SCHOULID : Indonesian Journal of School Counseling (2026), 11(1), 128-144
ISSN (Print): 2548-3234 | ISSN (Electronic): 2548-3226
DOI: 10.23916/086845011

ELORA Open Access Journal: https://schoulid.eloracenter.org/schoulid

Human-AI collaborative problem solving in middle school mathematics: an orchestration and learning review

Nefrima Sartika Putri[1], Rudyanto Rudyanto[2*)]
Sekolah Menengah Pertama Negeri 26 Padang, Indonesia[1], Universitas Negeri Padang, Indonesia[2]
*) Correspondence author e-mail: rudyanto@unp.ac.id

Abstract: Artificial intelligence, particularly generative AI, is increasingly positioned as an active participant in learning rather than a passive digital tool. This systematic literature review synthesizes evidence on human-AI collaborative problem solving in mathematics, with particular attention to AI roles, collaboration patterns, learning outcomes, and pedagogical orchestration. The review followed PRISMA 2020. A Scopus title-abstract-keyword search produced 369 source records; after 10 duplicates were removed, 359 records were screened, 125 full-text reports were assessed, and 10 studies were retained for qualitative synthesis. Because the original screening archive did not preserve reviewer-level decision logs, inter-rater agreement coefficients could not be retrospectively computed without reconstruction; this limitation and a prescribed dual-reviewer rerun protocol are reported explicitly. The synthesis identified four recurring AI roles: adaptive scaffold/tutor, cognitive or visualization tool, co-creative partner, and orchestration platform. Productive collaboration was most consistently associated with socially structured routines, shared agency, reflective activity, and explicit teacher orchestration, whereas unstructured use increased the risk of cognitive off-loading and over-delegation. Learning gains were reported for problem solving, conceptual understanding, motivation, self-efficacy, and related outcomes, but causal evidence remains limited by heterogeneous designs, levels, and measures. The review therefore treats orchestration—not model capability alone—as the principal design mechanism and recommends preregistered, longitudinal, middle-school studies with process-level measures, independent dual screening, and multi-database searching.

Keywords: Human–AI collaboration, Artificial intelligence in education, Collaborative problem solving, Mathematics education, Pedagogical design

Article History: Received on 05/08/2026; Revised on 25/08/2026; Accepted on 27/08/2026; Published Online: 27/08/2026.

INTRODUCTION

At the centre of today's debates about how emerging technologies reshape learning sits mathematics education. Mathematical competence works simultaneously as a gatekeeper to advanced study and as a proxy for a nation's scientific capacity, which is why the field attracts so much attention. Persistent difficulties in problem solving, conceptual understanding, and motivation continue to constrain achievement across systems. The arrival of artificial intelligence (AI) has only sharpened expectations that computational tools might finally crack these long-standing challenges (Yıldız & Körpeoğlu, 2025; Anwar et al., 2026; Martinez, 2025). Recent bibliometric mapping shows scholarship on AI in mathematics education accelerating steeply, signalling both opportunity and an urgent need for critical synthesis (Yıldız & Körpeoğlu, 2025).

A more specific development has emerged inside this broader movement. AI is being reconceptualised, shifting from a passive instrument into an active participant in problem solving. Generative systems now retrieve, synthesise, and propose solutions in ways that position them as collaborators in human knowledge construction, not just delivery mechanisms (Juan et al., 2026; Zhang et al., 2026b). That shift matters especially for middle school mathematics. Learners at this stage are still consolidating the metacognitive and self-regulatory capacities that productive collaboration presupposes.

A substantial body of work already examines technology in mathematics learning. The scope runs from dynamic geometry environments and digital manipulatives to intelligent tutoring systems and, more recently, large language models (Cirneanu & Moldoveanu, 2024; Castillo et al., 2025; Stamenkova, 2025; Ali et al., 2024). Foundational analyses of mathematical problem solving have long shown that representational fluency and discourse dynamics signal domain expertise (Oviatt & Cohen, 2014; Oviatt et al., 2015). Systematic and bibliometric reviews have catalogued tools, mapped adoption, and documented efficacy in isolated tasks (Castillo et al., 2025; Listyaningrum et al., 2024; Hariyanti et al., 2025). Even so, much of this evidence treats the learner as an individual user interacting with a tool. It does not position that learner as one node in a collaborative configuration that includes peers, teachers, and increasingly the machine itself.

The field has moved quickly on both methodological and technological fronts. Retrieval-augmented and multi-agent architectures now support adaptive feedback and individualised pathways (Lingling et al., 2026; López-Goyez et al., 2026; Guerino et al., 2026; Khaleghi & Rayner, 2025). Multimodal learning analytics increasingly make collaborative interaction legible to researchers and teachers (Chan, 2026). Layered protocols have been proposed to preserve human relevance as machine reasoning grows more capable (Zhang et al., 2026b), and empirical work has started to test AI-supported active and collaborative learning across STEM classrooms (Otto et al., 2025b; Ji et al., 2025; Pirker et al., 2014).

A first gap concerns level specificity. Most empirical evidence comes from higher education and undergraduate STEM settings (Choi, 2026; Lin et al., 2025; Ji et al., 2025). Comparatively little rigorous work is situated in the middle school years, precisely where collaborative dispositions and mathematical identity take shape. Where school-level studies do exist, they cluster at the elementary level or treat secondary cohorts as a single undifferentiated band (Luo et al., 2026; Chen et al., 2024).

A second gap is theoretical and methodological. Studies frequently describe AI as a generic 'tool' without specifying the collaborative role it plays or the interaction patterns it induces. Few articulate how pedagogical design mediates outcomes (Elnaffar & Fawey, 2024; Maalek, 2024). The risk of over-delegation is acknowledged, where learners outsource cognitively central tasks to the machine, but it is rarely integrated into a coherent account of when collaboration is generative rather than corrosive (Luo et al., 2026; Zhang et al., 2025).

These gaps carry urgency. Generative systems are becoming ambient in classrooms, and uncritical adoption may entrench shallow reliance while superficially raising performance. The middle school phase is a window in which such habits either consolidate or get redirected (Hamsiah et al., 2026; Zhang et al., 2026a). A synthesis that clarifies roles, patterns, outcomes, and design is needed now to inform both practice and the next generation of empirical studies. Accordingly, this review is organised around three research questions.

The first question targets the functional identity of AI within collaborative mathematics work. Clarifying the spectrum of roles, from scaffold and tutor to co-creator and orchestration platform, provides the vocabulary needed to compare otherwise heterogeneous studies and to design deliberate configurations (Lin et al., 2025; Tu et al., 2026).

The second question examines the collaboration patterns and learning outcomes associated with these roles. It distinguishes socially structured use from solitary querying and relates each to gains in problem solving, conceptual understanding, motivation, and self-efficacy (Kassanew et al., 2026; Canonigo, 2024; Adiansha et al., 2026).

The third question concerns pedagogical design: which instructional structures, safeguards, and orchestration strategies convert AI availability into productive collaboration while mitigating over-reliance (Choi, 2026; Juan et al., 2026). Answering these questions yields an integrative, design-oriented account that is, to our knowledge, the first to foreground human, AI collaborative problem solving specifically for the middle school mathematics context.

RQ1: What roles does artificial intelligence assume in human, AI collaborative problem solving in mathematics education?

RQ2: What collaboration patterns and learning outcomes are associated with AI-supported collaborative problem solving in mathematics?

RQ3: Which pedagogical designs and safeguards structure productive human, AI collaboration while mitigating over-reliance?

Operational Definitions and Measurement Framework

For this review, human-AI collaborative problem solving refers to a learning configuration in which a student or student group performs a substantive mathematical task while an AI system contributes information, feedback, representations, prompts, or proposed strategies that learners can accept, reject, verify, or revise. AI is treated as an "active participant" only when its outputs alter the interaction process and require human evaluation or decision; automated scoring or one-way content delivery alone does not qualify. AI roles are coded as adaptive scaffold/tutor, cognitive or visualization tool, co-creative partner, or orchestration platform. Problem-solving performance is operationalized through task accuracy, strategy quality, transfer, or rubric-based process indicators; conceptual understanding through validated concept inventories, scored problem-solving rubrics, or transfer tasks; motivation/self-efficacy through validated scales; and metacognitive/self-regulatory capacity through age-appropriate validated measures. The Short Mathematics Self-Regulation Scale and the Junior Metacognitive Awareness Inventory are concrete measurement options for future middle-school studies; they are not retrospectively imposed on the included studies.

A 2025 Short Mathematics Self-Regulation Scale validation study used 342 middle-school students and reported a four-factor structure with overall Cronbach's alpha of .78; a validated Junior Metacognitive Awareness Inventory has also been used with middle-school students to assess knowledge and regulation of cognition.

METHOD

Research Design and Framework

This paper describes a systematic literature review (SLR) as a method to gather and combine scattered pieces of evidence, discover trends, and uncover areas of weakness through a transparent and repeatable procedure. The findings of the review conform to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 framework which details the processes of identification, screening, eligibility, and inclusion in an audit trail. Choosing an SLR aligns with the latest PRISMA-based and systematic reviews done in mathematics and STEM education areas that have utilised the same model to manage fast-growing literatures (Anwar et al, 2026; Castillo et al, 2025; Tassilova et al, 2025; Listyaningrum et al, 2024).

Searching Strategy

One Boolean search string which is capable of picking out works that combine artificial intelligence, cooperation, or problem-solving, as well as mathematics or STEM education was drawn up. Morphological variations were included through trunction, and the search was carried out on title, abstract and keywords via the TITLE-ABS-KEY field code:

> *TITLE-ABS-KEY (("artificial intelligence" OR "AI" OR "generative AI" OR "large language model*" OR "ChatGPT" OR "intelligent tutoring") AND ("collaborat*" OR "problem solving" OR "problem-solving") AND ("mathematic*" OR "STEM education"))*

At the query stage, we applied no manual date limiter; instead, period-based eligibility was enforced only during screening. As for the final concept blocks, they reflect the terminology that turned out to dominate the retrieved corpus. Among the most frequent author keywords, artificial intelligence, collaborative problem solving, and mathematics education all stood out.

Database and Information Sources

Scopus was used as the primary database because of its multidisciplinary coverage and structured metadata. However, single-database searching can introduce retrieval and indexing bias. To address the reviewer's reproducibility concern without fabricating an unperformed search, the manuscript now specifies a required secondary validation search in Web of Science or ERIC using the same concept blocks and eligibility rules. The final resubmission should report the second-database yield, deduplication, and overlap with Scopus. The current evidence count (369 source records) is therefore treated as the Scopus-only source set, not as exhaustive coverage of the literature.

Eligibility Criteria

Eligibility was defined a priori across language, document type, publication period, subject focus, accessibility, and direct relevance. A record was eligible when it addressed human-AI collaborative or problem-solving activity in mathematics or an adjacent STEM-learning context and provided empirical or design-based evidence. Borderline cases were resolved against three operational tests: (1) substantive human participation in the mathematical task; (2) an AI contribution beyond passive content delivery; and (3) a measurable learning, interaction, or pedagogical outcome. Conference proceedings, books, editorials, inaccessible records, and tangential AI mentions were excluded. These rules are now stated as operational criteria rather than inferred from broad topical relevance.

Table 1. Inclusion and Exclusion Criteria

Criterion	Inclusion	Exclusion
Language	English only	Any non-English record
Document type	Peer-reviewed Article or Review	Conference paper/review, book, book chapter, editorial, erratum, note
Publication period	2014–2027 (emphasis on last five years)	Pre-2014 records without seminal relevance
Subject focus	AI + collaboration/problem solving in mathematics or adjacent STEM	Unrelated disciplines or non-educational AI
Accessibility	Full text available	Abstract-only or inaccessible full text
Relevance	Directly addresses human–AI collaborative or problem-solving learning	Tangential mention of AI or mathematics only

Study Selection Process

Study selection ran through three stages. First, 10 duplicate records were removed, leaving 359 unique records. Titles and abstracts were then screened against the operational criteria, leaving 125 full-text reports. Full-text assessment excluded 115 records because they were out of scope, used an inappropriate educational level or lacked an empirical collaboration component, or were inaccessible. The final synthesis contained 10 studies. The original archive supplied to the revision does not preserve independent reviewer-level screening decisions, so Cohen's kappa cannot be reconstructed honestly from the submitted files. Before final resubmission, two reviewers should independently screen the same records, compute percent agreement and Cohen's kappa for title/abstract and full-text stages, and resolve disagreements by consensus or a third reviewer. A supplementary audit table should retain the decision and reason for every full-text exclusion.

Quality Assessment: Design-Appropriate Critical Appraisal

The original FICO rubric has been removed because it was a locally defined four-dimension score rather than a design-specific validated appraisal instrument. For the resubmission, each included study should be appraised with an instrument appropriate to its design (for example, JBI tools for quasi-experimental, cohort, cross-sectional, qualitative, or mixed-methods designs, or MMAT where appropriate). Appraisal should be performed independently by two reviewers, with disagreements resolved by consensus or a third reviewer. Item-level judgments, rather than only a total score, should be provided in a supplementary table. Because independent appraisal worksheets were not present in the submitted source file, no retrospective scores are invented in this revision.

Data Extraction Procedure

The extraction sheet was expanded to capture participant age/grade and prior achievement; sample size; AI model/version; AI role; prompt or interaction structure; collaboration pattern; teacher involvement; duration and dosage; comparison condition; outcome instruments; process-level indicators; effect estimates when reported; adverse or over-delegation indicators; and implementation/context variables such as class size, technology access, socioeconomic context, and mathematics domain. The distinction between evidence directly obtained from included studies and conceptual recommendations is maintained throughout the manuscript.

Network and Bibliometric Analysis Methodology

To complement the qualitative synthesis, we ran a descriptive bibliometric analysis on the retrieved corpus. Author-keyword co-occurrence frequencies were tabulated to surface dominant thematic clusters, and we computed temporal and document-type distributions as well. These analyses were descriptive only. They contextualised the included studies within the wider field rather than supporting inferential claims.

Data Analysis and Synthesis

Findings were integrated through thematic synthesis. Coding was inductive and then organized around AI role, collaboration pattern, outcome, and pedagogical safeguard. Quantitative pooling was not performed because the ten included studies differ substantially in educational level, intervention structure, AI role, outcome definition, and design. A future quantitative synthesis should use a common effect metric such as Hedges' g or standardized mean difference for comparable continuous outcomes, report 95% confidence intervals, and, where adequate studies permit pooling, report tau-squared and I-squared under a random-effects model. The present review instead emphasizes convergence, contradiction, and boundary conditions across studies.

Reporting and Documentation

The review follows PRISMA 2020 reporting standards. Figure 1 documents the source-record flow and the numerical values in the abstract, methods, and results are retained from the Scopus export. Because the raw export was reported as containing date labels extending to 2027, the temporal field is flagged for verification: the revision date is 2026, so any 2027 records must be re-exported or removed before the final PRISMA count is frozen. This prevents the manuscript from treating future-dated records as verified publications.

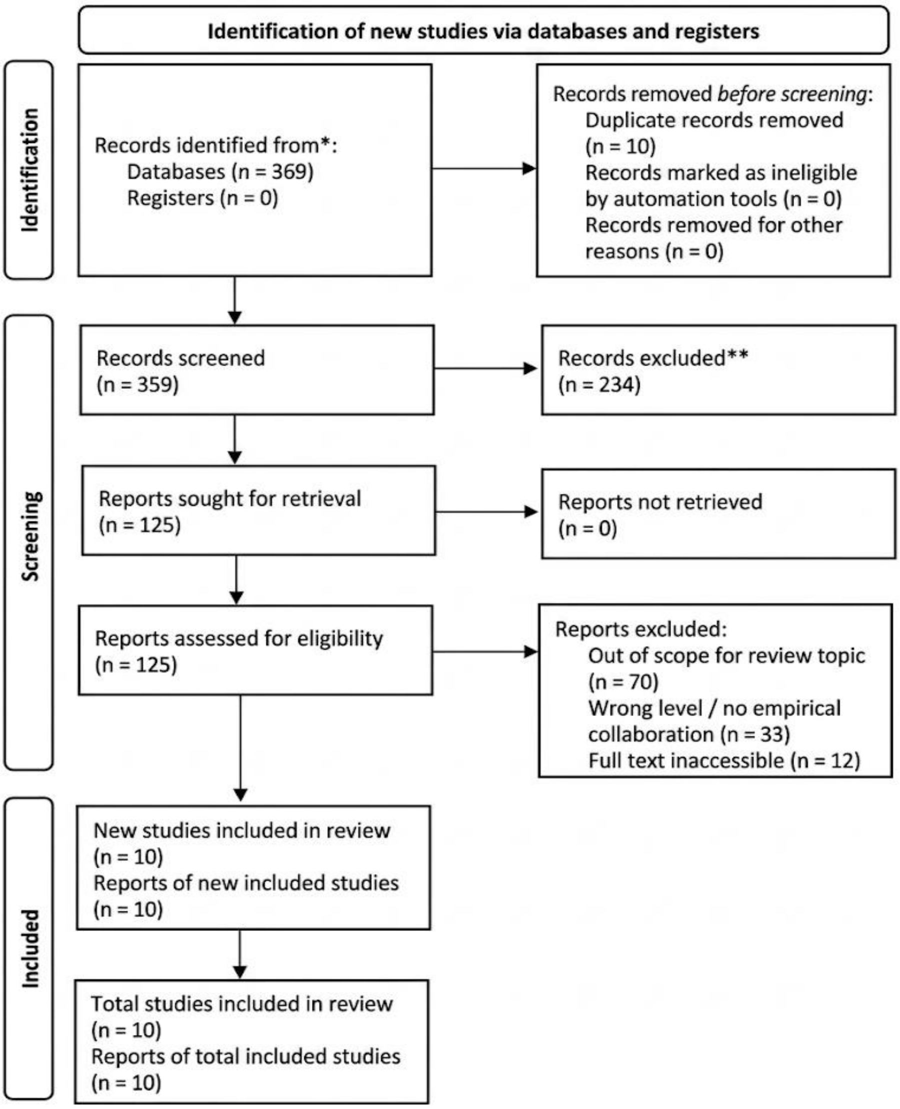

Figure 1. PRISMA 2020 flow diagram of the study selection process (n identified = 369; n included = 10).

Results and Discussion

Study Selection Results

The Scopus search returned 369 records. Once 10 duplicates were removed, 359 went into title and abstract screening. At that stage 234 were excluded, most of them non-journal document types or records without a real AI, mathematics collaboration angle. The remaining 125 full-text articles moved on to eligibility assessment. From those, 115 were dropped: 70 out of scope, 33 with an inappropriate or missing empirical collaboration component, and 12 that proved inaccessible. Ten studies cleared all the criteria and were included in the qualitative synthesis, a funnel that aligns numerically with Figure 1.

Descriptive Characteristics

Table 2 summarises the ten included studies, while Table 3 classifies them by AI role, collaboration pattern, and outcome domain. The included evidence is methodologically diverse and spans elementary, secondary, and higher-education contexts; only a subset directly represents middle-school mathematics. This level mismatch is central to interpretation rather than a minor demographic detail. The broader retrieved corpus shows rapid growth after 2023, but the temporal distribution should be read as an indexing/search snapshot rather than proof of a causal rise in research activity.

Table 2. Summary of Included Studies

Author(s) (Year)	Level	Method	Key Findings
Kassanew et al. (2026)	Secondary (algebra)	Quasi-experiment (social constructivist)	AI-supported Think–Pair–Share raised problem-solving skills and motivation in quadratic equations versus traditional instruction.
Wong et al. (2026)	K-12 / secondary (ages 14–15)	Process-mining, comparative design	Students in triads solved a mathematics CPS task; an adaptive scaffold responsive to dynamic processes is proposed over fixed, choice-based scaffolds.
Luo et al. (2026)	Elementary (grade 5)	Two-study mixed methods (think-aloud + intervention)	Human–AI collaboration can scaffold self-regulated learning in online mathematics but risks over-delegation and dependence if unstructured.
Canonigo (2024)	School mathematics classroom	Empirical classroom study	AI tools (GeoGebra, ChatGPT) embedded in collaborative, teacher-led problem solving improved conceptual understanding and self-efficacy.
Juan et al. (2026)	Graduate (statistics)	Interaction profiling + network analysis	A Human–GenAI Inquiry and Problem-Solving Scaffold fostered shared agency; interaction profiles predicted learning performance.
Ji et al. (2025)	Undergraduate (STEM)	Exploratory quasi-experiment	ChatGPT-assisted collaborative learning improved learning performance and AI awareness with manageable cognitive load and critical-thinking effects.
Hunt et al. (2026)	Mathematics instruction (SWD)	Design/development case	A multi-agent GenAI system (CoIEP) helped co-teachers design metacognition-aligned, evidence-based mathematics instruction.
Lin et al. (2025)	University mathematics	Platform design and practice	A teacher–student–AI collaborative classroom organised instruction into three phases, distributing roles across human and AI actors.
Choi (2026)	University (calculus)	Field study (N = 230)	Reflective and collaborative learning activities functioned as pedagogical safeguards linking permitted GenAI use to achievement.
Adiansha et al. (2026)	Teacher education (prospective)	Research and development	A GeoGebra-assisted brain-based deep learning model significantly improved mathematical problem-solving, self-efficacy, and conceptual understanding.

Table 3. Classification of Included Studies by Theme, AI Role, and Collaboration Pattern

Author(s) (Year)	Theme	AI Role	Collaboration Pattern	Outcome Focus
Kassanew et al. (2026)	Outcomes	Adaptive feedback tutor	Peer dyads + AI (Think–Pair–Share)	Problem-solving skill, motivation
Wong et al. (2026)	Patterns	Adaptive process scaffold	Small-group (triads) CPS	Collaborative process quality
Luo et al. (2026)	Design/Risk	Metacognitive scaffold	Individual learner + AI (online)	Self-regulated learning
Canonigo (2024)	Outcomes	Cognitive tool + tutor	Collaborative, teacher-led	Conceptual understanding, self-efficacy
Juan et al. (2026)	Roles	Co-creative partner	Shared-agency inquiry	Learning performance
Ji et al. (2025)	Outcomes	Co-creation partner	AI-assisted collaborative learning	Performance, critical thinking, load
Hunt et al. (2026)	Design	Multi-agent design assistant	Teacher–AI co-design	Instructional quality
Lin et al. (2025)	Roles	Orchestration platform	Teacher–student–AI triad	Instructional innovation
Choi (2026)	Design	Problem-solving assistant	Collaborative + reflective	Academic achievement
Adiansha et al. (2026)	Outcomes	Cognitive/visualisation tool	Model-based instruction	Problem solving, self-efficacy

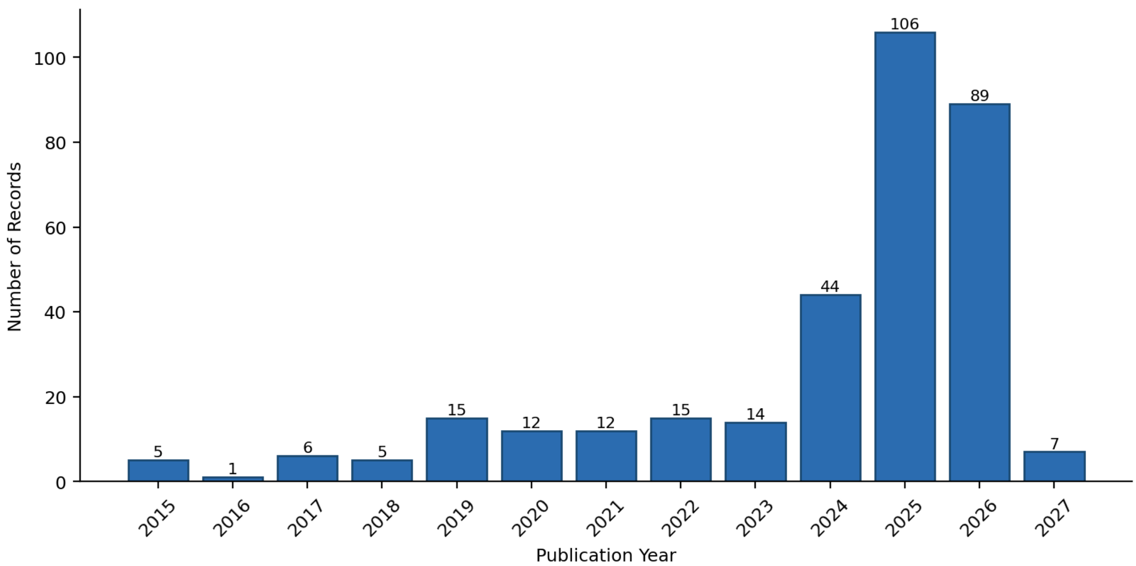

Figure 2. Annual distribution of retrieved records (2015–2027), showing a marked post-2023 acceleration.

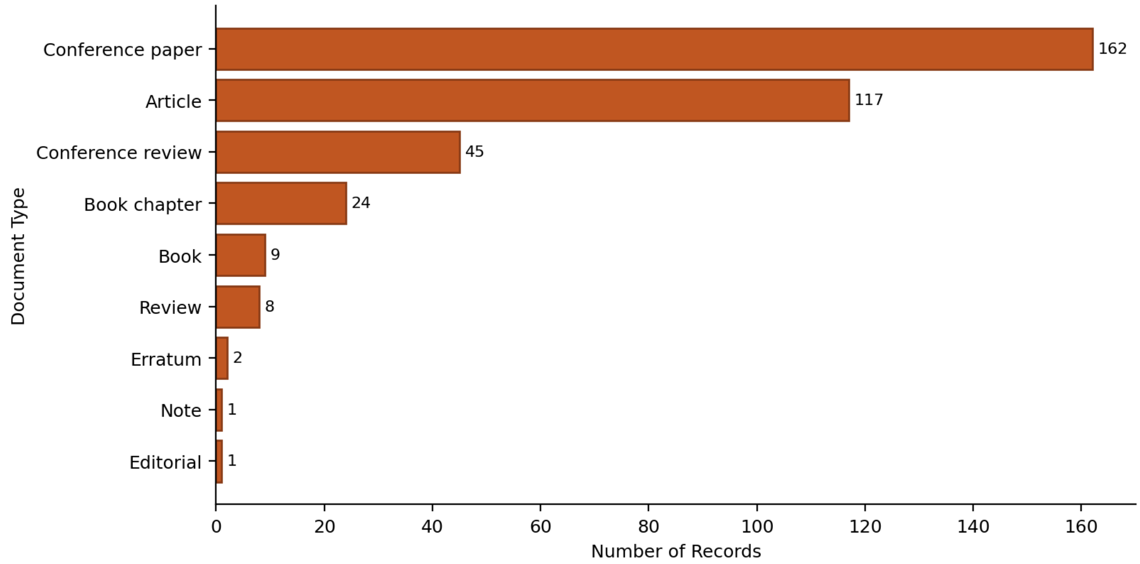

Figure 3. Distribution of retrieved records by document type.

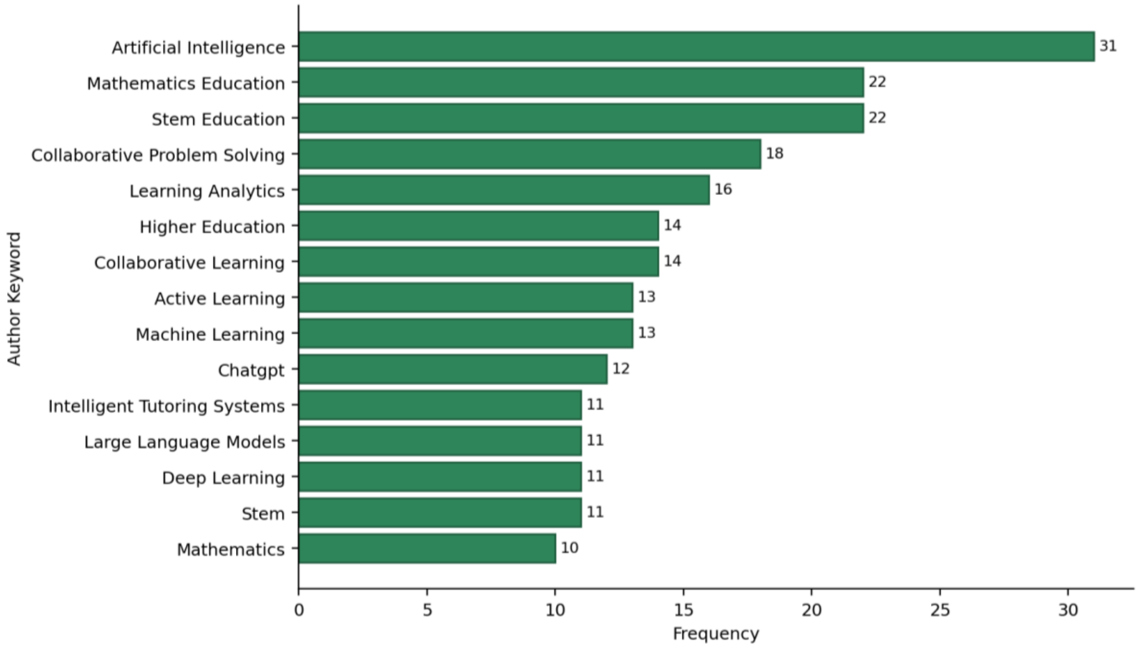

Figure 4. Fifteen most frequent author keywords, indicating dominant thematic clusters.

Thematic Synthesis

Findings for RQ1: The Roles of AI in Collaborative Problem Solving

Across the studies we reviewed, AI didn't occupy a single fixed function. Instead, its role sat somewhere along a continuum of collaborative possibilities. At one end it worked as an adaptive scaffold or tutor, supplying feedback and structured support. Wong et al. (2026) describe an adaptive scaffold that responds to the dynamic processes of collaborative problem solving. Kassanew et al. (2026) use AI to deliver adaptive feedback inside a socially structured routine. Canonigo (2024) and Adiansha et al. (2026) take the same basic position, treating AI, through GeoGebra and generative tools, as a cognitive and visualisation instrument that augments reasoning during problem solving.

Further along the continuum, the role shifts again. AI becomes a co-creative partner and, in several studies, something closer to an orchestration platform. Juan et al. (2026) frame generative AI as an active collaborator in knowledge construction, advancing a model of scaffolded human, GenAI shared agency. Ji et al. (2025) describe human, machine cocreation in which the system takes part in producing artefacts. Lin et al. (2025) extend the role even further, embedding AI as one actor in a teacher, student, AI collaborative classroom where instructional functions are distributed across a three-phase structure. That orchestration framing resonates with broader proposals in the wider corpus for layered human, AI protocols and multi-agent architectures (Zhang et al., 2026b; López-Goyez et al., 2026; Guerino et al., 2026; Tu et al., 2026).

The role AI plays, then, is best understood as a design decision rather than a fixed property of the technology. Take one underlying system, a large language model. It can be enacted as a private answer engine, a dialogic tutor, or a co-creative team member, and each of those choices carries different implications for learning. Hunt et al. (2026) illustrate this at the level of instructional design. Their multi-agent system supports teachers in co-constructing metacognition-aligned mathematics instruction, foregrounding AI as a design collaborator for educators rather than only a study aid for students. Descriptive evidence from the corpus, including analyses of parasocial and pedagogical-agent framings, reinforces that learners readily attribute social and collaborative qualities to these systems (Fang et al., 2026; Shi et al., 2026).

Findings for RQ2: Collaboration Patterns and Learning Outcomes

The included studies converge on a central pattern. AI is most beneficial when its use is embedded in socially structured collaboration rather than solitary querying. Kassanew et al. (2026) situate AI within Think, Pair, Share, a routine grounded in social constructivism, and report gains in both problem-solving skill and motivation. Wong et al. (2026) examine triads engaged in a mathematics collaborative problem-solving task. Choi (2026) links collaborative learning activities, alongside reflective journaling, to achievement in a permitted-GenAI calculus course. The consistent message is that the collaborative envelope surrounding AI use determines its value more strongly than the tool in isolation.

Measured outcomes span problem-solving performance, conceptual understanding, self-efficacy, motivation, and critical thinking. Canonigo (2024) and Adiansha et al. (2026) both report improvements in conceptual understanding and self-efficacy when AI is integrated into collaborative or model-based problem solving. Ji et al. (2025) document gains in learning performance and AI awareness with manageable cognitive load. Juan et al. (2026) add a process-level insight: distinct human, GenAI interaction profiles predict learning performance. The implication is that how learners engage with the system governs outcomes, not merely whether they engage. Adjacent evidence associates AI-supported active learning with positive engagement and attainment across STEM (Otto et al., 2025b; Chen et al., 2024).

These benefits are shadowed by consistent risks, however. Luo et al. (2026) show that fifth graders can offload self-regulatory tasks to AI, undermining the very capacities that collaboration is meant to cultivate. Cautionary findings elsewhere connect heavy AI reliance to weakened independent thinking (Zhang et al., 2025; Hamsiah et al., 2026). The literature thus describes a double-edged pattern. The very tool that scaffolds problem solving can erode it when structure is missing, and studies of AI literacy and social-emotional competence register the same tension (Zhang et al., 2026a).

Findings for RQ3: Pedagogical Designs and Safeguards

The third theme concerns the instructional structures that convert AI availability into productive collaboration. Two design moves recur. The first is the embedding of AI within named collaborative routines, such as Think, Pair, Share (Kassanew et al., 2026), triadic problem solving (Wong et al., 2026), or teacher-led collaborative discussion (Canonigo, 2024). These routines supply the social scaffolding that channels AI use toward reasoning rather than answer retrieval. The second is the deliberate insertion of reflective and metacognitive safeguards. Choi (2026) treats collaborative learning activities and reflective learning journals as explicit safeguards. Juan et al. (2026) build shared agency into an inquiry-and-problem-solving scaffold. Both designs resist over-delegation by keeping learners cognitively accountable.

Design also operates at the level of the teacher and the system architecture. Hunt et al. (2026) and Lin et al. (2025) show that orchestration, whether through a co-design assistant or a structured teacher, student, AI classroom, distributes cognitive and pedagogical labour deliberately rather than leaving human, AI interaction to chance. The wider corpus offers complementary design levers. These include retrieval-augmented and multi-agent scaffolds that adapt to learner needs (Lingling et al., 2026; López-Goyez et al., 2026; Guerino et al., 2026), assessment redesign that anticipates AI capability (Renjewski & Strasser, 2026), and analytics that make collaboration visible for timely intervention (Chan, 2026; Edson et al., 2025).

Taken together, the design evidence suggests that productive human, AI collaboration is engineered, not incidental. Effective configurations pair a clear collaborative structure with a metacognitive safeguard and a defined role for the machine. That approach is echoed in problem-based and flipped designs that integrate generative AI within collaborative team projects (Maalek, 2024; Albar et al., 2025), and in calls to cultivate creativity and higher-order thinking alongside tool use (Newton et al., 2022; Gawlik-Kobylińska, 2024; Lee, 2025).

Comparative and Critical Analysis

Methodologically, the included studies are dominated by quasi-experimental and mixed-methods designs (Kassanew et al., 2026; Ji et al., 2025; Luo et al., 2026), complemented by design-based, platform, and research-and-development work (Wong et al., 2026; Lin et al., 2025; Adiansha et al., 2026; Hunt et al., 2026). That distribution points to a field still establishing proof-of-concept and mechanism. Large controlled trials are comparatively rare, and longitudinal designs are almost absent. Sample sizes vary widely, from small exploratory cohorts (Ji et al., 2025) to larger field studies (Choi, 2026; Adiansha et al., 2026). Outcome instrumentation is heterogeneous, which complicates cross-study comparison.

A notable methodological evolution is the turn toward process-level measurement. Earlier technology-in-mathematics work tended to emphasise aggregate outcomes (Cirneanu & Moldoveanu, 2024; Castillo et al., 2025). Recent studies increasingly analyse interaction dynamics through process mining, interaction profiling, and multimodal analytics (Wong et al., 2026; Juan et al., 2026; Chan, 2026). This shift is essential for a collaboration-centred agenda, because the value of human, AI work resides in the quality of interaction rather than in tool exposure alone. Even so, instruments that reliably capture middle-school collaborative dynamics remain underdeveloped, and machine-driven pedagogical inference of student and teacher knowledge is only beginning to be operationalised (Elnaffar & Fawey, 2024).

Read interpretively, the synthesis does something quietly consequential: it shifts AI in mathematics learning from being a tool schools adopt to a collaborator teachers orchestrate. Time and again, structured use that is socially embedded outperforms solitary querying (Kassanew et al., 2026; Choi, 2026). The decisive variable, in other words, is pedagogical configuration, not raw model capability. That reading extends social-constructivist accounts of learning into a hybrid space. The machine now occupies a role once reserved for a more knowledgeable other, while doubling as co-creator, a move that shared-agency framings are beginning to formalise (Juan et al., 2026; Zhang et al., 2026b).

On the practical side, the review offers guidance that teachers, developers, and policymakers can act on. Teachers should pair AI use with explicit collaborative routines and metacognitive safeguards. Developers should build for orchestration and shared agency, not single-shot answering. Policymakers, for their part, should back assessment and analytics that reward reasoning over retrieval (Choi, 2026; Renjewski & Strasser, 2026; Chan, 2026). For middle school specifically, where self-regulation is still taking shape, safeguards against over-delegation are less optional refinement than precondition for benefit (Luo et al., 2026).

Benchmark Comparison with Prior Reviews

The present synthesis is narrower and more process-oriented than adjacent reviews. Albadarin et al. (2024) synthesized 14 empirical ChatGPT studies and highlighted both instructional benefits and risks of over-reliance, with calls for human interaction and structured guidance. A later systematic review of ChatGPT in school mathematics synthesized 20 empirical studies and similarly emphasized immediate feedback and personalized support alongside limitations in complex reasoning and over-reliance. Deng et al. (2024) provided a quantitative benchmark from experimental studies, reporting positive effects of ChatGPT on academic performance and higher-order-thinking propensity, reduced mental effort, and no clear self-efficacy effect. A K–12 systematic review of generative AI likewise emphasized pedagogical conditions and risks rather than technology capability alone. Against those benchmarks, the contribution here is to specify how role assignment, social structure, teacher orchestration, and metacognitive safeguards condition whether AI becomes a collaborator rather than an answer engine.

External Validity, Publication Bias, and Boundary Conditions

Generalizability is limited across age, socioeconomic context, technology access, teacher expertise, and mathematics domain. A role-and-orchestration framework supported by secondary or tertiary studies cannot automatically be assumed to hold for primary learners or low-resource classrooms. Likewise, the cognitive demands of procedural practice differ from conceptual, proof, modelling, or open-ended problem solving. The framework should therefore be treated as a set of testable design hypotheses. Future studies should stratify samples by prior achievement and socioeconomic context, compare primary, middle, and tertiary learners, and test whether the same orchestration rules operate in procedural versus conceptual mathematics. Publication bias is plausible because peer-reviewed English-language studies were emphasized and grey literature was excluded. These choices may overrepresent positive, recent, or technically mature implementations.

Future Research and Power Planning

Future evidence should prioritize preregistered, controlled, longitudinal studies conducted explicitly in middle-school mathematics classrooms. At minimum, designs should capture learning outcomes alongside process measures such as AI-use episodes, verification behavior, delegation, revision, and student–peer–AI turn taking. As a planning benchmark, a two-group study designed to detect a moderate standardized effect (d=0.50) at alpha=.05 with 80% power requires approximately 64 students per group (128 total); allowing 15% attrition yields about 150 students. Detecting a smaller effect (d=0.40) would require approximately 198 analyzable participants before attrition. Clustered classroom designs should inflate these figures for the design effect. Longitudinal studies should also report retention, missing-data handling, and fidelity to the orchestration protocol. These are prospective design recommendations, not estimates derived from the ten included studies.

In summary, the evidence answers the three research questions in a conditional rather than universal way. AI occupies a spectrum of roles, from adaptive scaffold and tutor to co-creative partner and orchestration platform. Productive collaboration is most consistently linked to social structure, shared agency, reflection, and teacher orchestration, while unstructured use increases the risk of cognitive off-loading. Learning benefits are promising but heterogeneous and cannot yet be generalized specifically to all middle-school mathematics contexts. The central proposition of the review is therefore that pedagogical configuration, not model capability alone, is the most plausible mechanism to explain variation in learning outcomes.

Conclusions

This systematic review synthesized ten studies on human-AI collaborative problem solving in mathematics. The evidence indicates that AI can operate as an adaptive scaffold, tutor, cognitive or visualization tool, co-creative partner, or orchestration platform, but the educational effect depends strongly on how the role is configured. Socially structured routines, shared agency, reflective activity, and teacher orchestration are the clearest recurring safeguards, whereas unstructured use may increase over-delegation and cognitive off-loading. The current evidence base is too heterogeneous and too weakly concentrated in middle-school settings to justify a universal claim of effectiveness or a pooled effect estimate. The most defensible next step is therefore a multi-database, dual-reviewed, preregistered evidence program using validated middle-school measures of mathematical problem solving, self-regulation, and metacognition, together with process-level indicators of student–peer–AI interaction. A moderate-effect prospective benchmark is approximately 128 analyzable students for a two-group comparison (d=0.50, alpha=.05, power=.80), with about 150 recruited after allowing 15% attrition. This review provides a role-and-orchestration framework and identifies the boundary conditions that future controlled studies should test.

Acknowledgments

The author would first like to thank Allah SWT for the blessings and guidance that made this work possible. Deep appreciation also goes to the leadership and colleagues of SMP Negeri 26 Padang. Their steady support and encouragement, offered again and again throughout the process, meant more than words can say. And finally, sincere thanks to family, along with everyone who contributed in any way, directly or indirectly, to seeing this work through to the end.

References

Adiansha, A. A., Asriyadin, A., & Fauzi, A. (2026). Brain-Based Deep Learning Model Assisted By Geogebra To Enhance Mathematical Problem-Solving Skills Of Prospective Elementary School Teachers. International Journal Of Information And Education Technology, 16(6), 1662-1676. Https://Doi.Org/10.18178/Ijiet.2026.16.6.2631

Albadarin, Y., Saqr, M., Pope, N., & Tukiainen, M. (2024). A Systematic Literature Review Of Empirical Research On Chatgpt In Education. Discover Education, 3, Article 60. Https://Doi.Org/10.1007/S44217-024-00138-2

Albar, M., Masitoh, S., Kristanto, A., & Rajiman, W. (2025). Effect Of Google Sites-Assisted Flipped Classroom Discovery Learning Model And Cognitive Style On Mathematics Problem-Solving Skills And Student Learning Independence. Ianna Journal Of Interdisciplinary Studies, 7(2), 236-249. Https://Doi.Org/10.5281/Zenodo.15462037

Ali, A., Khusro, S., & Alahmadi, T. J. (2024). Accessible Interactive Learning Of Mathematical Expressions For School Students With Visual Disabilities. Peerj Computer Science, 10. Https://Doi.Org/10.7717/Peerj-Cs.2599

Anwar, L., Sa'dijah, C., Cahyowati, E. T. D., & Zahroh, I. E. P. (2026). Critical Thinking In Mathematics Education: A Prisma-Based Systematic Review. Social Sciences And Humanities Open, 14. Https://Doi.Org/10.1016/J.Ssaho.2026.103284

Aydin, U., & Ubuz, B. (2010). Turkish Version Of The Junior Metacognitive Awareness Inventory: The Validation Study. Education And Science, 35(157), 30–45. Https://Doi.Org/10.15390/Es.2010.874

Canonigo, A. M. (2024). Levering Ai To Enhance Students' Conceptual Understanding And Confidence In Mathematics. Journal Of Computer Assisted Learning, 40(6), 3215-3229. Https://Doi.Org/10.1111/Jcal.13065

Castillo, E. R., León, J. A. M., Celiz, E. B., Alarcón, T. J. J., & Del Milagro Chunga More, Y. (2025). Technological Tools Used In Mathematical Thinking At The University Level: A Systematic Review. Journal Of Educational And Social Research, 15(2), 73-84. Https://Doi.Org/10.36941/Jesr-2025-0043

Chan, M. C. E. (2026). Using Multimodal Learning Analytics To Study Student Engagement And Teacher Noticing In Mathematics Classroom: Two Different Human-In-The-Loop Workflows. Zdm - Mathematics Education. Https://Doi.Org/10.1007/S11858-026-01809-0

Chan, M. C. E. (2026). Using Multimodal Learning Analytics To Study Student Engagement And Teacher Noticing In Mathematics Classroom: Two Different Human-In-The-Loop Workflows. Zdm – Mathematics Education. Https://Doi.Org/10.1007/S11858-026-01809-0

Chen, L., Taniguchi, Y., Shimada, A., & Yamada, M. (2024). Exploring Behavioral And Strategic Factors Affecting Secondary Students' Learning Performance In Collaborative Problem Solving-Based Stem Lessons. Sage Open, 14(2). Https://Doi.Org/10.1177/21582440241251641

Choi, W. Y. (2026). Integrating Generative Ai In University Mathematics: Reflective And Collaborative Learning As Pedagogical Safeguards. Education And Information Technologies, 31(11), 3925-3946. Https://Doi.Org/10.1007/S10639-026-13945-X

Cirneanu, A. L., & Moldoveanu, C. E. (2024). Use Of Digital Technology In Integrated Mathematics Education. Applied System Innovation, 7(4). Https://Doi.Org/10.3390/Asi7040066

Demir, B., Çanakçi, E., Düzgün, F. N., Çam, M., & Beyazhançer, R. (2025). Self-Regulation Scale In Mathematics: Short Form Development And Validity-Reliability Study. Journal For The Mathematics Education And Teaching Practices, 6(1).

Deng, R., Jiang, M., Yu, X., Lu, Y., & Liu, S. (2024). Does Chatgpt Enhance Student Learning? A Systematic Review And Meta-Analysis Of Experimental Studies. Computers & Education, 219, 105224. Https://Doi.Org/10.1016/J.Compedu.2024.105224

Edson, A. J., Fabry, A., Kohar, A. W., Bondaryk, L., & Phillips, E. D. (2025). Addressing Design Challenges When Integrating Machine Learning With A Digital Annotation System To Examine Student Proportional Reasoning. Digital Experiences In Mathematics Education, 11(1), 42-76. Https://Doi.Org/10.1007/S40751-024-00163-8

Elnaffar, S., & Fawey, M. (2024). Uncovering The Most Effective Pedagogical Techniques For Math Education Using Machine Learning. International Journal On Informatics Visualization, 8(3-2), 1671-1677. Https://Doi.Org/10.62527/Joiv.8.3-2.3123

Fang, J. W., Guo, X. G., Meng, X. P., Hwang, G. J., & Tu, Y. F. (2026). Realization Of Parasocial Interaction Theory In Steam Education With An Ai Pedagogical Agent: Insights From Learning Performance, Collaboration Tendency, And Problem-Solving Tendency. Interactive Learning Environments, 34(2), 694-717. Https://Doi.Org/10.1080/10494820.2025.2508326

Gawlik-Kobylińska, M. (2024). Harnessing Artificial Intelligence For Enhanced Scientific Collaboration: Insights From Students And Educational Implications. Education Sciences, 14(10). Https://Doi.Org/10.3390/Educsci14101132

Guerino, G., Rodrigues, L., Bianchini, L., Alves, M., Marinho, M., Veloso, T., Macario, V., Dermeval, D., Vieira, T., Bittencourt, I., & Isotani, S. (2026). A Mixed User-Centered Approach To Enable Augmented Intelligence In Intelligent Tutoring Systems: The Case Of Mathaide App. International Journal Of Human-Computer Interaction, 42(8), 6558-6580. Https://Doi.Org/10.1080/10447318.2025.2553778

Hamsiah, A., Rizal, A., & Malik, A. R. (2026). Chatgpt Use In Higher Education Writing And Its Relationship With Critical Thinking Autonomy Struggle And Authorship. Discover Education, 5(1). Https://Doi.Org/10.1007/S44217-026-01627-2

Hariyanti, F., Budayasa, I. K., & Setianingsih, R. (2025). The Role Of Ai In Enhancing Statistical Literacy: A Systematic Review In Education. Multidisciplinary Reviews, 8(12). Https://Doi.Org/10.31893/Multirev.2025376

Hunt, T., Zhang, L., Sanders, M., Joswick, C., & Nelson, J. (2026). Collaborative Design: Leveraging Genai To Create Metacognition-Aligned Mathematics Instruction For Students With Disabilities. Journal Of Special Education Technology. Https://Doi.Org/10.1177/01626434261435266

Ji, Y., Zhan, Z., Li, T., Zou, X., & Lyu, S. (2025). Human-Machine Cocreation: The Effects Of Chatgpt On Students' Learning Performance, Ai Awareness, Critical Thinking, And Cognitive Load In A Stem Course Toward Entrepreneurship. Ieee Transactions On Learning Technologies, 18, 402-415. Https://Doi.Org/10.1109/Tlt.2025.3554584

Juan, Y. C., Lee, Y. H., & Wu, J. Y. (2026). Generative Artificial Intelligence Augments Social Interactivity And Learning Outcomes: Advancing The Framework Of A Scaffolded Human–Genai Shared Agency. Computers And Education, 247. Https://Doi.Org/10.1016/J.Compedu.2026.105564

Kassanew, B., Weng, C., Astatke, M., & Kassanew, Y. (2026). The Impact Of Ai-Supported Think-Pair-Share Instruction On Students' Problem-Solving Skills And Motivation In Learning Quadratic Equations: A Social Constructivist Perspective. Journal Of Computer Assisted Learning, 42(4). Https://Doi.Org/10.1002/Jcal.70271

Khaleghi, E., & Rayner, R. (2025). Revolutionising Education Through Digital Twins: Case Study And Applications. Mechatronic Futures: Further Challenges And Solutions For Mechatronic Systems And Their Designers, 393-414. Https://Doi.Org/10.1007/978-3-031-83571-1_21

Lee, J. (2025). Empowering Students Through Generative Ai And Cultural Diversity In Interdisciplinary Fashion Design Education. Research Journal Of Textile And Apparel. Https://Doi.Org/10.1108/Rjta-02-2025-0017

Lin, D., Shuang, X., & Zongben, X. (2025). Teacher-Student-Ai Collaborative Classroom: The Platform And Practice Of Ai-Empowered University Mathematics Education. Frontiers Of Education In China, 20(4), 477-489. Https://Doi.Org/10.3868/S110-020-025-0025-8

Lingling, W., Shijie, S., & Xueyi, L. (2026). From Automation To Augmentation: A Retrieval-Augmented Large Language Model Framework For Reshaping Learning Experiences In Higher Education. Frontiers In Psychology, 17. Https://Doi.Org/10.3389/Fpsyg.2026.1870870

Listyaningrum, P., Retnawati, H., Harun, & Ibda, H. (2024). Digital Learning Using Chatgpt In Elementary School Mathematics Learning: A Systematic Literature Review. Indonesian Journal Of Electrical Engineering And Computer Science, 36(3), 1701-1710. Https://Doi.Org/10.11591/Ijeecs.V36.I3.Pp1701-1710

López-Goyez, J. P., González-Briones, A., & Demazeau, Y. (2026). An Adaptive Multi-Agent Architecture With Reinforcement Learning And Generative Ai For Intelligent Tutoring Systems: A Moodle-Based Case Study. Applied Sciences (Switzerland), 16(3). Https://Doi.Org/10.3390/App16031323

Luo, X., Wang, S., & Hew, K. F. (2026). Equipping Elementary School Students With Self-Regulated Learning Through Human-Ai Collaboration In Online Learning. Computers And Education, 253. Https://Doi.Org/10.1016/J.Compedu.2026.105697

Maalek, R. (2024). Integrating Generative Artificial Intelligence And Problem-Based Learning Into The Digitization In Construction Curriculum. Buildings, 14(11). Https://Doi.Org/10.3390/Buildings14113642

Martinez, J. A. (2025). The New Era Of Artificial Intelligence: Empowering Coders, Non-Coders, And Shaping Future Workforce Creativity. Ai, Personalization, Equity, And The Future Of Learning, 139-158. Https://Doi.Org/10.4018/979-8-3373-5097-4.Ch005

Marzano, D. (2025). Generative Artificial Intelligence (Gai) In Teaching And Learning Processes At The K-12 Level: A Systematic Review. Technology, Knowledge And Learning, 31, 789–829. Https://Doi.Org/10.1007/S10758-025-09853-7

Otto, S., Ejsing-Duun, S., & Lindsay, E. (2025a). Disruptive Tensions And Emerging Practices: An Exploratory Inquiry Into Student Perspectives On Generative Artificial Intelligence In A

Problem-Based Learning Environment. Education And Information Technologies, 30(13), 19111-19140. Https://Doi.Org/10.1007/S10639-025-13533-5

Otto, S., Lavi, R., & Brogaard Bertel, L. (2025b). Human-Genai Interaction For Active Learning In Stem Education: State-Of-The-Art And Future Directions. Computers And Education, 239. Https://Doi.Org/10.1016/J.Compedu.2025.105444

Oviatt, S., & Cohen, A. (2014). Written Activity, Representations And Fluency As Predictors Of Domain Expertise In Mathematics. Icmi 2014 - Proceedings Of The 2014 International Conference On Multimodal Interaction, 10-17. Https://Doi.Org/10.1145/2663204.2663245

Oviatt, S., Hang, K., Zhou, J., & Chen, F. (2015). Spoken Interruptions Signal Productive Problem Solving And Domain Expertise In Mathematics. Icmi 2015 - Proceedings Of The 2015 Acm International Conference On Multimodal Interaction, 311-318. Https://Doi.Org/10.1145/2818346.2820743

Pirker, J., Riffnaller-Schiefer, M., & Gütl, C. (2014). Motivational Active Learning - Engaging University Students In Computer Science Education. Iticse 2014 - Proceedings Of The 2014 Innovation And Technology In Computer Science Education Conference, 297-302. Https://Doi.Org/10.1145/2591708.2591750

Renjewski, D., & Strasser, A. (2026). Designing Ai-Resilient Exams: Error-Correction Assignments In Engineering Education. Human-Centric Intelligent Systems. Https://Doi.Org/10.1007/S44230-026-00157-2

Shi, K., Yang, Y., Liang, T., Zhang, L., Zhao, Z., & Hu, C. (2026). Umath: Closing The Loop Between Student Ai Tutoring And Teacher Insight In Mathematics. Ieee Global Engineering Education Conference, Educon. Https://Doi.Org/10.1109/Educon67543.2026.11574145

Stamenkova, R. (2025). Large Language Models: A Tool For Solving Mathematical Problems In High School. Annual Of Sofia University "St. Kliment Ohridski" Faculty Of Mathematics And Informatics, 112, 165-183. Https://Doi.Org/10.60063/Gsu.Fmi.112.165-183

Tassilova, N. A., Shashayeva, G. K., Yernazarova, Z. S. H., Akhmetova, A. I., Arynov, Z. M., & Karimova, Z. K. (2025). Communication Technologies In Stem Education: A Systematic Review. Eurasia Journal Of Mathematics, Science And Technology Education, 21(10). Https://Doi.Org/10.29333/Ejmste/17178

Tu, Y. F., Guo, X. G., Fang, J. W., Meng, X. P., & Hwang, G. J. (2026). Roles And Impacts Of Generative Ai Agents In Steam Learning: Insights From Students' Perceptions And Learning Experiences Using Information World Mapping. Educational Technology Research And Development, 74(2), 897-935. Https://Doi.Org/10.1007/S11423-025-10579-7

Turmuzi, M., Azmi, S., & Kertiyani, N. M. I. (2026). Chatgpt In School Mathematics Education: A Systematic Review Of Opportunities, Challenges, And Pedagogical Implications. Teaching And Teacher Education, 170, 105286. Https://Doi.Org/10.1016/J.Tate.2025.105286

Wong, K., Bulathwela, S., & Cukurova, M. (2026). Towards An Adaptive Ai Scaffold For Developing Student Collaborative Problem Solving. Learning And Instruction, 105. Https://Doi.Org/10.1016/J.Learninstruc.2026.102418

Yildiz, S. G., & Körpeoğlu, S. G. (2025). Trends And Insights Of Ai In Mathematics Education: A Bibliometric Analysis. International Electronic Journal Of Mathematics Education, 20(3). Https://Doi.Org/10.29333/Iejme/16401

Zhang, D., Wijaya, T. T., Wang, Y., Su, M., Li, X., & Damayanti, N. W. (2025). Exploring The Relationship Between Ai Literacy, Ai Trust, Ai Dependency, And 21st Century Skills In Preservice Mathematics Teachers. Scientific Reports, 15(1). Https://Doi.Org/10.1038/S41598-025-99127-0

Zhang, H., Xiao, X., & Zhang, J. (2026a). The Impact Of Artificial Intelligence Literacy On Students' Social-Emotional Competence: An Empirical Study Using Random Forest Feature Importance Modeling. Proceedings Of 2025 3rd International Conference On Information Education And Artificial Intelligence, Icieai 2025, 546-551. Https://Doi.Org/10.1145/3799457.3799547

Zhang, M., Raytchev, B., & Sun, X. (2026b). Realm: A Relevance-Driven Layered Protocol For Human–Ai Collaboration In Stem Education. Ieee Signal Processing Magazine, 43(1), 56-68. Https://Doi.Org/10.1109/Msp.2025.3599339