Supplementary Table S1. Methodological Quality Assessment of Included Studies Using MMAT 2018

Study ID	Author & Year	S1	S2	Decision	1.1	1.2	1.3	1.4	1.5	Remarks	Category
55	Aruleba et al. (2025)	Y	Y	Proceed	Y	Y	Y	Y	Y	Short-term longitudinal qualitative study involving 21 computer science students, using programming artifacts, ChatGPT interaction records, reflective journals, and semi-structured interviews. Independent coding by two researchers, iterative analysis, peer debriefing, and reflexive documentation strengthen methodological credibility. However, the single-institution sample, four-week duration, and structured task sequence limit transferability to other learning contexts.	Category 1 - Qualitative
162	Picardal (2025)	Y	Y	Proceed	Y	Y	Y	Y	Y	Qualitative descriptive study involving 65 preservice science teachers, using AI interaction logs, reflections, research outputs, and interviews with 15 group representatives. Multiple data sources, thematic analysis, triangulation, and cross-checking of coding strengthen credibility. However, voluntary participation, a single-course setting, and reliance on selected interview representatives limit transferability and the breadth of individual perspectives.	Category 1 - Qualitative
Study ID	**Author & Year**	**S1**	**S2**	**Decision**	**2.1**	**2.2**	**2.3**	**2.4**	**2.5**	**Remarks**	**Category**
53	Dominguez et al. (2026)	Y	Y	Proceed	Y	CT	CT	Y	CT	Semester-long randomized crossover study involving 128 engineering students, with stratified assignment of AI-assisted and non-AI sessions, standardized tasks, interaction-log analysis, and blinded assessment. Randomization procedures and outcome assessment were adequately described; however, incomplete reporting of condition-specific baseline comparability, missing outcome data, and intervention adherence limits confidence in some findings. The study primarily assessed prompting behaviors and academic performance rather than direct changes in critical-thinking ability.	Category 2 - Quantitative Randomized Controlled Trial
Study ID	**Author & Year**	**S1**	**S2**	**Decision**	**3.1**	**3.2**	**3.3**	**3.4**	**3.5**	**Remarks**	**Category**
1	Gong et al. (2025)	Y	Y	Proceed	Y	CT	Y	Y	CT	Single-cohort pre–post intervention involving 44 undergraduate students, supported by learning analytics of student–GenAI interactions. The absence of a control group, convenience sampling, and short intervention duration limit causal inference and generalizability.	Category 3 - Quantitative non-randomized
25	Lin et al. (2026)	Y	Y	Proceed	Y	CT	Y	Y	Y	Quasi-experimental study involving two intact classes (n = 60), with baseline assessment, standardized instruction, and student–AI dialogue analysis. Non-random allocation, potential residual confounding, and the brief intervention limit causal interpretation and assessment of sustained learning effects.	Category 3 - Quantitative non-randomized
85	Garg et al.	Y	Y	Proceed	Y	CT	Y	CT	Y	Three-group educational intervention involving 157 engineering students, with pre–post assessment and objective programming tasks. Unclear group allocation, potential confounding, and low knowledge-test reliability (Cronbach's $\alpha = 0.42$) reduce confidence in the estimated intervention effects.	Category 3 - Quantitative non-randomized
160	Cruz-Angeles et al. (2026)	Y	Y	Proceed	CT	CT	CT	CT	Y	Two-phase study combining an exploratory survey (n = 94) and a non-randomized comparison (n = 70). Non-probabilistic allocation, differences in AI-tool access, and potential baseline confounding limit causal inference and generalizability.	Category 3 - Quantitative non-randomized
Study ID	**Author & Year**	**S1**	**S2**	**Decision**	**4.1**	**4.2**	**4.3**	**4.4**	**4.5**	**Remarks**	**Category**
152	Buendía-García and Piris-Ruano (2025)	Y	Y	Proceed	Y	CT	Y	CT	Y	Single-course quantitative descriptive case study involving 21 UX design students, using GenAI interaction logs, benchmark comparisons, task-performance measures, and student questionnaires. Multiple measures support the examination of prompting practices and task outcomes. However, limited sample representativeness, incomplete questionnaire participation, small analytical subsamples, and exploratory statistical analyses constrain generalizability. Critical-thinking improvement was not directly assessed.	Category 4 - Quantitative descriptive
Study ID	**Author & Year**	**S1**	**S2**	**Decision**	**5.1**	**5.2**	**5.3**	**5.4**	**5.5**	**Remarks**	**Category**
96	Reddy et al. (2024)	Y	Y	Proceed	Y	CT	Y	CT	CT	Small-scale educational study involving 23 students, combining quantitative survey responses, qualitative feedback, and verification of AI-generated references. The multiple data sources provide complementary evidence of students' prompting and content-evaluation practices. However, limited reporting of mixed-methods integration, a small single-course sample, and potential recall bias constrain the interpretation and transferability of findings. Critical-thinking improvement was not directly measured.	Category 5 - Mixed methods
132	Tassoti (2024)	Y	Y	Proceed	Y	Y	Y	CT	CT	Mixed-methods investigation involving 23 preservice chemistry teachers and 67 complete interaction logs, combining qualitative prompt coding with quantitative comparisons of response ratings. A structured coding framework and interrater reliability assessment ($\kappa = 0.82$ after refinement) strengthen analytical credibility. However, the small single-course sample, reliance on subjective response evaluations, and limited reporting of mixed-methods integration constrain generalizability and interpretation of critical-thinking-related outcomes.	Category 5 - Mixed methods

80	Gallart et al. (2026)	Y	Y	Proceed	Y	Y	Y	CT	CT	Quasi-experimental study involving 133 pre-service teachers, integrating interaction profiles and quantitative comparisons across Fermi-problem tasks. Convenience sampling, possible confounding, and insufficiently documented qualitative–quantitative integration limit methodological certainty.	Category 5 - Mixed methods
37	Matzakos and Moundridou (2025)	Y	Y	Proceed	Y	CT	CT	CT	CT	Classroom-based investigation of six structured mathematics laboratory activities integrating large language models and computer algebra systems. Quantitative responses and qualitative observations provide complementary evidence of students' verification and problem-solving practices. However, convenience sampling, a short intervention, and limited reporting of formal mixed-methods integration constrain methodological confidence and generalizability.	Category 5 - Mixed methods
47	Supriadi and Suherman (2026)	Y	Y	Proceed	Y	Y	Y	CT	CT	Explanatory sequential mixed-methods study involving 85 student teachers, combining Spearman correlation analysis of prompt quality and mathematical reasoning accuracy with thematic analysis of AI interactions. The qualitative component explicitly explains quantitative patterns, strengthening integration. However, purposive sampling, possible confounding, and limited component-level methodological information restrict causal interpretation and transferability.	Category 5 - Mixed methods
87	Nathaniel et al. (2025)	Y	Y	Proceed	Y	Y	Y	CT	CT	Explanatory sequential mixed-methods study involving 25 programming students, combining repeated performance assessments with thematic analysis of student reflections, AI interaction logs, and code submissions. Expert validation, reliability checks, and multiple qualitative sources strengthen the appraisal. However, convenience sampling, the absence of a comparison group, and incomplete reporting of component-level quality limit causal inference and generalizability.	Category 5 - Mixed methods
128	X. Zhang (2025)	Y	Y	Proceed	Y	CT	CT	CT	CT	Educational case study combining descriptive student surveys, three email interviews, and examples of students evaluating AI-generated database solutions. The data provide useful evidence of perceived learning benefits and critical evaluation practices. However, limited reporting of survey validity, qualitative analytical procedures, and formal integration between data components reduces methodological confidence. Critical-thinking improvement was not independently measured.	Category 5 - Mixed methods
Study ID	**Author & Year**	**S1**	**S2**	**Decision**	**-**	**-**	**-**	**-**	**-**	**Remarks**	**Category**
89	Štuikys et al. (2025)	Y	Y	Proceed	-	-	-	-	-	Methodological modelling study illustrated through a robotics design case involving 48 Grade 9 students. Teacher observations and five-point Likert impact ratings describe GenAI-supported design activities; however, the rating procedures, measurement validity, analytical methods, and comparative evidence are insufficiently documented. The study does not clearly fit a standard MMAT design category and claims of improved critical thinking cannot be independently verified.	Studies outside established MMAT categories
151	Chen et al. (2024)	Y	Y	Proceed	-	-	-	-	-	Descriptive educational implementation involving three student groups participating in ChatGPT-supported intelligent vehicle competition training. Examples of stepwise prompting, code generation, troubleshooting, and iterative optimization provide evidence of student–GenAI interaction. However, participant sampling, systematic data collection and analysis, and outcome measurement procedures are insufficiently documented. The study does not clearly fit a standard MMAT design category, and reported critical-thinking benefits remain unverified.	Studies outside established MMAT categories

Note. MMAT criterion-level assessments are organized by study design. Y = Yes; N = No; CT = Can't Tell; - = not applicable. Ratings marked CT indicate insufficient evidence for a confident judgment, not necessarily poor methodological quality. MMAT 2018 uses two initial screening questions (S1 and S2), followed by five design-specific criteria.

Supplementary Table S2. Studies Excluded After Full-Text Assessment

Authors (year)	**Article title**	**Specific reason for exclusion**
Abualrob & Ghannam (2026)	AI-Supported Authentic Assessment in Science Education: Overcoming Logistical Barriers and Enhancing Outcomes	Empirical study, but GenAI was primarily used by the teacher to support assessment design; no eligible empirical student–GenAI interaction data.
Hsu (2025)	From Programming to Prompting: Developing Computational Thinking through Large Language Model-Based Generative Artificial Intelligence	Proposed a constructionist prompting framework and instructional workshop without eligible empirical student–GenAI interaction findings.
Kwong et al. (2026)	Designing Life Science Assessments in the Era of Generative Artificial Intelligence	Empirically compared ChatGPT performance with student performance, but did not examine students interacting with GenAI during learning tasks.
Pramanik & Husin (2025)	The Retrieval-Augmented Pedagogical Assistant (RAPA): A Methodology for Enhancing Critical Thinking and Equity in AI-Augmented Education	Focused on a proposed pedagogical architecture and technical validation rather than eligible empirical student–GenAI interaction evidence.
Feldman-Maggor et al. (2025)	Perspectives of Generative AI in Chemistry Education Within the TPACK Framework	Position paper discussing instructional and pedagogical considerations without original eligible empirical evidence.
Paredes-Cayllahua (2026)	Pedagogical Prompt Engineering Patterns: A Vygotskian Approach to Adaptive Scaffolding in Computing Education with LLMs	Design science study involving 54 model executions, but no empirical observations of student–GenAI interaction or student learning outcomes.
Bell & Birceanu (2025)	Educators' Experience and Guide to Scaffolding Generative AI Applications throughout a Physiology and Pharmacology Undergraduate Laboratory Course	Instructional guidance and educator experience without eligible empirical student–GenAI interaction findings.

Shan et al. (2026)	Can Large Language Model-Based Chatbots Support Sustainable AI-Enhanced Education? Evidence from Young Learners' Conceptions	Study involved 72 primary-school learners, outside the review's eligible secondary- and tertiary-level STEM education contexts. Although critical-thinking tendencies were surveyed, the study did not provide eligible student–GenAI critical-thinking process evidence for the target population.
Malhotra et al. (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	Technical and pedagogical validation relied primarily on simulated AI–patient-case interactions and rubric scoring of the simulation, rather than analyzable authentic student–GenAI interactions or student critical-thinking outcomes.
Zhao et al. (2026)	AI tutors and the transformation of education: Opportunities, challenges, and future directions	Platform implementation and course case studies reported deployment, student questions, engagement, and satisfaction, but did not systematically analyze or measure student critical-thinking-related processes or outcomes attributable to student–GenAI interaction.

Supplementary Table S3. Study-Level Evidence and Thematic Mapping of CT-Related Cognitive Processes

No.	**Keyword / cognitive process/ Initial Code**	**Subtheme**	**Primary CT Higher-order theme**	**Secondary CT higher-order theme**	**1**	**25**	**37**	**47**	**53**	**55**	**80**	**85**	**87**	**89**	**96**	**128**	**132**	**151**	**152**	**160**	**162**	**Student–GenAI Interaction**	**Student–GenAI Interaction Subpattern**	**Higher-order Interaction Pattern**
1	Analytical thinking / analytical skills	Analytical interpretation	T2. Interpretation and analytical decomposition	T3	D	D	O	O	D	O	O	M	M	D	O	O	O	O	D	M	D	Interpreting AI explanations; examining generated code, calculations, and task information; analyzing solution steps and relationships	Analytical interpretation of AI outputs	Evaluation & verification
2	Challenging or objecting to AI responses	Challenging AI-generated information	T3. Comparison, evaluation, and verification	T5	M	D	O	O	-	O	-	-	O	-	O	O	O	O	-	D	O	Objecting to AI-generated claims; questioning responses; debating or contesting generated explanations	Critical evaluation of AI outputs	Modification, challenge, or rejection of AI-generated outputs
3	Comparison of alternatives / alternative solutions	Comparative analysis	T3. Comparison, evaluation, and verification	T2	M	-	O	D	D	O	D	D	M	O	D	O	-	O	-	-	D	Comparing AI-generated solutions with independent answers, alternative mathematical procedures, or computational tools	Comparative analysis of solutions	Evaluation & verification
4	Computational thinking	Computational problem decomposition	T2. Interpretation and analytical decomposition	T4	M	-	-	-	D	D	-	D	D	M	-	D	-	D	D	-	-	Developing programming solutions; using progressive prompts; interpreting and	Problem decomposition and task structuring	Problem decomposition

																						modifying code; examining computational problem-solving processes		
5	Conceptual clarification	Conceptual interpretation	T2. Interpretation and analytical decomposition	T1	M	O	D	O	D	O	M	D	O	O	D	O	O	O	O	D	M	Requesting explanations and clarification of concepts; using guided questions to resolve uncertainty	Guided inquiry and exploratory dialogue	Contextual & iterative prompting
6	Conceptual understanding and knowledge construction	Knowledge construction	T2. Interpretation and analytical decomposition	T5	O	M	D	O	O	O	D	M	M	M	D	M	D	O	D	M	D	Using AI explanations to explore disciplinary concepts; connecting explanations with task knowledge; elaborating understanding through follow-up interactions	Analytical interpretation of AI outputs	Contextual & iterative prompting
7	Constraint specification	Task and constraint framing	T1. Strategic questioning and task framing	T6	D	-	-	M	D	O	M	O	O	O	O	D	M	O	O	-	O	Providing task requirements and contextual instructions; specifying conditions and refining prompts to obtain suitable responses	Contextual prompting and task specification	Contextual & iterative prompting
8	Creativity and creative thinking	Creative interpretation and solution generation	T2. Interpretation and analytical decomposition	T3	M	-	D	D	D	D	-	D	M	M	-	M	-	O	D	-	D	Exploring alternative designs or solutions; using AI suggestions during programming and engineering design; refining generated ideas	Iterative solution refinement	Modification, challenge, or rejection of AI-generated outputs
9	Critical evaluation of AI-generated outputs	Critical evaluation of AI outputs	T3. Comparison, evaluation, and verification	T5	M	D	O	O	D	O	D	D	O	D	M	M	M	O	D	O	D	Assessing the accuracy, relevance, and suitability of AI responses; questioning generated claims; comparing outputs	Critical evaluation of AI outputs	Evaluation & verification

																						with independent evidence		
10	Critical thinking	Critical evaluation and judgment	T3. Comparison, evaluation, and verification	T4	M	D	D	D	D	O	D	D	M	M	O	M	D	D	D	D	D	Questioning, critiquing, comparing, and examining AI responses; modifying outputs rather than accepting them without scrutiny	Critical evaluation of AI outputs	Evaluation & verification
11	Critiquing AI-generated responses	Critical appraisal of AI responses	T3. Comparison, evaluation, and verification	T4	M	D	O	O	-	O	-	-	O	-	O	O	O	O	-	D	O	Critiquing generated explanations; identifying weaknesses; questioning responses and requesting improvement	Critical evaluation of AI outputs	Modification, challenge, or rejection of AI-generated outputs
12	Data analysis and interpretation	Data interpretation	T2. Interpretation and analytical decomposition	T3	-	-	O	-	O	-	O	M	-	M	-	O	-	O	-	M	M	Using GenAI in data-analysis tasks; interpreting generated analytical information; examining task results	Analytical interpretation of AI outputs	Evaluation & verification
13	Debugging and code optimization	Debugging and corrective optimization	T4. Error detection and corrective modification	T3	O	-	-	-	-	O	-	-	M	O	-	D	-	O	O	-	-	Examining generated code; identifying programming problems; requesting corrections; testing and refining code	Error identification and diagnosis	Modification, challenge, or rejection of AI-generated outputs
14	Decision-making and independent judgment	Independent decision-making	T5. Justification, judgment, and independent decision-making	T3	O	O	O	O	D	O	M	D	D	O	O	O	O	O	D	M	D	Evaluating AI recommendations; accepting, modifying, or rejecting suggested solutions; retaining responsibility for final decisions	Independent judgment and decision-making	Modification, challenge, or rejection of AI-generated outputs

15	Empirical testing and experimentation	Empirical testing and validation	T4. Error detection and corrective modification	T3	O	D	O	-	-	D	D	O	M	O	D	D	-	O	O	D	-	Testing AI-generated engineering or robotics recommendations; examining real-world results; revising designs based on observed performance	Testing and validation of outputs	Evaluation & verification
16	Error detection and correction	Error identification and correction	T4. Error detection and corrective modification	T3	O	O	O	M	-	O	-	-	M	O	O	D	O	O	O	O	O	Identifying incorrect reasoning or AI-generated errors; using debugging prompts; correcting generated solutions	Error identification and diagnosis	Modification, challenge, or rejection of AI-generated outputs
17	Ethical reasoning and ethical reflection	Ethical judgment	T5. Justification, judgment, and independent decision-making	T3	-	-	-	-	-	O	-	-	D	M	D	D	D	-	D	D	D	Evaluating the appropriateness and responsible use of AI-generated information	Independent judgment and decision-making	Evaluation & verification
18	Evidence-based reasoning	Evidence-based evaluation	T3. Comparison, evaluation, and verification	T5	O	D	O	M	D	O	O	-	D	M	O	D	O	O	-	D	D	Comparing AI-generated claims with disciplinary knowledge, external sources, calculations, or other supporting evidence	Evidence and source verification	Evaluation & verification
19	Explanatory reasoning	Reasoning and justification	T5. Justification, judgment, and independent decision-making	T2	O	D	D	M	O	D	O	D	O	D	-	O	O	D	D	D	D	Requesting explanations of solution procedures; examining mathematical or programming reasoning; interpreting why a proposed solution works	Guided inquiry and exploratory dialogue	Contextual & iterative prompting
20	Feedback seeking and utilization	Feedback-driven modification	T4. Error detection and corrective modification	T6	M	O	O	O	O	O	O	D	O	O	O	D	O	O	O	D	M	Requesting corrective feedback; responding to AI suggestions; modifying solutions or code based on feedback	Corrective prompting and modification	Modification, challenge, or rejection of AI-generated outputs

21	Higher-order thinking skills (HOTS)	Higher-order cognitive processing	T2. Interpretation and analytical decomposition	T3	D	D	D	D	D	D	D	M	M	D	D	D	-	D	-	D	D	Engaging in AI-supported programming activities involving analysis, evaluation, modification, and solution development	Analytical interpretation of AI outputs	Evaluation & verification
22	Hypothesis generation and testing	Hypothesis formulation	T1. Strategic questioning and task framing	T3	-	D	M	-	-	-	-	-	-	M	-	-	-	D	-	-	-	Exploring possible explanations through guided inquiry; examining proposed solutions and testing their suitability	Guided inquiry and exploratory dialogue	Contextual & iterative prompting
23	Information reliability and accuracy assessment	Information credibility assessment	T3. Comparison, evaluation, and verification	T5	D	D	O	O	D	O	D	D	O	D	M	M	M	O	D	M	D	Checking generated information against independent sources; identifying inaccuracies; examining the credibility of generated claims	Evidence and source verification	Evaluation & verification
24	Inquiry and investigation	Inquiry-based exploration	T1. Strategic questioning and task framing	T2	M	O	O	O	D	O	O	D	D	M	O	O	O	O	O	O	O	Engaging in guided questioning; requesting explanations and hints; exploring problems through successive AI interactions	Guided inquiry and exploratory dialogue	Contextual & iterative prompting
25	Iterative refinement and improvement	Iterative modification	T4. Error detection and corrective modification	T1	O	O	O	M	M	O	O	D	O	O	O	D	M	O	O	-	M	Revising prompts; examining successive AI responses; refining generated code, solutions, or designs	Iterative solution refinement	Modification, challenge, or rejection of AI-generated outputs
26	Logical reasoning	Logical analysis	T2. Interpretation and analytical decomposition	T3	D	D	O	M	O	O	O	D	M	D	-	O	-	O	D	D	D	Examining mathematical solution steps; identifying logical flaws; using corrective prompts to improve reasoning	Analytical interpretation of AI outputs	Evaluation & verification

27	Mathematical modelling and analytical interpretation	Mathematical modelling and interpretation	T2. Interpretation and analytical decomposition	T3	-	-	O	O	D	-	M	-	-	D	-	-	-	D	-	-	-	Using GenAI to support Fermi modelling; developing mathematical solutions; interpreting generated modelling procedures	Analytical interpretation of AI outputs	Evaluation & verification
28	Mathematical reasoning	Mathematical analysis	T2. Interpretation and analytical decomposition	T3	-	D	M	M	D	-	O	-	-	D	-	-	-	D	-	D	-	Examining mathematical explanations; comparing alternative solution methods; checking mathematical reasoning	Analytical interpretation of AI outputs	Evaluation & verification
29	Metacognition and metacognitive awareness	Metacognitive awareness	T6. Reflection, metacognition, and self-regulation	T5	D	D	D	D	D	O	D	D	D	D	D	D	O	D	D	D	D	Reflecting on understanding and AI dependence; recognizing inadequate responses; reconsidering interaction strategies	Reflective and self-regulated interaction	Contextual & iterative prompting
30	Metacognitive planning and monitoring	Metacognitive planning and monitoring	T6. Reflection, metacognition, and self-regulation	T1	D	D	-	D	D	O	O	D	O	D	D	D	O	D	D	-	D	Planning prompts; monitoring generated responses; adjusting prompting and problem-solving strategies	Reflective and self-regulated interaction	Contextual & iterative prompting
31	Model construction and refinement	Model development and modification	T4. Error detection and corrective modification	T2	O	-	-	-	-	-	O	-	-	O	-	O	-	O	-	-	-	Developing AI-assisted mathematical or engineering models; modifying models or designs after evaluation	Iterative solution refinement	Modification, challenge, or rejection of AI-generated outputs
32	Problem decomposition	Problem decomposition and structuring	T2. Interpretation and analytical decomposition	T1	M	D	-	O	M	O	O	D	M	M	O	O	O	O	O	D	D	Breaking complex programming or modelling tasks into components; addressing subtasks through AI assistance	Iterative solution refinement	Problem decomposition

33	Problem-solving	Solution development	T4. Error detection and corrective modification	T2	M	M	M	M	M	O	M	M	M	M	O	M	D	O	O	D	D	Requesting solution assistance; exploring approaches; interpreting AI responses; developing and refining solutions	Direct answer and information seeking	Simple prompting / answer seeking
34	Procedural reasoning	Procedural analysis	T2. Interpretation and analytical decomposition	T5	M	M	D	M	M	O	O	D	M	D	-	D	-	O	O	D	D	Requesting stepwise explanations; interpreting generated procedures; examining programming or mathematical solution sequences	Analytical interpretation of AI outputs	Contextual & iterative prompting
35	Questioning and strategic questioning	Strategic inquiry	T1. Strategic questioning and task framing	T6	M	O	O	M	M	O	M	O	O	O	O	O	M	O	M	O	M	Formulating progressive prompts; requesting clarification; probing AI responses; adapting questions to task requirements	Iterative questioning and prompt refinement	Contextual & iterative prompting
36	Reflection and reflective thinking	Reflective evaluation	T6. Reflection, metacognition, and self-regulation	T3	D	D	D	O	O	O	D	D	O	D	D	M	M	O	D	M	D	Reviewing AI interactions; reflecting on generated responses; reconsidering prompting strategies and learning experiences	Reflective and self-regulated interaction	Contextual & iterative prompting
37	Scientific reasoning	Scientific interpretation	T2. Interpretation and analytical decomposition	T1	-	D	-	-	D	-	-	-	-	D	D	-	D	O	-	D	D	Exploring scientific explanations through guided inquiry; interpreting disciplinary information and examining generated responses	Analytical interpretation of AI outputs	Contextual & iterative prompting

38	Self-regulation and autonomous learning	Learning regulation and autonomy	T6. Reflection, metacognition, and self-regulation	T5	D	D	D	D	D	O	M	D	D	D	O	O	O	O	D	M	D	Managing AI assistance; determining when to use or question generated outputs; retaining or delegating responsibility during problem-solving	Reflective and self-regulated interaction	Contextual & iterative prompting
39	Socratic questioning	Dialogic and probing inquiry	T1. Strategic questioning and task framing	T2	-	O	-	-	-	-	-	-	-	-	-	-	-	-	-	D	-	Engaging in guided questioning; responding to prompts and hints; progressively examining explanations rather than relying only on direct answers	Guided inquiry and exploratory dialogue	Contextual & iterative prompting
40	Solution validation	Solution verification	T3. Comparison, evaluation, and verification	T4	O	O	O	M	D	O	M	O	M	O	O	O	O	O	O	O	O	D	Comparing generated solutions with independent answers; testing code or physical systems; examining whether proposed solutions work	Testing and validation of outputs	Evaluation & verification
41	Source and reference checking	Source credibility verification	T3. Comparison, evaluation, and verification	T2	-	-	-	-	-	-	-	-	-	-	O	-	O	-	-	O	D	Checking AI-generated references; verifying claims against external information; examining whether supporting sources are reliable	Evidence and source verification	Evaluation & verification
42	Synthesis and integration of information	Knowledge synthesis	T2. Interpretation and analytical decomposition	T3	M	D	O	O	M	O	O	M	D	O	O	O	-	O	O	M	O	Summarizing AI-supported discussions; integrating generated explanations with other information; producing revised outputs	Analytical interpretation of AI outputs	Evaluation & verification
43	Task organization and sequencing	Task structuring and sequencing	T1. Strategic questioning and task framing	T2	O	O	O	O	M	O	O	O	M	O	O	O	O	O	M	D	O	Structuring tasks through progressive prompts;	Problem decomposition and task structuring	Problem decomposition

																					organizing problem-solving activities; addressing subtasks in a planned order		
44	Troubleshooting	Problem diagnosis and correction	T4. Error detection and corrective modification	T3	O	O	O	O	-	O	-	-	O	O	O	D	O	O	O	-	D	Diagnosing problems in generated code or engineering recommendations; requesting corrective assistance; testing revised solutions	Error identification and diagnosis	Modification, challenge, or rejection of AI-generated outputs
45	Verification and cross-checking	Cross-validation of information	T3. Comparison, evaluation, and verification	T2	O	O	O	M	-	O	M	D	O	O	O	O	O	O	-	O	-	Comparing AI outputs with computational tools, independent solutions, external sources, or real-world observations	Comparative analysis of solutions	Evaluation & verification

Note. Supplementary Table S3 maps 45 initial critical-thinking (CT)-related cognitive process codes to six higher-order interpretive themes and presents the corresponding evidence classifications across 17 included studies. Each study–code intersection was classified as D (Discussed) when the process was addressed in the study's findings or self-reported by participants without direct behavioral documentation; O (Observed) when the process was supported by documented student interactions, prompts, learning artifacts, or other empirical observations; M (Measured) when the specific process was explicitly operationalized and assessed using a defined instrument, rubric, coding procedure, or quantitative measure; or - (Not established) when the available evidence did not sufficiently establish the process. Where more than one evidence level was supported, the highest applicable classification was recorded (M > O > D). Classifications were based on evidence from the individual studies rather than instructional intentions, theoretical assumptions, or findings cited from other research. Quantitative interaction indicators, self-reported perceptions, and task-performance outcomes were not automatically treated as direct measures of cognitive proficiency. Accordingly, M denotes explicit measurement of the coded process or its clearly defined operational indicator, not necessarily demonstrated improvement in critical-thinking ability. The classifications describe the nature of the reported evidence rather than the effectiveness of GenAI in developing the process. A classification of - indicates insufficient reported evidence and does not imply that the process was absent.

Supplementary Table S4. Consolidated distribution by primary theme

Theme	**Assigned keywords**	**Count**
1. Strategic questioning and task framing	7, 22, 24, 35, 39, 43	6
2. Interpretation and analytical decomposition	1, 4, 5, 6, 8, 12, 21, 26, 27, 28, 32, 34, 37, 42	14
3. Comparison, evaluation, and verification	2, 3, 9, 10, 11, 18, 23, 40, 41, 45	10
4. Error detection and corrective modification	13, 15, 16, 20, 25, 31, 33, 44	8
5. Justification, judgment, and independent decision-making	14, 17, 19	3
6. Reflection, metacognition, and self-regulation	29, 30, 36, 38	4
Total		**45**

Supplementary Table S5. Summary of student–GenAI interaction patterns

Interaction pattern	Representative initial codes	Summary of student–GenAI interaction	Representative studies and supporting evidence
1. Simple prompting / answer seeking	Problem-solving (specifically direct answer-seeking instances)	Students use direct or minimally elaborated prompts to obtain answers, explanations, calculations, or complete solutions. This pattern involves greater delegation of cognitive work, particularly when learners accept generated responses with limited independent examination.	Article 1 (Gong et al., 2025): guided interaction involving requests for AI assistance; Article 47 (Supriadi & Suherman, 2026): naive copy-paste prompting; Article 80 (Gallart et al., 2026): AI-use profiles involving different levels of delegation in mathematical modelling; Article 132 (Tassoti, 2024): initial prompts frequently copied existing chemistry questions; Article 160 (Cruz-Angeles et al., 2026): many students relied on a faculty-provided example prompt without substantial modification.
2. Contextual & iterative prompting	Conceptual clarification; conceptual understanding and knowledge construction; constraint specification; explanatory reasoning; hypothesis generation and testing; inquiry and investigation; metacognition and metacognitive awareness; metacognitive planning and monitoring; procedural reasoning; questioning and strategic questioning; reflection and reflective thinking; scientific reasoning; self-regulation and autonomous learning; Socratic questioning	Students provide contextual information, specify task requirements, formulate meaningful questions, request clarification, and progressively refine prompts in response to AI-generated outputs. This reflects more purposeful direction of GenAI interactions.	Article 1: progressive prompting and follow-up interaction; Article 25 (Lin et al., 2026): structured Socratic dialogue and guided inquiry; Article 47: contextual, algorithmic, and iterative/debugging prompting; Article 132: prompt refinement using structured prompting strategies; Article 162 (Picardal, 2025): contextual, iterative, file-based, and role-based prompting; Article 152 (Buendía-García and Piris-Ruano, 2025): incremental prompting and task organization.
3. Problem decomposition	Computational thinking; problem decomposition; task organization and sequencing	Students break complex STEM problems into manageable components, identify relevant subtasks, interpret generated explanations, and selectively use GenAI to support mathematical, scientific, programming, or engineering activities. Learners retain responsibility for organizing and integrating the problem-solving process.	Article 53 (Dominguez et al., 2026): AI-supported engineering activities involving multistep problem-solving, case analysis, and design tasks; Article 80: mathematical modelling activities involving varying degrees of learner autonomy and AI delegation; Article 87 (Nathaniel et al., 2025): structured programming activities involving logical planning and iterative coding; Article 152: task organization and interpretation of generated code; Article 55 (Aruleba et al., 2025): independent programming, decomposition, planning, and comparison of student-written and AI-generated code; Article 162: AI-supported conceptualization, research planning, and data interpretation.
4. Evaluation & verification	Analytical thinking / analytical skills; comparison of alternatives / alternative solutions; critical evaluation of AI-generated outputs; critical thinking; data analysis and interpretation; empirical testing and experimentation; ethical reasoning and ethical reflection; evidence-based reasoning; higher-order thinking skills; information reliability and accuracy assessment; logical reasoning; mathematical modelling and analytical interpretation; mathematical reasoning; solution validation; source and reference checking; synthesis and integration of information; verification and cross-checking	Students assess AI-generated responses by examining accuracy, relevance, consistency, and credibility. They compare outputs against independent solutions, disciplinary knowledge, external literature, computational tools, or empirical observations before deciding whether the results are appropriate.	Article 37 (Matzakos and Moundridou, 2025): comparison of LLM-generated mathematical results with Maxima; Article 96 (Reddy et al., 2024): checking AI-generated scientific information and references, including identification of nonexistent citations; Article 128 (X. Zhang, 2025): independently developed database solutions compared with AI-generated alternatives; Article 160: instructional requirements for cross-checking AI-generated claims against academic literature and institutional sources; Article 151 (Chen et al., 2024): testing and validation of AI-generated programming recommendations in intelligent-vehicle activities; Article 89 (Štuikys et al., 2025): robotics prototyping, testing, and evaluation; Article 47: mathematical reasoning supported by verification and questioning of AI-generated solutions.
5. Modification, challenge, or rejection of AI-generated outputs	Challenging or objecting to AI responses; creativity and creative thinking; critiquing AI-generated responses; debugging and code optimization; decision-making and independent judgment; error detection and correction; feedback seeking and utilization; iterative refinement and improvement; model construction and refinement; troubleshooting	Students identify inaccuracies, challenge unsupported claims, correct faulty reasoning, debug generated code, modify proposed solutions, and reject recommendations that do not satisfy disciplinary or task requirements. This pattern represents greater retention of cognitive responsibility through independent judgment and corrective action.	Article 1: objections, criticism, and follow-up questioning during AI interaction; Article 47: iterative debugging and correction of mathematical solutions; Article 55: critical comparison and refinement of AI-generated programming solutions; Article 37: cross-verification and correction of LLM-generated mathematical results; Article 89:troubleshooting and refinement during robotics development; Article 151: iterative testing and refinement of AI-generated code and control recommendations; Article 96: critical examination and correction of AI-assisted scientific writing and references.

Note. The five higher-order interaction patterns were derived through thematic synthesis of the reported student–GenAI interaction behaviors across the included studies. Representative keywords reflect the predominant interaction associated with each initial code rather than mutually exclusive behavioral categories. Individual studies may contribute evidence to multiple patterns. The reported evidence includes observed interactions, documented learning activities, and study authors' interpretations; therefore, the identified patterns should not be interpreted as demonstrating improvements in independently assessed critical-thinking ability.

Supplementary Table S6. Factors Associated with Student–GenAI Interaction Across the Reviewed Studies

Major category	Specific factors	Article ID	Relationship to student–GenAI interaction

1. Personal factors	Prompt quality and logical reasoning	47	Examines the relationship between prompt quality and students' logical reasoning.
	Prior problem-solving experience	80	Examines how experience relates to ChatGPT use during mathematical modelling.
	Prompting practices and performance	53	Investigates prompting behavior in relation to engineering learning performance.
	Epistemic trust and uncritical acceptance of AI responses	47	Examines how prompt quality relates to students' acceptance, verification, and questioning of AI-generated mathematical responses.
	Disciplinary knowledge and capacity to evaluate AI-generated information	47, 96	Provides evidence relevant to students' evaluation of AI-generated mathematical or scientific information; its independent influence requires cautious interpretation.
	Metacognitive reflection and monitoring	47, 162	Examines reflective processes associated with prompting and students' reported GenAI experiences.
	Prompt efficiency, refinement depth, and query frequency	53	Examines relationships between specific prompting behaviors and assignment performance.
	Confidence and comprehension needs	163	Reports students' experiences and difficulties when using AI-driven prompts.
	Existing computational-thinking proficiency	1	Examines interaction patterns among students with different computational-thinking levels.
	Initial prompting practices and prompt refinement	132	Identifies difficulties in developing effective prompts and changes in prompting practices.
	Perceived quality of AI-generated responses	132	Examines students' satisfaction with AI responses following structured prompting.
	Contextual, iterative, and role-based prompting strategies	162	Examines students' use of different prompting strategies in academic research activities.
	Cognitive load and comprehension difficulties	162	Identifies comprehension difficulties as a reported reason for seeking AI assistance.
	AI literacy	158, 162	Students' understanding of GenAI capabilities, limitations, and appropriate use may influence how they engage with AI-generated responses. The cited studies provide contextual evidence rather than direct measurement of AI literacy as a predictor.
	Prompt literacy and prompt quality	47, 53, 132	Examines how students formulate, structure, and refine prompts, including relationships with logical reasoning, response evaluation, and task performance.
	Disciplinary knowledge	47, 96	Provides evidence relevant to students' ability to interpret and evaluate AI-generated mathematical and scientific information, although disciplinary knowledge was not established as an independent predictor.
	Confidence in grammar and academic writing	162	Reports students' use of GenAI to address uncertainty about grammar and academic tone.
	Desire for validation and feedback	162	Identifies students' need for confirmation and feedback when interacting with GenAI.
	Personalized learning needs	162	Reports students' use of GenAI to obtain assistance suited to their learning needs.
2. Instructional factors	Structured prompt training	85	Examines prompt training in programming and data analysis.
	Guided inquiry and instructional scaffolding	25	Examines a 5E-structured GenAI coach designed to support inquiry-based learning.
	Socratic questioning, hints, and corrective feedback	25	Incorporates guided questioning and feedback into a structured GenAI learning environment.
	Incremental prompting and task scaffolds	152	Structures GenAI interactions during web-development learning activities.
	Few-shot prompting, chain prompting, and CLEAR framework	85	Examines explicit instruction in structured prompting techniques.
	Computational verification of AI-generated outputs	37	Uses mathematical software to check LLM-generated results or hypotheses.
	Physical or task-based validation of AI recommendations	89, 151	Places AI-generated recommendations within engineering or robotics tasks involving practical testing.

	Verification and critical evaluation requirements	96	Incorporates evaluation of AI-generated information into a biochemistry writing assignment.
	Inquiry-based learning and iterative testing	151	Examines GenAI integration within a structured robotics competition learning model.
	Progressive prompting support	1	Examines progressive prompting in programming education.
	Structured prompting framework	132	Examines the Five S prompting framework and changes in students' prompting practices.
	Clarity of acceptable GenAI use	158	Identifies uncertainty about instructors' expectations as a potential barrier to GenAI use.
	Course-level guidance and scaffolding	158	Describes proposed instructional guidance and structured support for GenAI use.
	Guidance on academic writing structure	162	Reports students' need for support in organizing and formatting academic work.
	Feedback and validation opportunities	162	Identifies feedback and validation as reasons for students' engagement with GenAI.
3. Contextual and task-related factors	Task complexity	80	Examines differences in ChatGPT use across mathematical modelling tasks of varying complexity.
	Web-development and UX design demands	152	Examines GenAI use during web-development and UX design activities.
	Mathematical and computational task demands	37	Investigates GenAI integration in mathematics learning among civil engineering students.
	Engineering design and sustainability-oriented problem solving	89	Examines GenAI-supported problem solving in design-based STEM learning.
	Availability and characteristics of GenAI support	152	Provides contextual evidence about students' experiences with AI-supported learning tasks.
	Time pressure and academic workload	162	Identifies workload and time constraints as reported reasons for seeking GenAI assistance.
	Academic research-writing requirements	162	Examines GenAI use for content development, methodological guidance, data interpretation, and writing refinement.
	Limited access to academic resources	162	Identifies resource limitations as a reported factor in students' use of GenAI.
	Writing structure and formatting demands	162	Examines students' use of GenAI to address academic writing and formatting challenges.

Note: Article IDs correspond to the numbered articles in the review dataset. Factors are grouped into personal, instructional, and contextual/task-related categories. Categories are not mutually exclusive, and a study may contribute to multiple factors. The relationships include investigated associations, reported experiences, and contextual observations rather than necessarily established causal effects.

Supplementary Table S7. Database-Specific Search Strategies, Execution Dates, and Applied Filters

Database	**Exact search string**	**Search date**	**Fields searched**	**Filters applied**	**Records retrieved**
Scopus	TITLE-ABS-KEY ("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) AND (LIMIT-TO (DOCTYPE,"ar"))	August 5, 2026	TITLE-ABS-KEY("prompt engineering") – Only the phrase “prompt engineering” is explicitly restricted to the title, abstract, and keywords.	Article (ar)	165
Web of Science	Using Web of Science (WoS) Advanced Search – Query Builder: TS=("prompt engineering") AND TS=(ChatGPT OR GPT OR LLM* OR "large language model*") AND TS=("critical thinking") AND TS=(STEM OR mathematics OR engineering OR science OR computing OR technology)	October 30, 2026	Topic	Includes journal articles	34

IEEE Xplore	Using IEEE Xplore – Advanced Search (Command Search): ("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)	October 30, 2026	All Metadata	Journal articles, excluding reviews	5