generative AI (GenAI) in higher education: A systematic review of case studies. *Computers and Education: Artificial Intelligence*, *8*, 100407. [CrossRef]

Bell, C. E., & Birceanu, O. (2025). Educators' experience and guide to scaffolding generative AI applications throughout a physiology and pharmacology undergraduate laboratory course. *Advances in Physiology Education*, *49*(3), 721–728. [CrossRef]

Bhuiyan, M. A., Rahman, M. K., Basile, V., Ping, H., & Mainul Bari, A. B. M. (2025). Adoption of ChatGPT for students' learning effectiveness. *The International Journal of Management Education*, *23*(3), 101255. [CrossRef]

Buendía-García, F., & Piris-Ruano, J. (2025). Using generative AI to support UX design students in web development courses. *Applied Sciences*, *15*(13), 7389. [CrossRef]

Chen, Q., Chen, H. C., & Lin, Y. L. (2024). ChatGPT-powered inquiry-based learning model of training for intelligent car racing competition. *Sensors and Materials*, *36*(3), 1147–1161. [CrossRef]

Cong-Lem, N., & Thuy-Dung, N. T. (2026). Towards a framework for understanding and assessing critical thinking in generative AI-enhanced learning environments: A scoping review. *European Journal of Psychology of Education*, *41*(2), 41. [CrossRef]

Cruz-Angeles, J., Elizondo-García, M. E., Zavala, G., & Membrillo-Hernández, J. (2026). Integration of intelligent assistants and adaptive learning in engineering pedagogy: A case study in nanotechnology and life sciences. *International Journal of Engineering Pedagogy*, *16*(4), 90–103. [CrossRef]

Ding, L., O'Berry, R., Tallent, H., Chhetri GC, S. G., & Gapud, A. (2026). Learning with large language models: Beyond prompt engineering. *Education and Information Technologies*, *31*(9), 2877–2901. [CrossRef]

Dominguez, L. G. I., Robles-Gómez, A., & Pastor-Vargas, R. (2026). An empirical study of ChatGPT use in engineering education: Prompting and performance. *The Internet and Higher Education*, *71*, 101105. [CrossRef]

Fang, L., & Zhou, X. (2025). From tool to co-learner: Exploring student engagement with GenAI through the lens of social constructivism. *Evaluation Review*, 1–33. [CrossRef]

Farrokhnia, M., Latifi, S., Papadopoulos, P. M., Hogenkamp, L., Gijlers, H., Khosravi, H., & Noroozi, O. (2026). Generative AI offers more, but students revise less: Comparing the effects of teacher and AI feedback on student essay revisions. *International Journal of Educational Technology in Higher Education*, *23*(1), 6. [CrossRef]

Feldman-Maggor, Y., Blonder, R., & Alexandron, G. (2025). Perspectives of generative AI in chemistry education within the TPACK framework. *Journal of Science Education and Technology*, *34*(1), 1–12. [CrossRef]

Gallart, C., Ferrando, I., Segura, C., & Albarracín, L. (2026). Modelling with AI: How complexity and experience shape ChatGPT use by pre-service teachers. *Eurasia Journal of Mathematics, Science and Technology Education*, *22*(4), em2816. [CrossRef]

Garg, A., Nisumba Soodhani, K., & Rajendran, R. (2025). Enhancing data analysis and programming skills through structured prompt training: The impact of generative AI in engineering education. *Computers and Education: Artificial Intelligence*, *8*, 100380. [CrossRef]

Gong, X., Xu, W., & Qiao, A. (2025). Exploring undergraduates' computational thinking and human-computer interaction patterns in generative progressive prompt-assisted programming learning. *International Journal of Educational Technology in Higher Education*, *22*(1), 51. [CrossRef]

Hashemi, S. S. (2026). Students' multimodal prompting practices as epistemic work in AI literacy development. *Computers and Education: Artificial Intelligence*, *11*, 100635. [CrossRef]

Hong, Q. N., 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., & Pluye, P. (2018). The Mixed Methods Appraisal Tool (MMAT) version 2018 for information professionals and researchers. *Education for Information*, *34*(4), 285–291. [CrossRef]

Hsu, H. P. (2025). From programming to prompting: Developing computational thinking through large language model-based generative artificial intelligence. *TechTrends*, *69*(3), 485–506. [CrossRef]

Isaeva, R., Caner, H. N., Caner, M., Giray, L., & Karadag, E. (2026). Students' engagement with generative AI in academic learning: A self-determination theory and epistemic network analysis study. *Computers and Education: Artificial Intelligence*, *10*, 100606. [CrossRef]

Kim, K., Kwon, K., Kwon, H., & Brush, T. (2026). How do students' AI interactions, dispositions, and prompt engineering skills shape human–AI collaboration in middle school STEM education? *Educational Technology Research and Development*, 1–39. [CrossRef]

Krupp, L., Steinert, S., Kiefer-Emmanouilidis, M., Avila, K. E., Lukowicz, P., Kuhn, J., Küchemann, S., & Karolus, J. (2024). Unreflected acceptance: Investigating the negative consequences of ChatGPT-assisted problem solving in physics education. *Frontiers in Artificial Intelligence and Applications*, *386*, 199–212. [CrossRef]

Kwong, A. C., Magnano, C., DeOliveira, C., Goglia, C., Loparo, J. J., & Peters, J. J. (2026). Designing life science assessments in the era of generative artificial intelligence. *PLoS ONE*, *21*(4), e0346127. [CrossRef]

Lee, D., & Palmer, E. (2025). Prompt engineering in higher education: A systematic review to help inform curricula. *International Journal of Educational Technology in Higher Education*, *22*(1), 7. [CrossRef]

Lin, T. C., Shih, Y. T., & Li, C. H. (2026). Designing and evaluating a 5E-structured GenAI coach for guided inquiry: A pedagogy-to-prompt engineering framework. *Education Sciences*, *16*(3), 384. [CrossRef]

Lu, H. C. C. (2026, April 13–17). *How do undergraduate students make decisions when adopting generative artificial intelligence for academic purposes?* [Conference contribution]. Conference on Human Factors in Computing Systems, Barcelona, Spain. [CrossRef]

Malhotra, A., Buller, M., Modi, K., Pajazetovic, K., & Wijesinghe, D. S. (2025). Blueprint for constructing an AI-based patient simulation to enhance the integration of foundational and clinical sciences in didactic immunology in a US Doctor of pharmacy program: A step-by-step prompt engineering and coding toolkit. *Pharmacy*, *13*(2), 36. [CrossRef]

Matzakos, N., & Moundridou, M. (2025). Exploring large language models integration in higher education: A case study in a mathematics laboratory for civil engineering students. *Computer Applications in Engineering Education*, *33*(3), e70049. [CrossRef]

Nathaniel, J., Oyelere, S. S., Suhonen, J., & Tedre, M. (2025). Investigating the impact of generative AI integration on the sustenance of higher-order thinking skills and understanding of programming logic. *Computers and Education: Artificial Intelligence*, *9*, 100460. [CrossRef]

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. [CrossRef]

Pangestu, F. N. N., Hamidah, I., Widaningsih, L., Abdullah, A. G., & Saputra, N. A. (2026). A scoping literature review of prompt engineering for bridging students' AI literacy in higher education. *Computers and Education: Artificial Intelligence*, *10*, 100581. [CrossRef]

Paredes-Cayllahua, E. L. (2026). Pedagogical prompt engineering patterns: A Vygotskian approach to adaptive scaffolding in computing education with LLMs. *Publicaciones de la Facultad de Educación y Humanidades del Campus de Melilla*, *56*(1), 39–60. [CrossRef]

Picardal, M. T. (2025). Utilization of AI-driven smart prompts in academic research in higher education institutions. *International Journal of Learning, Teaching and Educational Research*, *24*(9), 384–404. [CrossRef]

Pramanik, S., & Husin, M. B. H. (2025). The retrieval-augmented pedagogical assistant (RAPA): A methodology for enhancing critical thinking and equity in AI-augmented education. *International Journal of Advanced Computer Science and Applications*, *16*(12), 274–284. [CrossRef]

Qu, X., Sherwood, J., Liu, P., & Aleisa, N. (2025, April 26–May 1). *Generative AI tools in higher education: A meta-analysis of cognitive impact* [Conference contribution]. Conference on Human Factors in Computing Systems, Yokohama, Japan. [CrossRef]

Raitskaya, L., & Tikhonova, E. (2025). Enhancing critical thinking skills in ChatGPT-human interaction: A scoping review. *Journal of Language and Education*, *11*(2), 5–19. [CrossRef]

Reddy, M. R., Walter, N. G., & Sevryugina, Y. V. (2024). Implementation and evaluation of a ChatGPT-assisted special topics writing assignment in biochemistry. *Journal of Chemical Education*, *101*(7), 2740–2748. [CrossRef]

Reicho, M., Otrel-Cass, K., Ebner, M., Brünner, B., Freinhofer, D., & Irfan, J. (2026). Generative AI literacy across education and business: Competencies, obstacles, and benefits—A systematic literature review. *International Journal of Educational Technology in Higher Education*, *23*(1), 23. [CrossRef]

Schapiro, H. M., Tripepi, M., & Blackley, J. (2026). Reclaiming the writing process using AI: A scaffolded writing assignment that promotes critical thinking, reflection, and accountability. *Journal of Microbiology & Biology Education*, *27*, e00337-25. [CrossRef]

Shan, X., Ye, X., Tu, Y., He, Z., Liu, Z., Zheng, S., Zeng, T., & Zhu, H. (2026). Can large language model-based chatbots support sustainable AI-enhanced education? Evidence from young learners' conceptions. *Sustainability*, *18*(19), 10031. [CrossRef]

Sun, Q., & Yin, S. (2026). Exploring critical interaction with GAI in academic writing: A mixed-methods study. *TechTrends*, *70*(4), 834–845. [CrossRef]

Sun, X., Liu, C., Li, P., & Cai, Z. (2026). Effects of generative AI tools on students' critical thinking: A three-level meta-analysis. *Educational Psychology Review*, *38*(1), 84. [CrossRef]

Supriadi, N., & Suherman, S. (2026). Integration of generative AI in creative mathematics learning: A correlational analysis between the quality of prompts and the logical reasoning of student teachers in the education degree programme. *Revista de Educación a Distancia*, *26*(84), 1–16. [CrossRef]

Štuikys, V., Burbaitė, R., Binkis, M., & Ziberkas, G. (2025). Developing problem-solving skills to support sustainability in STEM education using generative AI tools. *Sustainability*, *17*(15), 6935. [CrossRef]

Tassoti, S. (2024). Assessment of students use of generative artificial intelligence: Prompting strategies and prompt engineering in chemistry education. *Journal of Chemical Education*, *101*(6), 2475–2482. [CrossRef]

Válek, J., Straková, N., & Prokeš, L. (2025). The use of generative AI in STEM education to develop scientific literacy and critical thinking. *Journal of Physics: Conference Series*, *3155*(1), 012006. [CrossRef]

Walter, Y. (2024). Embracing the future of artificial intelligence in the classroom: The relevance of AI literacy, prompt engineering, and critical thinking in modern education. *International Journal of Educational Technology in Higher Education*, *21*(1), 15. [CrossRef]

Yan, W., Nakajima, T., & Sawada, R. (2025). Beyond tool use: Tracking the evolution of generative AI literacy among university students through a process-oriented investigation. *Computers and Education: Artificial Intelligence*, *9*, 100465. [CrossRef]

Zhang, W., & Liu, X. (2025). Artificial intelligence-generated content empowers college students' critical thinking skills: What, how, and why. *Education Sciences*, *15*(8), 977. [CrossRef]

Zhang, X. (2025). Incorporating AI tools into database classes. *Journal of Information Systems Education*, *36*(1), 37–52. [CrossRef]

Zhao, H. V., Ren, Y., Shang, C., Yin, H., Xu, Z., Jiang, R., Jiang, B., Zhuo, Q., Zhang, H., Pei, X., Zhang, C., & Geng, H. (2026). AI tutors and the transformation of education: Opportunities, challenges, and future directions. *Cybernetics and Intelligence*, 1–12. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.