INPLASY

Can GenAI improve students' mathematics learning outcomes effectively ?—A meta-analysis of empirical research 2023-2025.

INPLASY2025110051

doi: 10.37766/inplasy2025.11.0051

Received: 18 November 2025

Published: 18 November 2025

Liu, BX; Zhang, WL; Wang, FF.

Corresponding author:
baoxin liu

13474710182@snnu.edu

Author Affiliation:
Shaanxi Normal University.

ADMINISTRATIVE INFORMATION

Support - This research received no external funding.

Review Stage at time of this submission - The data analysis is complete, and a first draft has been written, but it has not been carefully revised yet.

Conflicts of interest - None declared.

INPLASY registration number: INPLASY2025110051

Amendments - This protocol was registered with the International Platform of Registered Systematic Review and Meta-Analysis Protocols (INPLASY) on 18 November 2025 and was last updated on 18 November 2025.

INTRODUCTION

Review question / Objective Based on the provided manuscript, the review question/ objective formulated using the PICOS framework is as follows:

Objective:

This systematic review and meta-analysis aims to synthesize empirical evidence from 2023 to 2025 to evaluate the effectiveness of Generative Artificial Intelligence (GenAI) in enhancing students' mathematics learning outcomes. It seeks to quantify the overall effect size, compare impacts on cognitive versus non-cognitive skills, and identify key moderating variables influencing this effectiveness.

PICOS Framework:

P (Population): Students (from primary/elementary school to university/higher education) engaged in mathematics learning.

I (Intervention): Implementation of Generative Artificial Intelligence (GenAI), such as ChatGPT or other large language models, within mathematics education contexts. This includes various integration levels defined by the PIC-RAT model (e.g., Creative Transformation, Interactive/Passive Augmentation).

C (Comparison): Traditional teaching methods or non-GenAI instructional approaches used in control groups.

O (Outcomes): Student mathematics learning outcomes, categorized into:

Cognitive Skills: Further divided into higher-order (e.g., analysis, evaluation, creation) and lower-order (e.g., remember, understand, apply) thinking skills based on Bloom's taxonomy.

Non-cognitive Skills: Including factors such as learning motivation, self-efficacy, mathematics anxiety, and attitudes.

The primary quantitative measure is the standardized mean difference (Hedges' g) in these outcomes between intervention and control groups.

S (Study Design): Experimental or quasi-experimental studies published between January 2023 and October 2025.

Specific Research Questions:

What is the overall effect size of Generative Artificial Intelligence (GenAI) on students' mathematics learning outcomes? How does its effect differ between cognitive skills and non-cognitive skills?

Which variables (e.g., educational level, intervention duration, learning content, GenAI integration level, sample size) significantly moderate the effectiveness of GenAI on these outcomes?

By applying this PICOS structure, the review precisely defines its scope, ensuring a focused investigation into the impact of GenAI on mathematics education, the specific outcomes of interest, the context for comparison, and the types of evidence considered relevant for synthesis.

Rationale The integration of Generative Artificial Intelligence (GenAI) into education represents a paradigm shift, offering unprecedented capabilities for personalized and interactive learning. Mathematics, as a foundational discipline critical for innovation in fields like data science and engineering, is a primary domain where GenAI's potential is both promising and contested. The advent of powerful GenAI models like ChatGPT has accelerated the transition of mathematics education from a focus on knowledge transmission towards fostering critical thinking, creativity, and self-regulated learning. GenAI, with its superior interactivity, adaptability, and content-generation abilities compared to traditional AI, appears uniquely positioned to support this pedagogical transformation.

However, the empirical landscape investigating the actual efficacy of GenAI in mathematics education is fragmented and marked by inconsistent findings. On one hand, a body of research highlights GenAI's unique advantages in problem-solving, providing personalized feedback, and adapting tasks, thereby supporting the development of mathematical knowledge and skills. Positive outcomes reported include deepened conceptual understanding, reduced cognitive load, lowered mathematics anxiety, and enhanced self-efficacy and engagement across various educational levels. On the other hand, significant concerns and negative findings persist. Some studies warn that GenAI can supplant essential cognitive activities, potentially leading to negative long-term effects on mathematical learning ability. Instances have been reported where GenAI intervention groups underperformed compared to traditional instruction, particularly when the technology provides complete answers, potentially stifling students' opportunities for questioning and reflection. Further risks include the limitation of deep thinking on complex problems, the generation of incorrect mathematical explanations, and the potential for declining confidence and rising technical anxiety with prolonged use.

This state of contradictory evidence creates a critical gap in understanding. While individual studies offer valuable insights, their divergent conclusions make it difficult for educators, policymakers, and researchers to draw definitive conclusions about the overall impact and practical value of GenAI in mathematics learning. Existing meta-analyses have begun to synthesize evidence on GenAI's effects, but they are predominantly broad in scope, focusing on general academic performance or spanning multiple disciplines. This lack of a specialized synthesis for mathematics is a significant limitation, as the pedagogical needs and application of GenAI in this specific subject are distinct. Furthermore, prior syntheses often lack a comprehensive framework that simultaneously considers both cognitive skills (e.g., problem-solving, understanding) and non-cognitive skills (e.g., anxiety, motivation), which are both crucial for holistic mathematical proficiency.

Most notably, a key factor hypothesized to influence GenAI's effectiveness—its degree of integration into pedagogy—has not been systematically examined as a moderating variable. The PIC-RAT model provides a robust framework for classifying this integration based on student interaction and pedagogical transformation. Preliminary evidence suggests that higher integration levels, such as "Creative

Transformation," may lead to better outcomes, but this has not been rigorously tested through a meta-analytic approach.

Therefore, this study is motivated by the pressing need to provide a definitive, quantitative synthesis of the emerging evidence on GenAI in mathematics education. Its rationale is to move beyond isolated and conflicting studies by conducting a focused meta-analysis that: (1) quantifies the overall effect of GenAI on mathematics learning outcomes; (2) separately analyzes its impact on cognitive and non-cognitive skills to provide a holistic view; and (3) investigates key moderating variables—including the crucially under-researched variable of integration level via the PIC-RAT model—to identify the conditions under which GenAI is most effective. By doing so, this research aims to consolidate empirical evidence, resolve inconsistencies, and offer evidence-based guidance to inform educational practice, future research, and technological development.

Condition being studied The core condition under investigation in this study is students' learning outcomes in mathematics, specifically the changes observed following the integration of Generative Artificial Intelligence (GenAI). Mathematics, as a fundamental discipline, is seeing a shift in its educational objectives, moving from traditional knowledge transmission towards an emphasis on cultivating critical thinking, creativity, and self-regulated learning capabilities. However, mathematics learning is inherently abstract and logical, posing challenges for many students. These challenges manifest as difficulties in deep conceptual understanding, insufficient problem-solving abilities, and are often accompanied by non-cognitive issues such as mathematics anxiety and low motivation.

Traditional teaching methods can sometimes struggle to meet the individualized learning needs of all students, particularly regarding the provision of immediate feedback and adaptive support. Generative AI, as an emerging technology, is considered a potential tool to address these challenges due to its robust capabilities in interactivity, content generation, and personalized feedback. It can offer personalized tutoring through conversational interactions, concretize abstract concepts through multimodal presentations, and reduce students' cognitive load via instant feedback.

Therefore, this study does not focus on a "disease," but rather investigates an educational intervention context: the impact of integrating GenAI as a pedagogical tool or learning environment into mathematics instruction. The research concentrates on the effect of this intervention on the broad condition of "mathematics learning outcomes," which is systematically categorized into two dimensions:

Cognitive Skills: These include lower-order cognitive skills (e.g., remembering, understanding, applying mathematical facts and procedures) and higher-order cognitive skills (e.g., analyzing, evaluating, creating, involving complex problem-solving and reasoning), based on Bloom's taxonomy.

Non-cognitive Skills: These encompass the affective, motivational, and belief-related factors associated with mathematics learning, such as mathematics anxiety, self-efficacy, learning motivation, and academic engagement.

This study aims to systematically evaluate the overall effect of GenAI on this core condition (mathematics learning outcomes) and to thoroughly investigate how factors such as educational level, learning content, intervention duration, and the degree of technological integration moderate its effectiveness.

METHODS

Search strategy

Optimized English Version
Literature Search Strategy

To ensure a comprehensive and systematic literature retrieval, this meta-analysis strictly adhered to the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines. The search strategy was meticulously designed to balance sensitivity (retrieving all relevant studies) and specificity (minimizing irrelevant results).

1. Electronic Databases
To cover a broad spectrum of published and grey literature, the following electronic databases and platforms were systematically searched:

Web of Science Core Collection: For high-quality, international peer-reviewed journals.

EBSCOhost: Searching the Academic Search Complete, ERIC, and PsycINFO databases simultaneously.

China National Knowledge Infrastructure (CNKI): To ensure the inclusion of significant research published in Chinese.

Google Scholar: To capture additional grey literature and publications not indexed in the primary databases.

2. Search Terms and Syntax
The search strategy was developed based on the core elements of the PICO framework, comprising three sets of keywords related to "Intervention (GenAI)", "Context (Mathematics)", and "Outcome (Learning Results)". Terms within each set were combined with "OR", and the three sets were combined using "AND".

The core search string, adapted for syntax compatibility across databases, was as follows:
("generative AI" OR "generative artificial intelligence" OR ChatGPT OR "Gen-AI" OR "large language model*" OR "AI-powered" OR "AI-driven") AND (math* OR mathematic* OR algebra OR geometry OR calculus OR statistics OR "problem-solving") AND (learn* OR performance OR achievement OR outcome* OR anxiety OR attitude* OR motivation OR "computational thinking" OR skill*) AND (student* OR pupil* OR learner* OR "elementary school" OR "primary school" OR "middle school" OR "high school" OR "undergraduate" OR "higher education")

Truncation symbols (e.g., *) were used to capture variations in word endings. Searches were primarily conducted on titles, abstracts, and keywords.

3. Search Timeframe
Considering that Generative AI (notably ChatGPT) gained widespread public attention and educational application in late 2022, the search timeframe was set from January 1, 2023, to October 31, 2025. This period aligns with the emergence of the first wave of empirical studies following the technology's maturation.

4. Supplementary Search Strategies
To further enhance comprehensiveness and mitigate publication bias, the following supplementary searches were performed:

Manual screening of the reference lists of all included studies and relevant prior review articles.

Forward citation tracking for key included studies using Google Scholar.

Searching for publications by leading authors identified during the initial screening process.

5. Search Execution and Yield
The database searches were execut**.**

Participant or population Patients, Participants, or Population
The participants involved in this review are students from various educational stages, including:

Primary school students

Secondary school students (including both middle and high school students)

University students (undergraduates)

These students participated in empirical studies that investigated the effects of generative artificial intelligence (GenAI) on mathematics learning outcomes between 2023 and 2025.

Intervention Intervention:
The intervention examined in this systematic review and meta-analysis is the use of generative artificial intelligence in instructional settings to support mathematics learning.

Intervention Group: Students engaged in mathematics learning through the use of generative AI tools. These tools, typically based on large language models, provided supports such as personalized tutoring, immediate feedback, step-by-step problem-solving guidance, conceptual explanations, and multimodal content generation. Specific forms of intervention included, but were not limited to:

Interactive dialogues with GenAI to solve mathematical problems or explore concepts.

Receiving personalized practice problems or learning materials generated by GenAI.

Using GenAI for mathematical proof, reasoning, or visualization.

Iterating on their learning based on feedback provided by the GenAI.
Furthermore, the level of technology integration was coded based on the PIC-RAT model, primarily categorized as "Creative Transformation" or "Interactive/Passive Enhancement".

Control Group: Students received traditional mathematics instruction without the use of

generative AI. This typically involved conventional methods such as teacher-led lectures, use of standard textbooks, completion of paper-based exercises, and receiving feedback directly from the teacher.

Comparator Comparator:
The comparator intervention applied to the target population in this systematic review and meta-analysis is traditional mathematics teaching methods.

Specifically, students in the control group did not receive any generative AI-based learning support. Instead, they received conventional mathematics instruction that was not integrated with GenAI, typically including:

Teacher-led classroom lectures

Use of standard textbooks and paper-based exercises

Receiving human (non-AI) feedback from teachers or peers

Conventional classroom discussions and assignments.

Study designs to be included In order to systematically evaluate the impact of Generative Artificial Intelligence (GenAI) on students' mathematics learning outcomes and explore the role of relevant moderating variables, this meta-analysis strictly defined the types of empirical research designs to be included. The primary research designs incorporated were Randomized Controlled Trials (RCTs) and Quasi-Experimental Designs, particularly those featuring comparisons between an experimental group (receiving GenAI intervention) and a control group (receiving traditional teaching methods).

Eligibility criteria English
In addition to the inclusion criteria defined within the PICOS (Participants, Interventions, Comparators, Outcomes, Study design) framework, this meta-analysis applied the following additional eligibility criteria to ensure the quality, relevance, and data availability of the included studies:

Study Quality: Studies had to meet a minimum methodological quality threshold. Using the Medical Education Research Study Quality Instrument (MERSQI), only studies with a MERSQI score ≥ 10.5 were included to safeguard the robustness of the meta-analytic findings.

Data Availability: Studies must have reported complete effect size data (e.g., means, standard deviations, sample sizes) or provided sufficient information to allow for the calculation of an effect size (e.g., t-values, F-values, chi-square values along with their corresponding p-values and degrees of freedom). Studies failing to provide extractable or convertible data were excluded.

Publication Status and Language: The initial search did not restrict publication status (e.g., journal articles, preprints, theses, conference proceedings). However, for final inclusion, non-peer-reviewed research (such as some preprints or reports) required confirmation through quality assessment and bias analysis to ensure methodological rigor and no significant bias introduced to the overall results. The search was limited to Chinese and English publications.

Specificity of GenAI Tool: The generative AI tool used in the study had to be explicitly specified (e.g., ChatGPT, a specific large language model). Studies that only vaguely referred to "AI" or "intelligent technology" without clearly identifying it as generative AI were excluded.

Clarity of Mathematics Learning Outcomes: The learning outcomes assessed in the studies must be explicitly related to the mathematics discipline and could be clearly classified into one or more of the following dimensions: cognitive skills (higher-order/lower-order) or non-cognitive skills (e.g., anxiety, motivation, self-efficacy).

These additional criteria were implemented to enhance comparability across studies and to ensure that the meta-analysis was based on high-quality, synthesizable evidence.

Information sources To ensure the comprehensiveness of the systematic review and minimize publication bias, this study conducted a systematic literature search through the following complementary sources:

Systematic Electronic Database and Platform Searches
We searched multiple authoritative databases and academic platforms covering both Chinese and English literature, including:

English Databases: Web of Science (WoS), EBSCOhost.

Chinese Database: China National Knowledge Infrastructure (CNKI).

Academic Search Engine: Google Scholar.

Google Scholar was specifically included in the search strategy due to its effectiveness in indexing "grey literature" such as preprints, dissertations, and conference papers. This is crucial for capturing the latest empirical research in the rapidly evolving field of Generative AI, helping to mitigate publication bias that can arise from relying solely on published literature.

Reference Tracing
To minimize the omission of relevant studies, we implemented rigorous manual reference tracing, including:

Backward Tracing: Systematic examination of the reference lists of all initially eligible studies.

Key Literature Tracing: Review of references from high-impact review articles or meta-analyses closely related to the topic.

This tracing procedure is a recommended standard method in systematic reviews, enhancing search comprehensiveness and the methodological rigor of the study.

Handling of Non-Peer-Reviewed Literature
During the screening process, we identified one research report by Nakavachara et al. (2024) that had not yet undergone peer review. To uphold methodological rigor, we independently assessed its quality using the Medical Education Research Study Quality Instrument (MERSQI), which yielded a score of 13, meeting the high-quality literature threshold (≥10.5). Further publication bias tests (Egger's test, *p* = .107) and sensitivity analyses (leave-one-out analysis) confirmed that this study did not exert an undue influence on the overall results. Based on these considerations, it was included in the analysis to reflect the most current research trends, while ensuring the robustness and transparency of the conclusions.

Through this multi-channel, systematic search strategy and the careful handling of non-traditional literature, this study not only ensures methodological rigor but also provides a more comprehensive understanding of the actual impact of Generative AI in mathematics education, representing a significant strength of this research.

Main outcome(s) This study conducted a systematic review and meta-analysis of 22 empirical studies (46 independent effect sizes, N=5,132) from 2023 to 2025 to rigorously evaluate the impact of Generative AI (GenAI) on students' mathematics learning outcomes. The main findings are as follows:

Overall Effect: The random-effects model revealed a moderate positive overall effect of GenAI on mathematics learning outcomes (*g* = 0.539, *p* < 0.005), confirming its effectiveness as an auxiliary tool in mathematics education.

Breakdown by Outcome Type: GenAI demonstrated a significantly stronger promoting effect on cognitive skills (*g* = 0.596) than on non-cognitive skills (*g* = 0.320). Within cognitive skills, its positive impact on higher-order thinking (e.g., analysis, creation, *g* = 0.740) was particularly prominent and significantly greater than on lower-order thinking (e.g., memory, understanding, *g* = 0.562).

Moderator Analysis (Subgroup Analysis): The analysis identified grade level, GenAI integration degree, and sample size as significant moderators. Specifically:

Lower-grade students (elementary school, *g* = 0.800) benefited the most, with effects diminishing at higher grades (middle school *g* = 0.574, university *g* = 0.283).

Integration characterized as "Creative Transformation" yielded the optimal effect (*g* = 1.075), significantly outperforming the "Interactive/ Passive Enhancement" mode (*g* = 0.400).

Studies with smaller sample sizes reported significantly larger effect sizes (*g* = 0.749) than those with larger samples (*g* = 0.366), highlighting the importance of personalized intervention.

Intervention duration and learning content did not show significant moderating effects.

The highlights of this study include: being the first focused meta-analysis on GenAI's effect within the mathematics discipline; constructing a comprehensive evaluation framework encompassing both cognitive and non-cognitive skills; and innovatively examining the moderating role of integration degree based on the PIC-RAT model, providing refined empirical evidence for both theory and practice.

Additional outcome(s) Beyond the primary findings, a key methodological decision and highlight of this meta-analysis was the selection of a two-level random-effects model over a three-level model for data analysis, based on model fit statistics.

Although multiple effect sizes were extracted, potentially creating a hierarchical structure, the initial three-level model test indicated 0% heterogeneity at Level 2 (within studies). Further model comparison (see Table 3) showed that the two-level model had lower AIC and BIC values, and the likelihood ratio test was non-significant (p = 1.0000). This collective statistical evidence demonstrated that the three-level model did not provide a superior fit. The simplified two-level model was sufficient to accurately describe the data structure while adhering to the principle of parsimony.

Consequently, the 46 effect sizes were treated as independent, and the two-level random-effects model was adopted for final analysis. This rigorous methodological choice ensures the robustness of the subsequent effect size aggregation and moderator analyses.

Table 3 Model Fitting Comparison Results

Model df AIC BIC AICc logLik LRT p-value QE
Three-Level 3 94.98 100.40 95.57 -44.49 90.91
Two-Level 2 92.98 96.60 93.27 -44.49 0.0000 1.0000 90.91.

Data management To ensure the systematic, transparent, and reproducible processes of literature screening and data extraction, this study implemented a structured data management workflow utilizing a combination of specialized tools.

During the literature retrieval and record-keeping phase, all initially identified records from databases including Web of Science, EBSCO, CNKI, and Google Scholar were uniformly managed using Zotero reference management software. Its automatic deduplication and online synchronization features effectively prevented duplicate entries and ensured the completeness of the literature sources.

In the literature screening phase, we utilized the online collaborative screening platform Rayyan. Two researchers independently performed blinded screening of titles and abstracts, marking and discussing discrepant items until consensus was reached. This system efficiently supported the rapid and transparent screening of a large volume of literature.

For data extraction and coding, a structured coding form was designed using Microsoft Excel, encompassing basic document information and study characteristics. Two coders worked independently according to a pre-defined coding protocol. Discrepancies in coding were resolved through discussion, and inter-coder reliability for key variables (e.g., moderator classifications) was calculated (Cohen's Kappa = 0.918), ensuring the accuracy and reliability of the extracted data.

All intermediate data, screening decisions, and final coding results for included studies were archived and backed up in Excel, creating a complete and traceable data audit trail. This integrated data management mechanism, combining Zotero, Rayyan, and Excel, not only improved workflow efficiency but also significantly enhanced the methodological rigor and the credibility of the findings through multiple rounds of verification and consistency checks, representing a key methodological strength of this study.

Quality assessment / Risk of bias analysis In terms of quality assessment, this study employed the Medical Education Research Study Quality Instrument (MERSQI) to systematically evaluate the 22 included empirical studies. The MERSQI tool consists of 10 items covering six key dimensions, such as study design, data collection methods, data analysis, and outcome reporting, and has demonstrated good reliability and validity in assessing quantitative research in educational settings. Two coders independently scored each study using MERSQI, and any discrepancies in ratings were resolved through in-depth discussion until consensus was reached, ensuring rigor and consistency in the assessment process. All included studies had MERSQI scores ≥10.5, indicating high overall literature quality and providing a reliable foundation for the meta-analysis.

For bias risk analysis, multiple methods were used to examine publication bias. The funnel plot showed a largely symmetric distribution of effect sizes; Egger's regression test yielded p = 0.10676 (> 0.05), indicating no significant publication bias; and the fail-safe N of 3030 far exceeded the 5k + 10 (k = 22) criterion, further supporting a low likelihood of bias influencing the results. Additionally, sensitivity analysis (leave-one-out method) revealed that the overall effect size remained stable after excluding any single study (z = 6.397, p < 0.005; 95% CI = [0.374, 0.704]), confirming the robustness of the pooled results.

The rigor of this study is reflected in: strict adherence to PRISMA guidelines for literature search and screening, use of standardized tools for quality assessment, multi-method validation of

bias risks, and high consistency in data coding (Cohen's Kappa = 0.918). These measures comprehensively ensure the reproducibility and credibility of the study, highlighting methodological strengths.

Strategy of data synthesis

Data Synthesis Strategy

To ensure the rigor of the meta-analysis and the reproducibility of its conclusions, this study employed a systematic and transparent data synthesis strategy. Initially, raw data (e.g., means, standard deviations, sample sizes) extracted from the included studies were used to calculate standardized effect sizes. We consistently used Hedges's *g* as the effect size metric, which corrects for small sample bias, thereby providing a more accurate estimate of the overall effect. The calculated effect sizes were interpreted against Cohen's (2009) benchmarks, where *g* ≈ 0.2, 0.5, and 0.8 represent small, medium, and large effect sizes, respectively.

Data analysis was performed using CMA (Comprehensive Meta-Analysis) Version 3.0. Anticipating heterogeneity among the studies due to variations in participants, interventions, and contexts, a random-effects model was employed for pooling effect sizes. This model allows for the possibility that the true effect size varies across studies, making the conclusions more generalizable. To validate the choice of model, we conducted model fit comparisons. An initial three-level model check revealed that the level-3 variance component (within-study, between-effect sizes) was dominant (84.18%), while the level-2 variance (between studies) was 0%. Further comparison of model fit indices (e.g., AIC, BIC) and a non-significant likelihood ratio test (LRT, *p* > 0.05) supported the use of a more parsimonious two-level random-effects model, treating each effect size as an independent data point.

Before pooling the overall effect size, we confirmed the presence of high heterogeneity among the studies via a significant Q-statistic (*p* < .05) and an I^2 statistic of 87.442%. This justified conducting moderator analyses to explore sources of this heterogeneity. Several moderators were pre-specified for subgroup analysis, including intervention duration, educational stage, learning content, GenAI integration level (coded based on the PIC-RAT model), and sample size. For categorical moderators, between-group Q-tests were used to examine if effect size differences across subgroups were statistically significant.

To assess the robustness of the findings, comprehensive publication bias tests and sensitivity analyses were implemented. Publication bias was evaluated using a combination of three methods: visual inspection of the funnel plot symmetry, Egger's regression intercept test (*p* = 0.10676), and calculation of the fail-safe N (N = 3030). The results collectively suggested no significant publication bias. Sensitivity analysis was performed using the 'leave-one-out' method, recalculating the overall effect size after sequentially removing each study. The results showed no substantial fluctuations in the pooled effect size, demonstrating that the findings were stable and not driven by any single study.

The entire data synthesis process, from literature screening and data coding (inter-coder reliability Cohen's Kappa > 0.7) to statistical analysis, strictly adhered to the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines. All decision steps were d.

Subgroup analysis To thoroughly investigate the sources of heterogeneity, this study conducted subgroup analyses on several pre-specified moderator variables. All analyses were performed based on a pre-established coding protocol using a random-effects model. The results identified the integration level of GenAI as a significant moderator: the effect size was large when integrated as "Creative Transformation" (*g* = 1.075), significantly surpassing the effect observed under the "Interactive/Passive Augmentation" mode (*g* = 0.400). Educational stage was another significant moderator, with the largest effect size found at the elementary school level (*g* = 0.800), compared to secondary (*g* = 0.574) and university levels (*g* = 0.283). Sample size also played a significant role, with smaller sample studies yielding a markedly higher effect size (*g* = 0.749) than larger sample studies (*g* = 0.366).

The methodological rigor of this subgroup analysis is highlighted by two key aspects. First, all variable categories were explicitly defined prior to data extraction, preventing post-hoc hypotheses and ensuring the reproducibility of the analytical process. Second, the "GenAI Integration Level" was innovatively conceptualized and operationalized according to the PIC-RAT model, providing a solid theoretical lens and practical implications for understanding the efficacy differences across application modes. The entire analytical procedure strictly adhered to meta-analytic standards, thereby underpinning the robustness and reliability of the research conclusions.

Sensitivity analysis To assess whether the meta-analytic pooled result was overly dependent on any single study and to test the robustness of the overall effect, this study employed the "leave-one-out" method for sensitivity analysis. This procedure involved iteratively removing each of the 22 included studies and recalculating the overall effect size of Generative AI (GenAI) on mathematics learning outcomes based on the remaining 21 studies.

The results indicated that the recalculated overall effect sizes (Hedges' g) ranged from 0.502 to 0.566 after the sequential exclusion of each study. All recalculated effect sizes remained within the original overall confidence interval of [0.374, 0.704], and their statistical significance remained unchanged (all p-values < 0.005). This finding demonstrates that the pooled effect size (g = 0.539) obtained in this meta-analysis is highly stable and not disproportionately influenced or driven by any single primary study.

The successful application of the "leave-one-out" sensitivity analysis strengthens the reliability and robustness of the primary conclusion—that GenAI has a moderately positive impact on students' mathematics learning outcomes. The process is transparent and employs a standard methodology, ensuring the replicability of the findings. This means that other researchers following the same procedure should arrive at consistent conclusions, thereby significantly enhancing the methodological rigor and persuasiveness of this meta-analysis.

Country(ies) involved China.

Other relevant information I. Research Background and Contemporary Significance
Generative Artificial Intelligence (GenAI) rapidly entered the educational landscape starting in late 2022. Its capabilities for multimodal interaction, personalized feedback, and content generation offer new possibilities for the transformation of mathematics education. This meta-analysis specifically focuses on the period from 2023 to 2025, capturing the critical phase of GenAI's transition from technological emergence to practical application in education. During this time, educational policies and ethical guidelines gradually took shape worldwide, and related empirical research experienced explosive growth. Conducting this meta-analysis within this timeframe ensures strong timeliness and clear practical orientation, providing systematic evidence for understanding the initial effectiveness of GenAI in mathematics education.

II. Methodological Strengths and Innovations
Systematic Construction of Moderator Variables:
This study moves beyond simply asking "is it effective" to deeply investigate moderating factors such as integration degree (based on the PIC-RAT model), type of learning content, and grade level. This provides granular insights for subsequent instructional design and technological optimization.

Holistic Analytical Framework:
Learning outcomes are clearly distinguished between cognitive skills (higher-order/lower-order) and non-cognitive skills (motivation, anxiety, self-efficacy, etc.), with their effects evaluated separately. This addresses a common limitation in previous research that overemphasized academic achievement while neglecting affective and attitudinal dimensions.

Rigorous Literature Screening and Quality Assessment:
The use of the MERSQI tool for quality assessment and the inclusion of unpublished yet high-quality research not only ensure the reliability of the evidence but also demonstrate inclusivity towards the evolving research ecosystem in this emerging field.

III. Extended Practical Implications of the Findings
Promotional Value of the "Creative Transformation" Integration Model:
The finding that GenAI is most effective when integrated as a "Creative Transformation" (g = 1.075) suggests that educators should move beyond a "tool replacement" mindset. Instead, GenAI should be leveraged as a cognitive partner to stimulate student inquiry, support project-based learning, and facilitate interdisciplinary integration.

Implications from Small-Sample Studies:
The larger effect sizes found in small-sample studies indicate that GenAI shows greater advantages in small-class, personalized teaching environments. This offers valuable insights for resource allocation and technology deployment strategies, particularly in resource-limited areas or special education contexts.

Deeper Interpretation of Grade-Level Differences:
The strongest effects in primary school may be due to the high compatibility between GenAI's visual, interactive, and gamified features and children's cognitive characteristics. The weaker effects at the university level suggest that GenAI's support capabilities for advanced mathematical content (e.g., proofs, modeling) still require further enhancement.

IV. Supplementary Research Limitations and Future Directions

Limitations in Literature Sources and Language:
Relying primarily on Chinese and English databases might have led to the omission of research in other languages or regions. Future work could expand to include multilingual literature to enhance the cross-cultural representativeness of the conclusions.

"Timeliness Attenuation" Due to Rapid Technological Iteration:
GenAI technology is evolving extremely rapidly. Models and educational applications post-2025 may have already surpassed the scope of the current studies. It is recommended to conduct updated analyses every 1–2 years to track the impact of technological advancements on educational outcomes.

Insufficient Exploration of Teacher Role and Training Mechanisms:
While the "irreplaceability" of teachers is mentioned, the relationship between teacher GenAI literacy, training support, and teaching effectiveness is not systematically analyzed. Future research could include qualitative or mixed-methods studies focusing on teacher cognition, acceptance, and professional development.

Inadequate Discussion of Ethical and Equity Issues:
Issues such as the access barrier to GenAI, data privacy, and resource accessibility, while not the core focus of this study, are critical factors influencing its widespread adoption in education. It is recommended to emphasize these aspects in the policy recommendations section.

V. Recommendations for Policy and the Educational Ecosystem

Promote a "AI + Teacher" collaborative teaching model, clarifying the teacher's irreplaceable role in guidance, supervision, and emotional support.

Strengthen the development of standards and evaluation mechanisms for GenAI educational products to ensure their scientific validity, educational value, and safety.

Encourage longitudinal tracking studies, particularly on the long-term impact of GenAI on students' mathematical thinking development and innovation capabilities.

Keywords Mathematics, Generative Artificial Intelligence, Learning Outcomes, Meta-Analysis.

Dissemination plans The research, "*Can Generative Artificial Intelligence Effectively Enhance Students' Mathematics Learning Outcomes?——A Meta-Analysis of Empirical Studies from 2023-2025*," is now complete. The following pragmatic dissemination plan is proposed to foster academic exchange.

This study was rigorously conducted following PRISMA guidelines, emphasizing methodological transparency to ensure reproducibility. We have provided a detailed account of the comprehensive search strategy, explicit inclusion/exclusion criteria, systematic coding procedures (Kappa = 0.918), and meta-analytic techniques using a two-level model, thereby guaranteeing the robustness of the findings.

Dissemination will focus on the following accessible channels for a doctoral student:

Academic Conferences: Prioritizing presentation (poster or parallel session) at domestic and international conferences in educational technology, mathematics education, or meta-analysis to disseminate initial findings, gather immediate feedback, and build academic networks.

Preprint Servers: Submitting the manuscript to platforms like arXiv, EdAriv, or their Chinese counterparts prior to formal journal submission to establish precedence and solicit broad peer commentary.

Academic Social Media: Sharing the abstract and key findings on platforms like ResearchGate or within relevant academic online communities to cost-effectively increase visibility among specialized peers.

Peer Collaboration Networks: Circulating the study among interested scholars via supervisor referrals and conference interactions to explore potential opportunities for future collaboration and research development.

This plan aims to leverage the solid methodological foundation to enhance the academic visibility and impact of this research through feasible and targeted channels.

Contributions of each author

Author 1 - baoxin liu - Led research design, data analysis, and manuscript drafting; coordinated coding and model validation to ensure methodological rigor andreproducibility.
Email: 1241083586@qq.com

Author 2 - wenlan zhang - Guided research framework and model selection; reviewed methodological quality, result interpretation, and academic standards to oversee the overall direction.
Email: wenlan19@163.com
Author 3 - fangfang wang - Assisted in literature retrieval, screening, and data coding; created tables/figures and formatted references, supporting data organization and workflow execution.
Email: iwishthatyou@163.com