

Exploring the Integration of Generative AI and Game-Based Learning in Mathematics Education: A Narrative Literature Review

Martina Ragusa

Institute for Educational Technology (ITD) of the
Italian National Research Council (CNR)

Flavio Manganello

Institute for Educational Technology (ITD) of the
Italian National Research Council (CNR)

Abstract

This narrative literature review synthesises emerging evidence on the integration of generative artificial intelligence, particularly large language models such as ChatGPT, with game-based and gamified approaches in mathematics education. Nine empirical studies published in 2024-2025 were analysed to examine usage modalities, pedagogical affordances, and implications for learner engagement and educational equity. Three modalities of LLM integration emerged from the corpus: authoring tools for content generation, pedagogical agents providing scaffolding, and problem-solving environments enabling interactive simulations. The corpus reports statistically significant reductions in mathematics anxiety in two quasi-experimental studies and increased participation among students who tend to be less engaged in class, alongside persistent challenges that include scaffolding difficulties for students with knowledge gaps (with one study reporting zero task completion among low-performing learners), low authenticity rates in AI-generated problems (6.8% passing all quality criteria in the one study that systematically assessed them), and demographic skew in generated content. From the synthesis, three interpretive frameworks are developed as working hypotheses rather than validated models: a socio-emotional mediation framework positioning psychological safety as a candidate mechanism supporting learning, a dimensional authenticity framework distinguishing dimensions that may carry differential pedagogical weight, and a dual alignment framework articulating architectural directions for educational AI simulation. The review distinguishes access equity from outcomes equity, noting that the available evidence is more compatible with improvements in access than with assured outcomes for students with foundational knowledge gaps. Priority directions for future research include larger studies with adequately powered comparison groups, intervention work on adaptive scaffolding and interaction competencies, systematic bias auditing, and longitudinal designs examining learning persistence and transfer.

Keywords: Generative artificial intelligence, game-based learning, mathematics education

DOI: <https://doi.org/10.4148/edconsiderations.20913>

Introduction

The development of generative artificial intelligence tools, including large language models such as ChatGPT, Claude, and Gemini, has drawn considerable interest in mathematics education (Wardat et al., 2023), particularly when these technologies are integrated with pedagogical approaches designed to promote active participation and authentic problem-solving (Baidoo-Anu & Ansah, 2023). Research examining the educational effectiveness, implementation constraints, and theoretical grounding of such integration, however, remains limited and methodologically dispersed (Hwang & Chang, 2023).

Mathematics continues to present persistent challenges across educational contexts, with sustained patterns of mathematical anxiety, diminished motivation, and inequitable access to competence (Boaler, 2016), and technological availability alone has not resolved them. Indonesian secondary students, for instance, continue to demonstrate difficulties with contextual mathematical reasoning despite expanded access to digital resources (Susilwati et al., 2025; Zulkardi & Kohar, 2018), suggesting that the central challenge is the absence of pedagogically coherent models that integrate adaptive technological support with established principles of mathematical learning.

Research on artificial intelligence in education has documented potential benefits for personalised learning and adaptive feedback, while game-based learning scholarship has established theoretical frameworks explaining how mechanics produce engagement and motivation. Large language models add capabilities for scaffolding problem-solving processes, generating contextually appropriate content, and simulating collaborative environments. Their convergence with game-based learning, however, remains theoretically underspecified: game-based learning theory explains how mechanics produce engagement, and AI scaffolding theory explains how adaptive support enhances learning, but their interaction is poorly characterised. It is not clear whether AI enhancement amplifies existing game-based mechanisms, introduces affordances of a different kind, or undermines intrinsic motivation through over-scaffolding.

Contemporary mathematics education scholarship emphasises engagement, personalisation, and authentic problem-solving as mechanisms for addressing persistent challenges (Schoenfeld, 2013), and game-based learning strategies have shown capacity to improve motivation, persistence, and understanding across diverse learner populations. Several frameworks support this potential. Self-Determination theory (Deci & Ryan, 2000) proposes that game mechanics can address psychological needs for autonomy, competence, and relatedness, with implications for intrinsic motivation. Flow theory (Csikszentmihalyi, 1990/2008) describes how balanced challenge maintains learners between anxiety and boredom. The Mechanics-Dynamics-Aesthetics framework (Hunicke et al., 2004) provides analytical tools for connecting specific game components to emotional and cognitive experiences that may support learning. Deterding et al. (2011) distinguish full-fledged games (game-based learning) from game-like elements applied to non-game contexts (gamification), though both share theoretical foundations in motivation and experiential learning.

Recent conceptual work proposes that LLMs may serve not only as learning tools but also as pedagogical design instruments, enabling translation and adaptation of learning activities across modalities (Haiman et al., 2025, framed as the rationale for a planned intervention rather than as

empirical findings). Adjacent design-oriented contributions raise the hypothesis that, when contextual mathematical tasks are supported by LLM-driven adaptive scaffolding and integrated with gamification mechanics, learning motivation may mediate the link between instructional design and gains in problem solving (Susilwati et al., 2025). Questions of equitable design remain inadequately addressed, with proposals raising the possibility of differential effects across learner subgroups, including by gender and prior mathematical knowledge.

The current evidence base is heterogeneous in theoretical framing, methodological approach, and empirical context, with available studies predominantly recent and often appearing as preprints or conference proceedings. This diversity, while challenging for meta-analytic aggregation, reflects the exploratory stage of the field. Following methodological guidance for emerging research domains (Ferrari, 2015), a narrative literature review provides an appropriate framework for synthesising emerging evidence, identifying theoretical patterns and tensions, and establishing directions for future research at the intersection of generative AI and game-based mathematics learning.

Research Questions. Given the exploratory stage of this research domain and the need to synthesize heterogeneous evidence, this narrative review adopts a thematic approach organized around three research questions. These questions address usage modalities, pedagogical dimensions, and implications for future research and practice:

- **RQ1.** How are generative AI tools, particularly large language models, currently being employed to support mathematics education through game-based or gamified learning strategies?
- **RQ2.** What are the main pedagogical affordances, design approaches, and challenges reported in the literature regarding the integration of generative AI with game-based mathematics learning?
- **RQ3.** What implications and directions for future research emerge from existing evidence, particularly concerning learner engagement, educational equity, and instructional design?

Methods

This review adopts a structured narrative format rather than a scoping or systematic review design, a choice that reflects both the developmental stage of the research domain and the interpretive aims of the synthesis. Two considerations motivate this positioning. First, the corpus addressing generative AI integration with game-based mathematics learning remains small, theoretically heterogeneous, and methodologically dispersed, conditions under which narrative synthesis is better suited than scoping mapping or quantitative aggregation (Ferrari, 2015; Greenhalgh et al., 2018). Second, the review pursues conceptual integration across distinct theoretical traditions, including self-determination theory, flow, the MDA framework, and emerging accounts of LLM-mediated scaffolding, an analytic task that depends on interpretive judgement rather than on descriptive enumeration.

The review nonetheless incorporates procedural elements typically associated with systematic and scoping reviews. This hybridisation follows Ferrari's (2015) explicit recommendation that narrative reviews benefit from borrowing structured search and screening procedures from systematic methodologies to reduce selection bias and improve transparency. Specifically, the review adopts the Population-Concept-Context framework to anchor inclusion decisions (Aromataris, 2024; Peters et al., 2020; Pollock et al., 2023), applies a documented two-phase search across multiple databases, employs independent dual screening with inter-rater agreement assessment, and reports the selection process through a flow diagram. These features serve transparency and replicability rather than the evidentiary logic of systematic review. The synthesis itself remains interpretive and thematic, organised around the three research questions, and does not aggregate effect sizes, assess risk of bias across studies in formalised terms, or claim exhaustive coverage of the domain. Readers should accordingly interpret the findings as a structured account of an emerging field rather than as a definitive evidentiary synthesis.

Throughout the manuscript, citations in the introductory and theoretical sections may include conceptual, theoretical, or design-oriented contributions that fall outside the empirical evidence base; such citations are used to frame the research questions and situate the review within the wider literature, and are not treated as evidence supporting the claims advanced in the synthesis.

Analytical Framework. The PCC framework was operationalised as follows. The population comprises students in mathematics learning across primary, secondary, and early tertiary education, with some studies also examining teachers as designers or facilitators of AI-enhanced environments. The concept is the integration of generative AI, specifically large language models, with game-based or gamified learning strategies, with attention to effects on mathematical problem-solving, problem-posing, engagement, and anxiety. The context spans formal classroom instruction, supplementary or after-school interventions, and virtual learning platforms, in both physically present and digitally mediated environments, provided the activities are structured and focused on mathematics learning.

Eligibility Criteria. Studies were eligible if they examined LLM integration with game-based or gamified learning in mathematics education, involved students in primary, secondary, or early tertiary education, and documented effects on problem-solving, conceptual understanding, or affective dimensions such as engagement or anxiety. The restriction to generative AI excluded earlier rule-based systems. Eligible contexts comprised formal classroom instruction, supplementary programs, and structured online learning environments; informal settings and purely theoretical papers without empirical data were excluded.

Studies published between 2021 and the present were included, reflecting the emergence of contemporary LLMs after ChatGPT's 2022 release. Publications were primarily in English, with studies in other languages not excluded. Primary sources comprised peer-reviewed journal articles and conference proceedings from recognised educational technology, mathematics education, and learning sciences venues; preprints and working papers with sufficient methodological documentation were also included, given the rapid evolution of the field. Diverse methodological designs were accommodated provided they generated original empirical data rather than synthesising existing literature or proposing untested frameworks.

Information Sources and Search Strategy. The literature search was conducted in two sequential phases. The first phase applied a structured query across six sources: five academic databases (Scopus, Web of Science, ERIC, IEEE Xplore, ACM Digital Library) and Google Scholar (filtered to include only peer-reviewed articles). The search targeted peer-reviewed journal articles and conference proceedings published between 2021 and 2025, focusing on three thematic axes: (1) artificial intelligence (generative AI, large language models, ChatGPT, Claude, Gemini, AI tools); (2) game-based learning (digital games, educational games, gamification, GBL, serious games); and (3) mathematics education (mathematical thinking, mathematics education, mathematical problem solving, mathematical problem posing).

This initial phase returned 226 records: Scopus (22), Web of Science (10), ERIC (7), IEEE Xplore (10), ACM Digital Library (75), and Google Scholar (102). After removing 16 duplicates, 210 records were independently screened by two researchers based on title and abstract, with disagreements resolved through consensus discussion. Only one study met all inclusion criteria: Haiman et al. (2025).

Given the limited results from the structured phase, a second exploratory search was conducted using Google Scholar through Publish or Perish (version 8.0) (Harzing, 2011). Six specific queries combined terms such as generative AI/LLM/ChatGPT with game-based learning/gamification and mathematics education/mathematical problem solving. For each query, two searches were performed (unrestricted and peer-reviewed only), with a predetermined limit of 20 results per search based on relevance ranking. This phase yielded 228 documents.

All retrieved records were exported in RIS format and imported into Rayyan for deduplication, resulting in 174 unique entries. These records were transferred to Zotero for manual enrichment of missing metadata. The 174 records were independently screened by two reviewers using a three-level coding scheme (include, exclude, maybe). Raw agreement was 97.1% (169 of 174), with Cohen's $\kappa = 0.801$ (95% CI [0.633, 0.969]); the wide interval reflects prevalence imbalance rather than substantive disagreement. The five disagreements and maybe codes were resolved by consensus, and studies not intersecting all three thematic dimensions were excluded.

Thirteen studies from this exploratory phase were retained for inclusion. Notably, Haiman et al. (2025) appeared in both the structured and exploratory search phases.

Data Charting and Synthesis Approach. Following initial screening, full-text assessment was conducted for studies meeting eligibility criteria. Data charting employed a structured extraction template documenting study characteristics (design, sample, context), AI integration modalities, game component, and reported outcomes. This process facilitated systematic comparison across studies while preserving contextual richness essential for narrative synthesis.

The analytical categories used to compare studies in subsequent tables and figures, including LLM modalities, game-based learning mechanics, and outcome dimensions, were developed bottom-up through iterative thematic coding of the extracted data, with category labels chosen to align with established terminology in the game-based learning and gamification literature where applicable.

The synthesis adopted a thematic approach organised around the three research questions, identifying convergent findings, theoretical tensions, and methodological patterns across studies that conceptualised and operationalised LLM-game integration in different ways. This interpretive approach permitted integration of diverse study designs while maintaining focus on theoretical coherence and practical implications.

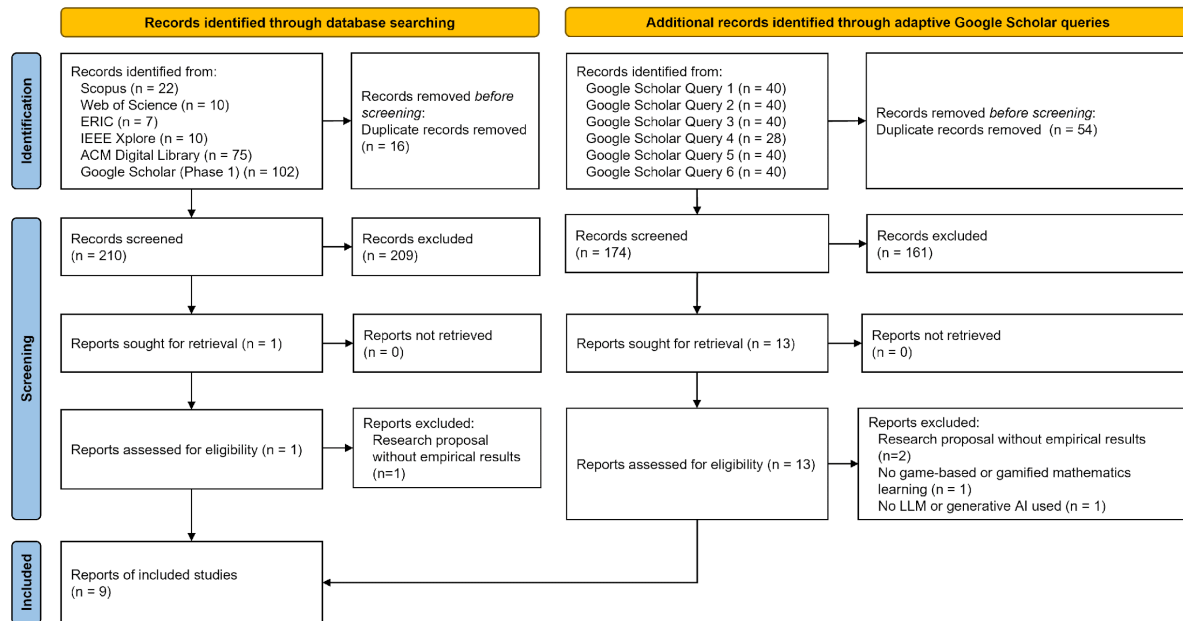
Results

Selection of Sources of Evidence. Thirteen studies underwent full-text screening. Four studies were excluded: Haiman et al. (2025) was a research proposal describing a planned intervention without pilot data or preliminary findings; Rühl & Thurm (2024) was a design study proposing RAG-enhanced principles without empirical outcome data; Karlström (2024) focused on AI-supported tutoring without game-based components; Kulkarni et al. (2025) employed rule-based AI (minimax algorithm) rather than generative models. One Turkish study (Divrik et al., 2025) was retained after translation using Google Translate. The final corpus comprised 9 empirical studies. The PRISMA flow diagram (Figure 1) summarizes the selection process.

Studies excluded from the empirical corpus, such as Haiman et al. (2025), are referenced elsewhere in the manuscript as conceptual or design-oriented contributions where appropriate to the framing of the review, with their non-empirical status made explicit at the point of citation.

Figure 1

PRISMA Flow Diagram Outlining the Screening Process



Overview of Included Studies. Nine empirical studies published between 2024-2025 examined LLM integration with game-based mathematics learning across eight countries. Sample sizes ranged from 5 to 51 participants (median=24). Most studies (n=7) employed qualitative or design-based methodologies; three used quasi-experimental designs with control groups (Divrik et al.,

2025; Mazuru & Havârneanu, 2025; Sayeg-Sánchez & Rodriguez-Paz, 2024). The corpus is composed predominantly of small-scale pilots with brief interventions: median duration was one session, with only two studies exceeding one month. Table 1 summarizes population characteristics, game component, AI role, and key findings across studies.

Table 1
Summary of Included Studies

Study	Population/Context	Game Component	AI Role	Key Finding
Choi & Chang (2025)	5 elementary teachers, Macau	Browser-based serious games with non-digital mechanics	WebSim authoring tool	High teacher satisfaction; prompt design learning curve
De Santis et al. (2025)	12 middle schoolers, Italy	Minetest sandbox (“Superflat Math”) with post-gameplay reflection worksheets	GPT-4o as struggling learner	Stimulated metacognitive reflection; gaps in scaffolding
Divrik et al. (2025)	51 fourth graders, Turkey	Scratch-based games (physical mazes, narrative escape game)	ChatGPT-4 game design	Attitude improved ($p < 0.05$); no motivation differences
Eckhardt & Thurm (2025)	Middle schoolers, Germany	Digital escape game “The Scavenger Hunt” (museum setting, speech-to-text)	LLM character “Justin”	Deep explanations; scaffolding gaps for struggling students
Mazuru & Havârneanu (2025)	Students ages 10-11, Romania	Scratch/Blockly programming games (curricular educational program)	ChatGPT in Scratch/Blockly	Higher performance (8.69 vs 6.61); reduced anxiety
Rühl & Thurm (2025)	14 low-performers, Germany	Digital escape game (ancient temple, reflection symmetry task)	GPT as “Kleo” character	Zero task completion; students lack LLM interaction skills
Sayeg-Sánchez & Rodriguez-Paz (2024)	34 engineering students, Mexico	Text-based role-playing game (SIR epidemiologic model simulation)	ChatGPT role-playing	20% performance gain; low engagement
Walkington et al. (2025)	36 middle school girls, USA	Video game pitch project (math story problems with visual illustrations)	ChatGPT/DALL-E2 authoring	168 prompts; cultural authenticity concerns
Yue et al. (2024)	14 middle schoolers, China	Chat-based collaborative problem solving with 2 virtual peers (Alice & Bob)	Multi-persona classroom	High constructive interaction ($M = 4.36/5$)

RQ1. Usage modalities of LLM integration: Three distinct modalities. Analysis revealed three distinct modalities through which LLMs support mathematical learning, each reflecting different pedagogical assumptions and technical implementations.

- 1) LLM as Authoring Tool.** Three studies positioned LLMs as content generators where teachers or students create game materials through prompt engineering (Choi & Chang, 2025; Divrik et al., 2025; Walkington et al., 2025). Teachers using WebSim valued real-time feedback and rapid differentiation but reported prompt design learning curves (Choi & Chang, 2025). Divrik et al. (2025) found significant within-group attitude improvement ($p < 0.05$) but no between-group motivation differences, with external motivation paradoxically higher in the control group. Walkington et al. (2025) exemplified student-driven problem-posing, where 36 middle school girls used ChatGPT and DALL-E2 to create personalized story problems embedding cultural interests (zombies, wizards, Spiderman). Analysis of 168 prompts revealed extensive cultural references ($n=109$), academic vocabulary ($n=54$), and chain prompting ($n=45$ instances) seeking authentic voice (“make it sound like me”). However, systematic evaluation revealed only 6.8% of problems passed all authenticity criteria, with failures in realistic data (42.5% pass rate) and appropriate vocabulary (38.8% excessively difficult). Students identified bias in names (61.7-84.7% White) and images (seven of 14 groups detected racial bias, four detected gender bias) yet remained optimistic about ChatGPT’s engagement benefits, raising questions about whether authenticity matters when students find problems motivating.
- 2) LLM as Pedagogical Agent.** Four studies configured LLMs as learning companions employing non-telling scaffolding strategies (De Santis et al., 2025; Eckhardt & Thurm, 2025; Rühl & Thurm, 2025; Yue et al., 2024). Implementations varied in technical sophistication, from single-character interactions to architecturally complex multi-persona virtual classrooms. In simpler implementations, GPT-4o configured as a struggling learner stimulated metacognitive reflection but sometimes overvalued well-argued incorrect answers (De Santis et al., 2025). Students anthropomorphized LLM characters such as Justin and Kleo, treating them as friends (Eckhardt & Thurm, 2025; Rühl & Thurm, 2025). However, despite six prompt engineering iterations, zero groups completed tasks among 14 low-performing learners, suggesting students lack LLM interaction competencies (Rühl & Thurm, 2025). Yue et al. (2024) advanced this modality through a multi-persona virtual classroom where students collaborated with two LLM-simulated peers (Alice and Bob) with different proficiency levels. This architecture addressed two challenges naive prompting cannot solve: persona alignment (simulating authentic student proficiency levels and error patterns through symbolic schemas with error injection) and conversational procedural alignment (orchestrating multi-stage collaborative problem-solving through meta planning modules). The system generated 695 messages across 40 sessions with high constructive interaction ratings ($M=4.36/5$). Preliminary exploration demonstrated why architectural sophistication was necessary: vanilla approaches produced agents claiming limited proficiency while providing flawless calculations and bypassing collaborative stages.
- 3) LLM as Problem-Solving Environment.** Two studies integrated ChatGPT into programming or text-based simulations (Mazuru & Havârneanu, 2025; Sayeg-Sánchez &

Rodriguez-Paz, 2024). Mazuru and Havârneanu (2025) achieved significantly higher school efficiency (mean=8.69 vs. control=6.61) with reduced anxiety. Sayeg-Sánchez and Rodriguez-Paz (2024) employed ChatGPT in a pandemic simulation role-playing game, producing 20% performance improvement ($t=2.78$, $p=0.010$) with skill-specific gains in graph interpretation ($t=3.17$, $p=0.002$) and decision making ($t=3.05$, $p=0.002$) but no improvement in understanding rates of change ($t=-0.22$, $p=0.585$). Despite performance gains, students reported low enjoyment and curiosity, with strong correlations between interactivity and engagement ($r=0.601-0.823$).

RQ1. Usage modalities of LLM integration: Critical note on transparency and replicability.

Five of nine studies did not specify exact LLM model versions (Eckhardt & Thurm, 2025; Mazuru & Havârneanu, 2025; Sayeg-Sánchez & Rodriguez-Paz, 2024 version unspecified; Walkington et al., 2025 used GPT-3.5 now outdated; Yue et al., 2024 specified GPT-4o but acknowledged evolution). Only Choi and Chang (2025) explicitly declared using an API model picker. This lack of transparency creates replicability challenges, as model capabilities evolve rapidly and researchers cannot reproduce findings without knowing which version generated results.

RQ2. Pedagogical affordances, design approaches, and challenges: Pedagogical affordances.

Cross-study synthesis identified affordances organized into four thematic clusters.

- 1) **Cognitive scaffolding** included real-time immediate feedback (Choi & Chang, 2025; Divrik et al., 2025; Sayeg-Sánchez & Rodriguez-Paz, 2024; Yue et al., 2024). Metacognitive reflection appeared through non-telling approaches (De Santis et al., 2025; Eckhardt & Thurm, 2025; Rühl & Thurm, 2025; Yue et al., 2024). Task comprehension support appeared (Rühl & Thurm, 2025; Yue et al., 2024). Reduced cognitive load through collaboration appeared when distributing reasoning across virtual peers (Yue et al., 2024). Reciprocal teaching emerged when students helped error-prone virtual peers (Yue et al., 2024).
- 2) **Motivational and affective affordances** spanned the three psychological needs of self-determination theory. Autonomy support emerged through student choice in problem contexts and customisation (Walkington et al., 2025) and self-paced progression (Mazuru & Havârneanu, 2025). Competence support emerged through tiered challenge calibration, immediate feedback, and visible progress indicators. Relatedness support emerged through collaborative mechanics (Yue et al., 2024) and narrative character relationships in which students anthropomorphised LLM characters (Eckhardt & Thurm, 2025; Rühl & Thurm, 2025). Narrative framing, escape-room mechanics, simulation, and culturally relevant problem posing supported engagement across studies. Reduced mathematics anxiety reached statistical significance in the two quasi-experimental studies (Divrik et al., 2025: $p=-0.281$, $p=0.05$; Mazuru & Havârneanu, 2025: $\text{Gamma}=-0.223$, $p=0.043$), and Yue et al. (2024) reported confidence gains and a degree of psychological safety in interactions with virtual peers.
- 3) **Equity and accessibility affordances** included increased participation when shy or low-performing students became more active (Choi & Chang, 2025; Eckhardt & Thurm, 2025; Rühl & Thurm, 2025). Rapid differentiation allowed teachers to generate parallel versions

for different ability levels (Choi & Chang, 2025). Student-centered learning shifted pedagogical approaches from direct instruction to facilitation.

- 4) **Efficiency and scalability** appeared through speed of generation, with students creating problems in under 30 seconds (Walkington et al., 2025). Adaptivity emerged through difficulty adjustment and dynamic persona schema updates (Yue et al., 2024).

RQ2. Pedagogical affordances, design approaches, and challenges: Design approaches. Studies employed diverse design strategies organized into four categories.

- 1) Scaffolding approaches included non-telling scaffolding guiding through questions rather than solutions (De Santis et al., 2025; Eckhardt & Thurm, 2025; Rühl & Thurm, 2025). Tiered hints provided progressive support levels: motivational, strategic, then content specific (De Santis et al., 2025; Rühl & Thurm, 2025).
- 2) Engagement strategies included narrative framing and NPC integration embedding mathematics in stories (Choi & Chang, 2025; Eckhardt & Thurm, 2025; Rühl & Thurm, 2025; Sayeg-Sánchez & Rodriguez-Paz, 2024; Walkington et al., 2025). Iterative prompt engineering appeared essential. Sayeg-Sánchez and Rodriguez-Paz (2024) conducted six iterations using the CLEAR framework (Concise, Logical, Explicit, Adaptive, Reflective). Walkington et al. (2025) documented 45 chain prompting instances by students.
- 3) Technical implementations addressed multimodal interaction and language accessibility. Visual and textual multimodality combined diagrams with text. Speech-to-text interaction reduced writing barriers (Eckhardt & Thurm, 2025). Technical language bridging connected everyday, educational, and technical language registers (Rühl & Thurm, 2025; Yue et al., 2024).
- 4) Quality assurance frameworks included Palm's (2008) authenticity framework evaluating generated problems on six dimensions: Event, Question, Existence, Realism, Specificity, and Language Use (Walkington et al., 2025). Five-stage collaborative problem-solving formulation structured interaction through explicit stages adapted from PISA 2015 (Yue et al., 2024). Symbolic schema representations encoded mathematical thinking as structured sub-tasks, with task schemas representing solutions and persona schemas containing error-injected versions reflecting student proficiency. Error injection from teacher-specified misconceptions created pedagogically meaningful mistakes. Heterogeneous skill levels among virtual peers created diversity (Yue et al., 2024).

RQ2. Pedagogical affordances, design approaches, and challenges: Reported challenges. Studies reported challenges across eight categories.

- 1) Pedagogical challenges included LLM overvaluation of well-argued incorrect answers (De Santis et al., 2025). Three studies reported students with knowledge gaps not progressing without scaffolding (De Santis et al., 2025; Eckhardt & Thurm, 2025; Rühl & Thurm, 2025) with zero task completion rate among low-performing learners (Rühl & Thurm, 2025). Learning curves for prompt design appeared consistently. Students struggled to initiate

LLM conversations (Rühl & Thurm, 2025; Walkington et al., 2025; Yue et al., 2024). LLM generation of mathematically inaccurate or developmentally inappropriate content appeared (Walkington et al., 2025).

- 2) Technical challenges included session limits, device availability, and absence of analytics dashboards (Choi & Chang, 2025).
- 3) Engagement challenges showed low enjoyment and curiosity despite performance gains (Sayeg-Sánchez & Rodriguez-Paz, 2024).
- 4) Authenticity challenges were substantial. Only 6.8% of ChatGPT-generated problems passed all six authenticity criteria (Walkington et al., 2025). Specific failures included unrealistic events (48.5%), inappropriate questions (40%), nonexistent information (49.4%), unrealistic data (42.5%), and excessively difficult vocabulary (38.8%). Only 66 of 103 problems asked questions and only 87 of 103 contained mathematical data.
- 5) Bias and equity challenges emerged when analysis revealed 61.7-84.7% White demographic representation in ChatGPT-generated names (Walkington et al., 2025). Seven of 14 student groups identified racial bias in DALL-E2 images. Four groups identified gender bias. DALL-E2 also generated mathematically incorrect diagrams.
- 6) Linguistic style challenges appeared when virtual peers sounded overly formal compared to real classmates (Yue et al., 2024).
- 7) Duration and sample limitations showed all studies demonstrated short-term interventions with maximum one month duration and small sample sizes with the largest being 51 participants.
- 8) Task completion challenges revealed zero task completion despite LLM support (Rühl & Thurm, 2025). Among 14 low-performing learners, no groups successfully completed the reflection symmetry task despite tiered hints, visual aids, and language bridging.

Table 2 provides an overview of how the nine empirical studies combine LLM modalities with game-based learning mechanics across different mathematical domains. The categories used in the three analytical dimensions (LLM modalities, game-based learning mechanics, outcome categories) were generated through bottom-up thematic analysis of the included studies, with category labels drawing on established vocabularies in the game-based learning and gamification literature (Deterding et al., 2011; Hunicke et al., 2004) to support comparability with prior work. Three LLM modalities recur in the corpus: authoring tools (3 studies), pedagogical agents (4 studies), and problem-solving environments (2 studies). In the distribution of game mechanics, narrative framing and gamification elements appear most frequently, while multi-persona collaboration and competition mechanics appear in fewer studies. The outcome mapping shows that critical challenges were reported in eight of nine studies (89%) and that engagement improvements were reported across all three modalities. Recurring patterns include reported difficulties with knowledge gaps in pedagogical agent contexts (De Santis et al., 2025; Eckhardt

& Thurm, 2025; Rühl & Thurm, 2025) and reductions in mathematics anxiety in both authoring and problem-solving environment contexts (Divrik et al., 2025; Mazuru & Havârneanu, 2025).

Table 2

Summary and Comparison of LLM-GBL Integration Techniques in Mathematics Education

LLM Modalities			Game-Based Learning Mechanics								Outcome Categories					Study Context		
																Target Users	Mathematical Domain	Study
✓								✓	✓				✓	✓	✓	Teachers	Problem solving (diverse topics)	Choi & Chang (2025)
✓								✓		✓			✓		✓	Students	Arithmetic (basic operations)	Divrik et al. (2025)
✓														✓	✓	Students	Problem posing (word problems)	Walkington et al. (2025)
	✓				✓										✓	Students	Geometric reasoning	De Santis et al. (2025)
	✓		✓											✓	✓	Students	Combinatorics (counting)	Eckhardt & Thurm (2025)
	✓		✓				✓								✓	Students	Proportional reasoning	Rühl & Thurm (2025)
	✓			✓		✓								✓	✓	Students	Algebra (linear equations)	Yue et al. (2024)
		✓								✓	✓	✓			✓	Students	Computational thinking	Mazuru & Havârneanu (2025)
		✓	✓					✓		✓					✓	Students	Functions & Modeling	Sayeg-Sánchez & Rodríguez-Paz (2024)

Legend: Category Definitions

LLM Modalities

- Authoring Tool: LLM used by teachers or students to generate educational content (games, problems, scenarios)
- Pedagogical Agent: LLM acts as conversational tutor, virtual peer, or instructional character
- Problem-solving Environment: LLM integrated into interactive problem-solving or simulation platforms

Game-Based Learning Mechanics

- Narrative/Quest: Story-driven frameworks, escape games, role-play scenarios
- Multi-persona: Multiple AI agents with differentiated roles or expertise levels
- Sandbox/Exploration: Open-ended environments for reflection and self-directed inquiry
- Competition: Competitive elements between students or with virtual peers
- Tiered Hints: Progressive scaffolding through graduated hint systems
- Gamification Elements: Points, badges, rewards, participation incentives
- Simulation/Role-play: Real-world modeling scenarios requiring decision making
- Programming/Block-based: Coding environments (Scratch, Blockly) for algorithmic thinking

Outcome Categories

- Performance Gains: Documented improvements in mathematical achievement or task completion
- Anxiety/Affect: Changes in mathematics anxiety, confidence, or emotional engagement
- Engagement/Participation: Increased curiosity, sustained interaction, or active participation
- Equity/Inclusivity: Impacts on marginalized learners, bias recognition, cultural identity work
- Critical Challenges: Documented limitations (authenticity failures, knowledge gaps, technical issues)

RQ3. Implications for engagement, equity, and instructional design: Engagement and motivation. Yue et al. (2024) identified four mechanisms through which multi-persona collaboration enhanced engagement. Reduced cognitive load emerged when distributing reasoning across virtual peers made problems more tractable, with students contrasting brief peer-like conversational turns favorably with prolonged teacher-led instruction. The importance of interactivity receives reinforcement from Sayeg-Sánchez and Rodríguez-Paz (2024), who found strong correlations between interactivity and both enjoyment ($r=0.601$, $p<0.05$) and curiosity

($r=0.823$, $p<0.05$), suggesting text-based formats lacking interactivity undermine affective engagement.

Scaffolding deeper reasoning occurred through heterogeneous skill levels creating cognitive conflict, with participants explicitly valuing disagreements and observing alternative mathematical methods (Yue et al., 2024). The reciprocal teaching dynamic further scaffolded learning, with students noting that diversity made the experience better because they could help classmates while learning themselves. Friendly competition driving proactive participation emerged when students described competitive motivation to contribute first while still working toward shared solutions, as evidenced by high agreement on reaching shared understanding ($M=4.43/5$). Immediate targeted feedback promoting persistence appeared when students contrasted process-oriented feedback favorably with answer-giving tendencies of tutor-mode LLMs, occupying a pedagogically optimal middle ground.

RQ3. Implications for engagement, equity, and instructional design: Confidence building.

Two distinct forms of confidence gains emerged (Yue et al., 2024). Supportive scaffolding confidence developed when students felt more capable due to virtual peers providing reliable assistance, with some finding the virtual environment less intimidating than human peer interactions. Metacognitive confidence through articulation and error-handling emerged when students gained belief in their capacity to monitor cognitive processes, reporting increased confidence explaining problem solving to others.

Critically, students reframed errors positively, with errors becoming diagnostic information revealing fixable gaps rather than ego-threatening failures, representing growth mindset characteristics. Virtual peers' fallibility contributed to error reframing, with students who corrected Alice's errors experiencing validation of their own knowledge.

RQ3. Implications for engagement, equity, and instructional design: Equity considerations.

Anxiety reduction appeared in both quasi-experimental studies with statistically significant effects (Divrik et al., 2025: $\rho=-0.281$, $p=0.05$; Mazuru & Havârneanu, 2025: $\text{Gamma}=-0.223$, $p=0.043$). The convergence across two studies with different populations and game types is suggestive but limited; the candidate mechanisms (privacy of failure, self-paced progression, immediate feedback, narrative distraction) remain to be tested directly.

Increased participation for marginalized students appeared when students who rarely spoke in class became active (Choi & Chang, 2025). However, zero task completion rates provide a critical equity cautionary tale: increased participation does not guarantee success without calibrated scaffolding (Rühl & Thurm, 2025).

Bias recognition and critical AI literacy emerged when middle school girls of colour recognised and articulated AI bias in Walkington et al. (2025). Seven of 14 groups identified racial bias in DALL-E2 outputs, with students articulating that AI is human-made and can carry comparable biases. Four groups identified gender bias. The finding is consistent with the possibility that middle school students can develop critical AI engagement when explicit instruction is provided, although the privileged setting of the study and the substantial researcher mediation involved limit how broadly this can be generalised. However, bias pervasiveness in both ChatGPT names

(predominantly White at 61.7-84.7%) and DALL-E2 images signals need for continuous auditing and mitigation.

RQ3. Implications for engagement, equity, and instructional design: Skill-specific development. Within the single study reporting skill-disaggregated outcomes, LLM-game integration was associated with differential effects across mathematical competencies. Graph interpretation improved significantly ($t=3.17$, $p=0.002$), as did decision-making ($t=3.05$, $p=0.002$), while understanding rates of change did not ($t=-0.22$, $p=0.585$) (Sayeg-Sánchez & Rodríguez-Paz, 2024). The pattern is consistent with the possibility that some mathematical processes are less responsive to game-based simulation, but the single-study basis precludes broader generalisation.

Discussion

This synthesis of nine empirical studies suggests that LLM integration with game-based mathematics learning may operate through mechanisms not fully captured by existing theoretical frameworks, although the small and heterogeneous evidence base supports interpretive proposals rather than firm conclusions. Across the corpus, recurring patterns point to affordances, limitations, and tensions that warrant closer theoretical attention from both game-based learning scholarship and research on AI in education. The discussion that follows develops three interpretive frameworks generated from the reviewed studies, examines implementation difficulties that complicate optimistic accounts of technological enhancement, and identifies directions for empirical work needed to test the proposals advanced here. The frameworks should be read as hypotheses derived from a limited evidence base, not as validated models.

Theoretical Contribution 1: Socio-Emotional Mediation as Primary Mechanism.

Contemporary mathematics education scholarship emphasizes engagement, personalization, and authentic problem solving as mechanisms for addressing persistent challenges (Schoenfeld, 2013). Game-based learning theory proposes that mechanics produce engagement through psychological need satisfaction (Deci & Ryan, 2000), while AI scaffolding theory proposes that adaptive support enhances learning through cognitive load reduction and immediate feedback. The patterns observed across the reviewed studies suggest that affective mediation may play a more prominent role than these accounts emphasise, with cognitive mechanisms appearing to operate alongside rather than independently of affective pathways. This reading is tentative: it draws on a limited number of studies with small samples and short interventions, and it should be tested against larger and more controlled evidence.

Two quasi-experimental studies reported statistically significant anxiety reduction across different populations and game types (Divrik et al., 2025: $p=-0.281$, $p=0.05$; Mazuru & Havârneanu, 2025: $\text{Gamma}=-0.223$, $p=0.043$). The candidate mechanism is psychological safety: narrative distance from mathematical threat, privacy from peer judgement, self-paced progression, and immediate correction that reframes errors as diagnostic. Yue et al. (2024) documented growth-mindset characteristics emerging from environmental structure rather than direct instruction, with students reframing errors as fixable gaps.

A tentative socio-emotional mediation framework can be drawn from these patterns. As a working hypothesis, LLM-game integration may enhance mathematical learning in part by creating

psychologically safer environments in which anxiety reduction, confidence building, and growth mindset cultivation support engagement, with cognitive affordances operating alongside affective conditions rather than in isolation. The framework awaits empirical testing and should not be read as displacing existing cognitive models of scaffolding effectiveness.

The framework offers one possible reading of an apparent paradox in the corpus. Sayeg-Sánchez and Rodríguez-Paz (2024) reported a 20% performance improvement alongside low enjoyment and curiosity, with notable correlations between interactivity and both constructs ($r=0.601-0.823$). Performance gains may have arisen from structured problem-solving practice independent of affective engagement, while the low engagement scores leave open the question of whether stronger affective conditions would have produced larger gains. The single-study evidence base does not allow this counterfactual to be resolved.

If borne out by further evidence, the framework would raise questions for both source domains. Game-based learning theory holds that mechanics produce engagement through psychological need satisfaction, but the reviewed studies suggest that mechanics may be insufficient when LLM integration introduces stochastic behaviour and procedural misalignments. AI scaffolding theory holds that adaptive support enhances learning through cognitive mechanisms, yet the same studies indicate that cognitive support may be insufficient for students with knowledge gaps, with effectiveness contingent on implementation conditions that create the affective space for productive engagement.

Theoretical Contribution 2: Dimensional Authenticity and Pedagogical Trade-Offs. Walkington et al. (2025) reported that only 6.8% of ChatGPT-generated problems passed all six authenticity criteria. Read alongside other studies in the corpus, this finding suggests that authenticity dimensions may not carry equal weight for learning, although the evidence supporting this reading is concentrated in a single study and the pattern remains speculative.

A tentative dimensional authenticity framework can be sketched on this basis, distinguishing dimensions that may matter more from dimensions that may matter less for pedagogical effectiveness. Candidate critical dimensions include data realism, where mathematical inconsistencies appear to undermine credibility (only 42.5% of problems provided realistic data); question appropriateness (only 66 of 103 problems asked questions, with 40% deemed inappropriate); and mathematical content presence (only 87 of 103 contained mathematical data). These dimensions are flagged because failures along them, in the Walkington et al. data, correspond to problems that are not pedagogically usable; the claim that they are the most consequential awaits broader comparison.

Three further bands of dimensions can be sketched. A second group appears to affect learning quality without compromising task validity, including collaborative behaviour authenticity, where students in Yue et al. (2024) valued disagreements and alternative methods from virtual peers, and scaffolding calibration, where the zero completion rate in Rühl and Thurm (2025) is consistent with the view that poorly calibrated scaffolding can hamper learning. A third group appears to affect engagement without clearly affecting learning outcomes: linguistic authenticity fits this description in Yue et al. (2024), where overly formal language drew criticism without preventing high ratings on constructive interaction. A fourth group appears to allow considerable latitude,

with event realism a likely candidate: students in Walkington et al. (2025) tolerated fantastical contexts, and the 48.5% failure rate on this criterion did not seem to undermine engagement when mathematical content remained sound.

The framework offers one reading of an apparent contradiction: students in Walkington et al. (2025) expressed enthusiasm despite only 6.8% of problems passing all criteria. If tolerance varies by dimension, failures involving fantastical events or formal language may be absorbed while failures involving impossible data or missing questions compromise the task itself. Should this hierarchy hold in further studies, design priorities would weight data realism and question appropriateness above linguistic casualness or event plausibility.

Theoretical Contribution 3: Dual Alignment Framework for Educational Simulation. Yue et al. (2024) provide observations relevant to the limits of naive prompting in pedagogically meaningful simulation. Their preliminary attempts produced agents that claimed limited proficiency while solving problems correctly, a misalignment the authors attribute to tensions between LLM pre-training objectives, which favour correct responses, and simulation objectives, which require realistic student behaviours including errors. These observations rest on a single development effort.

Drawing on Yue et al. (2024), a tentative dual alignment framework can be proposed that distinguishes two challenges plausibly affecting LLM-based educational simulation. The first, persona alignment, concerns the simulation of individual learner behaviours. LLMs tend to default to correctness, a tendency commonly attributed to pre-training on expert text and reinforcement learning prioritising accuracy. The Yue et al. design addresses this through symbolic schema representations: task schemas encode solution steps, while persona schemas contain error-injected versions reflecting student proficiency levels, with errors derived from teacher-specified misconceptions. The architecture queries persona schemas before generating responses, producing specified error types even when the underlying model would otherwise generate correct answers.

The second challenge, conversational procedural alignment, concerns the orchestration of multi-stage interactions. LLMs are well suited to contextually appropriate single turns but offer fewer mechanisms for maintaining state across turns or ensuring coverage of required stages. The Yue et al. design responds through meta planning modules that monitor dialogue state, track stage progression, and enforce conversational structure through explicit stage gates adapted from the PISA 2015 collaborative problem-solving framework. As with persona alignment, the proposal rests on a single architectural example, and its transferability remains to be demonstrated.

The framework offers one reading of why three studies implementing non-telling scaffolding reported similar difficulties with students who had knowledge gaps (De Santis et al., 2025; Eckhardt & Thurm, 2025; Rühl & Thurm, 2025). The reading is that these implementations achieved conversational coherence without achieving persona alignment: the agents could not simulate the specific knowledge configurations with which the students needed to engage. The zero completion rate in Rühl and Thurm (2025) is compatible with this interpretation, although task design, prerequisite knowledge, and intervention duration remain plausible alternative explanations.

If supported by additional evidence, the framework would suggest that educational simulation may benefit from architectural interventions addressing both individual behaviour fidelity and multi-stage process orchestration, rather than prompt engineering alone, since natural language instructions appear to have limited capacity to override model training objectives or to enforce structured state management across turns. Educational LLM applications oriented toward simulation may need explicit representation structures, error injection mechanisms, control modules, and stage gates of the kind described by Yue et al. (2024). Whether comparable architectures, tested in different settings, produce comparable results is the question on which the value of these design directions rests.

Implementation Failures and the Scaffolding Calibration Problem. Three studies implementing non-telling scaffolding reported similar difficulties, with students with knowledge gaps appearing not to progress through the intended pathways (De Santis et al., 2025; Eckhardt & Thurm, 2025; Rühl & Thurm, 2025). The convergence across three small studies in different settings and with different mathematical content suggests a recurring scaffolding calibration issue rather than isolated implementation problems, though three studies cannot establish its generality.

One reading of these difficulties is that effective scaffolding requires dynamic adjustment matching support levels to student knowledge states, a function human teachers perform through continuous informal assessment, and one for which the reviewed LLM implementations appear to offer limited diagnostic capacity. De Santis et al. (2025) configured GPT-4o as a struggling learner, a design presupposing students with sufficient understanding to teach the agent; when students themselves struggled, the persona did not shift toward explicit instruction. Rühl and Thurm (2025) provided tiered hints with visual aids and language bridging, and after six prompt engineering iterations, no groups completed the task. The observations are consistent with the view that carefully designed scaffolding sequences may be insufficient when prerequisite knowledge is missing.

The three studies align with the long-standing view that scaffolding effectiveness depends on the match between instructional strategy and prerequisite knowledge: non-telling approaches appear better suited to students with adequate foundations, while students lacking foundations may require more explicit instruction. The reviewed implementations tended to employ fixed strategies with limited flexibility to shift approach when students fell outside the target competency range. If this reading holds, the consequences would fall disproportionately on struggling learners who most depend on adaptive support, with the zero completion rate in Rühl and Thurm (2025) consistent with the concern that increased technological access does not by itself address equity when system design assumes competencies some students lack. Whether current LLM implementations widen achievement gaps in practice is a question the reviewed evidence raises but does not answer.

The Interaction Competency Gap and Motivation Paradox. Students in three of the reviewed studies appeared to find it difficult to engage productively with LLM pedagogical agents, with reported gaps in question formulation, iterative refinement, critical evaluation of responses, and strategic help seeking (Rühl & Thurm, 2025; Walkington et al., 2025; Yue et al., 2024). One reading is that effective LLM use draws on metacognitive and communicative competencies that should not be assumed at the start of an intervention. This offers a candidate explanation for findings in which increased participation did not translate into measurable learning outcomes,

though isolating the role of interaction competency would require studies designed for that purpose.

Two findings illustrate the broader question. Divrik et al. (2025) reported higher external motivation in the control group despite attitude improvements in the experimental condition; Sayeg-Sánchez and Rodriguez-Paz (2024) reported a 20% performance improvement alongside low enjoyment and curiosity. One reading of these findings distinguishes different motivational pathways, with the caveat that the distinction is interpretive and not empirically tested within the reviewed studies.

A performance pathway might operate through structured problem-solving opportunities, with improvements arising from increased practice independently of affective engagement, while an engagement pathway might draw on psychological safety, narrative immersion, and identity investment that transform mathematical activity into personally meaningful goals. The reviewed evidence is compatible with the possibility that the two pathways can dissociate: in Sayeg-Sánchez and Rodriguez-Paz (2024), performance gains coexisted with low enjoyment and curiosity, a pattern the authors associate with limited interactivity. If borne out by further research, the distinction would suggest different design priorities, with skill development possibly served by structural features and text-based interaction, and disposition development requiring richer experiential features and interaction modalities.

Equity Reconsidered: Participation, Learning, and Critical Literacy. The reviewed evidence points in two directions. Several studies report anxiety reduction, increased participation for students less engaged in class, and a degree of psychological safety in interactions with virtual peers. The same corpus contains a zero completion rate among low-performing students (Rühl & Thurm, 2025), demographic skew in generated content (Walkington et al., 2025), and interaction competency requirements that may function as barriers. A useful distinction for interpreting these findings is between access equity, the opportunity to participate, and outcomes equity—actual learning achievement. The reviewed studies are compatible with improvements in access equity through reduced anxiety and lower participation barriers, while leaving outcomes equity more open. The zero completion rate is the most cautionary signal in the corpus: students were engaged with the agent without completing the task. Whether scaffolding design failed them, or whether task difficulty and intervention duration were the binding constraints, cannot be settled from the available data.

The emergence of critical AI literacy in Walkington et al. (2025), where seven of 14 groups identified racial bias and four identified gender bias, is one of the more promising signals, although the study involved substantial researcher-led mediation that limits how broadly the observation generalises. Future scholarship in this area would benefit from foregrounding marginalised students in design and evaluation, separating access from outcomes when reporting impacts, and pairing continuous bias auditing with mitigation strategies rather than relying on one-time evaluations.

Practical Implications for Educators and Instructional Designers. The interpretive frameworks developed above suggest directions for educators and instructional designers,

presented as working considerations rather than evidence-based guidelines, each inheriting the limitations of the small corpus on which it rests.

If affective conditions function as the socio-emotional mediation framework suggests, design priorities may benefit from features that support psychological safety: private feedback, self-paced progression, and narrative framing that introduces distance from evaluative threat. The Sayeg-Sánchez and Rodríguez-Paz (2024) result, where measurable performance gains coexisted with low enjoyment and curiosity, is a reminder that learning gains and engagement do not always travel together. Whether students with the highest baseline anxiety would benefit disproportionately from such choices, and whether this would narrow existing disparities, remains an open empirical question.

The dimensional authenticity framework, if it holds in further studies, points to a heuristic for teachers using LLMs as authoring tools. Rather than weighting all six of Palm's (2008) dimensions equally, attention may be better directed to the conditions that determine whether a task functions mathematically at all. The Walkington et al. (2025) data are consistent with the concern that a meaningful share of generated problems may lack a question or sufficient mathematical content. A short screening focused on question presence, data realism, and mathematical content adequacy is a low-cost safeguard, leaving wider tolerance for fantastical scenarios and stylistic register.

The dual alignment framework speaks to instructional designers and to educators evaluating commercial LLM-based products. The Yue et al. (2024) observations, together with the recurring difficulties in three non-telling scaffolding studies (De Santis et al., 2025; Eckhardt & Thurm, 2025; Rühl & Thurm, 2025), are compatible with the view that off-the-shelf chatbots configured through prompts alone may not simulate authentic student misconceptions or sustain multi-stage collaborative problem solving with high fidelity. Systems oriented toward simulation may benefit from explicit task and persona schemas, error-injection mechanisms, and orchestration layers of the kind described by Yue et al. (2024), and educators selecting tools may find it useful to distinguish prompt-wrapped chatbots from architecturally designed pedagogical systems. Two further considerations cut across these directions. Non-telling scaffolding appears to presuppose threshold knowledge that cannot be assumed for all students, suggesting that classrooms deploying such strategies may need diagnostic procedures and fallback routes to more explicit instruction. Productive LLM use appears to depend on competencies (query formulation, iterative refinement, critical evaluation of outputs, recognition of bias) that are not acquired incidentally; teacher mediation is therefore a likely determinant of whether LLM exposure produces useful learning, with the conditions under which mediation suffices to offset bias and miscalibration remaining to be characterised.

Methodological Limitations and Future Research Directions. The nine studies offer methodological innovations (symbolic schemas with error injection, the CLEAR framework, systematic authenticity evaluation) alongside persistent constraints. All used small samples (median $n=24$) and brief interventions (median one session); only three adopted quasi-experimental designs. Five did not specify exact LLM model versions, which limits replicability given how rapidly model capabilities change, and none examined learning persistence beyond immediate post-intervention assessment. The available evidence therefore supports a cautious reading: LLM-game integration appears capable of producing positive outcomes under favourable

conditions for some students, but the conditions, mechanisms, and durability remain insufficiently characterised.

Several research directions would help address these gaps. Larger studies with adequately powered comparison groups could provide stronger causal evidence and examine heterogeneous treatment effects across student subgroups. Intervention studies that explicitly teach productive LLM engagement strategies could test whether interaction competencies can be developed systematically rather than assumed, while experimental designs that vary authenticity dimensions could test predictions of the dimensional authenticity framework. Work on adaptive scaffolding that assesses prerequisite knowledge and adjusts instructional strategy would speak directly to the calibration difficulties documented in three of the reviewed studies, and theory-driven design programmes drawing on Self-Determination theory, Flow theory, and the MDA framework could generate testable predictions about how specific design elements relate to psychological outcomes. Beyond these priorities, the field would benefit from longitudinal and transfer studies addressing learning persistence and reliance on AI assistance, systematic bias auditing paired with mitigation strategies, head-to-head comparisons across LLM models, and implementation research on how interventions perform under typical classroom constraints.

These directions outline a progression from exploratory pilots toward more systematic investigation of mechanisms, boundary conditions, and practical implementation. The current evidence base is consistent with the view that LLM-game integration can produce positive outcomes under favourable conditions; specifying when, how, and for whom such outcomes occur reliably at scale is the work that remains.

Conclusion

This narrative review synthesised nine empirical studies on the integration of large language models with game-based mathematics learning. The corpus offers theoretical openings alongside persistent challenges and remains too small and heterogeneous to support strong general claims. Three interpretive frameworks from the synthesis emerged, each offered as a working hypothesis rather than a validated model. The socio-emotional mediation framework proposes that effectiveness may rest in part on the creation of psychologically safer environments in which anxiety reduction, confidence building, and growth mindset cultivation support engagement with mathematical challenges. The dimensional authenticity framework proposes that some authenticity dimensions, notably data realism and question appropriateness, may carry greater pedagogical weight than others. The dual alignment framework distinguishes persona alignment from conversational procedural alignment, suggesting that educational simulation may benefit from architectural choices beyond what prompt engineering alone seems likely to deliver.

Several patterns in the corpus call for caution about optimistic narratives. Students with knowledge gaps did not appear to progress through the intended pathways even with carefully designed scaffolding (Rühl & Thurm, 2025). Only 6.8% of AI-generated problems passed all six authenticity criteria in the one study that systematically assessed them. Three studies pointed to interaction competencies that students appear to need but that cannot be assumed. Equity considerations remain prominent: anxiety reduction and increased participation are notable, but on the available evidence they represent necessary rather than sufficient conditions for equitable

learning, and the demographic skew documented in generated content signals that current systems may reinforce existing patterns without explicit mitigation. These cautions sit alongside methodological constraints that bear directly on what can be claimed, including small samples (median $n=24$), brief interventions, limited use of control groups, and the absence of longitudinal designs.

The three frameworks proposed here are intended as resources for future scholarship rather than as conclusions, and each generates predictions that controlled studies could test. The most consistent positive signal in the corpus concerns anxiety reduction and the conditions for psychological safety, a signal worth taking seriously while acknowledging its narrow empirical basis. Realising the wider potential of LLM-game integration will involve moving past assumptions that technological sophistication translates automatically into pedagogical effectiveness, through empirical work that specifies boundary conditions and mechanisms, theory-driven design that derives testable predictions from established frameworks, and critical attention to equity through examination of differential effects across demographic groups. Whether the frameworks proposed here hold up under that scrutiny is itself a question for the work that follows.

References

- Aromataris, E., Lockwood, C., Porritt, K., Pilla, B., & Jordan, Z. (Eds.). (2024). *JB I manual for evidence synthesis*. JBI. <https://doi.org/10.46658/JBIMES-24-01>
- Baidoo-Anu, D., & Ansah, L. O. (2023). Education in the era of generative artificial intelligence (AI): Understanding the potential benefits of ChatGPT in promoting teaching and learning. *Journal of AI*, 7(1), 52-62. <https://doi.org/10.61969/jai.1337500>
- Boaler, J. (2016). Designing mathematics classes to promote equity and engagement. *The Journal of Mathematical Behavior*, 41, 172-178. <https://doi.org/10.1016/j.jmathb.2015.01.002>
- Choi, W. C., & Chang, C. I. (2025). *WebSim: Benefits and challenges of designing teacher-authored serious games for mathematics learning in an AI world*. TechRxiv. <https://doi.org/10.36227/techrxiv.175756646.63900547/v1>
- Csikszentmihalyi, M. (2008). *Flow: The psychology of optimal experience*. Harper Perennial Modern Classics. (Original work published 1990)
- De Santis, C. M., Baccaro, S., Seniawska, W., Zarouri, I., Short, R., & Guidoni, L. (2025). Supporting students in completing mathematical reflection worksheets through an AI-based application. In A. Clark-Wilson & C. Bokhove (Eds.), *Proceedings of the 17th International Conference on Technology in Mathematics Teaching (ICTMT-17)*. https://www.matematicasuperpiatta.it/wp-content/uploads/2025/11/ICTMT17_4-DeSantis-1.pdf
- Deci, E. L., & Ryan, R. M. (2000). The “what” and “why” of goal pursuits: Human needs and the self-determination of behavior. *Psychological Inquiry*, 11(4), 227-268. https://doi.org/10.1207/S15327965PLI1104_01
- Deterding, S., Dixon, D., Khaled, R., & Nacke, L. (2011). From game design elements to gamefulness: Defining “gamification.” In *Proceedings of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments* (pp. 9-15). Association for Computing Machinery. <https://doi.org/10.1145/2181037.2181040>

- Divrik, R., Çelik, Ö., & Elmas, Z. (2025). ChatGPT ile tasarlanan matematiksel oyunların 4. sınıf öğrencilerinin matematik motivasyon ve tutumlarına etkisi. *Kocaeli Üniversitesi Eğitim Dergisi*, 8(2), 679-700. <https://doi.org/10.33400/kuje.1722769>
- Eckhardt, H., & Thurm, D. (2025). Fostering mathematical explanations through LLM-based tasks. In M. Bosch, G. Bolondi, S. Carreira, C. Spagnolo, & M. Gaidoschik (Eds.), *Proceedings of the Fourteenth Congress of the European Society for Research in Mathematics Education (CERME14)*. ERME. <https://hal.science/hal-05227461v1>
- Ferrari, R. (2015). Writing narrative style literature reviews. *Medical Writing*, 24(4), 230-235. <https://doi.org/10.1179/2047480615Z.000000000329>
- Gee, J. P. (2003). What video games have to teach us about learning and literacy. *Computers in Entertainment*, 1(1), 20-20. <https://doi.org/10.1145/950566.950595>
- Greenhalgh, T., Thorne, S., & Malterud, K. (2018). Time to challenge the spurious hierarchy of systematic over narrative reviews?. *European Journal of Clinical Investigation*, 48(6), e12931. <https://doi.org/10.1111/eci.12931>
- Haiman, A., Arroyo, I., & McLaren, B. M. (2025). Comparing the effectiveness of digital game-based learning and embodied learning. In A. I. Cristea, E. Walker, Y. Lu, O. C. Santos, & S. Isotani (Eds.), *Artificial intelligence in education. Posters and late breaking results, workshops and tutorials, industry and innovation tracks, practitioners, doctoral consortium, blue sky, and WideAIED* (Vol. 2590, pp. 409-414). Springer. https://doi.org/10.1007/978-3-031-99261-2_45
- Harzing, A. W. (2011). *The publish or perish book: Doing bibliometric research with Google Scholar*. Tarma Software Research.
- Hunicke, R., LeBlanc, M., & Zubek, R. (2004). MDA: A formal approach to game design and game research. In *Proceedings of the AAAI Workshop on Challenges in Game AI* (pp. 1-5). AAAI Press. <https://cdn.aaai.org/Workshops/2004/WS-04-04/WS04-04-001.pdf>
- Hwang, G. J., & Chang, C. Y. (2023). A review of opportunities and challenges of chatbots in education. *Interactive Learning Environments*, 31(7), 4099-4112. <https://doi.org/10.1080/10494820.2021.1952615>
- Karlström, K. (2024). *Exploring the possibility and implementation of AI-supported online math coaching* [Master's thesis, Stockholm University]. DiVA. <https://www.diva-portal.org/smash/get/diva2:1900580/FULLTEXT01.pdf>
- Kulkarni, R., Choudhari, G., Bhavik, M., Bura, R., & Bhosale, V. (2025). Gamifying early math education: An innovative AI game for children. *Journal of Information Technology Education: Innovations in Practice*, 24, 014. <https://doi.org/10.28945/5595>
- Mazuru, D. G., & Havârneanu, G. (2025). Building a curricular educational program focused on gamification, AI and ChatGPT to promote the study of mathematics. *TPM-Testing, Psychometrics, Methodology in Applied Psychology*, 32(Suppl. 7), 638-643. <https://tpmap.org/submission/index.php/tpm/article/view/2221>
- Palm, T. (2008). Performance assessment and authentic assessment: A conceptual analysis of the literature. *Practical Assessment, Research, and Evaluation*, 13(1). <https://doi.org/10.7275/0qpc-ws45>
- Peters, M. D. J., Marnie, C., Tricco, A. C., Pollock, D., Munn, Z., Alexander, L., McInerney, P., Godfrey, C. M., & Khalil, H. (2020). Updated methodological guidance for the conduct of scoping reviews. *JBIM Evidence Synthesis*, 18(10), 2119-2126. <https://doi.org/10.11124/JBIES-20-00167>

- Pollock, D., Peters, M. D. J., Khalil, H., McInerney, P., Alexander, L., Tricco, A. C., Evans, C., de Moraes, É. B., Godfrey, C. M., Pieper, D., Saran, A., Stern, C., & Munn, Z. (2023). Recommendations for the extraction, analysis, and presentation of results in scoping reviews. *JBIE Evidence Synthesis*, 21(3), 520-532. <https://doi.org/10.11124/JBIES-22-00123>
- Rühl, L., & Thurm, D. (2024). From problems to progress: LLM-based individual support in digital mathematical escape games. In E. Faggiano, A. Clark-Wilson, M. Tabach, & H.-G. Weigand (Eds.), *Proceedings of the 17th ERME Topic Conference MEDA 4* (p. 446). Università degli Studi di Bari Aldo Moro. <https://hal.science/hal-05022073>
- Rühl, L., & Thurm, D. (2025). Can you help us? Task-specific LLM-based support in a digital mathematical escape game. In M. Bosch, G. Bolondi, S. Carreira, C. Spagnolo, & M. Gaidoschik (Eds.), *Proceedings of the Fourteenth Congress of the European Society for Research in Mathematics Education (CERME14)*. ERME. <https://hal.science/hal-05283685>
- Sayeg-Sánchez, G., & Rodriguez-Paz, M. X. (2024). Use of game-based learning with ChatGPT to improve mathematical modeling competences in first-year engineering students. In *2024 ASEE Annual Conference & Exposition Proceedings*. American Society for Engineering Education. <https://doi.org/10.18260/1-2--48216>
- Schoenfeld, A. H. (2013). Reflections on problem solving theory and practice. *The Mathematics Enthusiast*, 10(1-2), 9-34. <https://doi.org/10.54870/1551-3440.1258>
- Susilawati, W., Sharov, S., Pasqa, M., & Malik, H. (2025). Integrating realistic mathematics education, AI, and gamification to enhance students' learning motivation and problem-solving skills. *Journal on Mathematics Education*, 16(4), 1257-1282. <https://doi.org/10.22342/jme.v16i4.pp1257-1282>
- Walkington, C., Pando, M., Lipsmeyer, L. L., Beauchamp, T., Sager, M., & Milton, S. (2025). Middle school girls using generative AI to engage in mathematical problem-posing. *Mathematical Thinking and Learning*, 1-22. <https://doi.org/10.1080/10986065.2025.2542724>
- Wardat, Y., Tashtoush, M. A., AlAli, R., & Jarrah, A. M. (2023). ChatGPT: A revolutionary tool for teaching and learning mathematics. *Eurasia Journal of Mathematics, Science and Technology Education*, 19(7), em2286. <https://doi.org/10.29333/ejmste/13272>
- Yue, M., Lyu, W., Suh, J., Zhang, Y., & Yao, Z. (2024). *MathVC: An LLM-simulated multi-character virtual classroom for mathematics education*. arXiv. <https://doi.org/10.48550/arXiv.2404.06711>
- Zulkardi, Z., & Kohar, A. W. (2018). Designing PISA-like mathematics tasks in Indonesia: Experiences and challenges. *Journal of Physics: Conference Series*, 947, 012015. <https://doi.org/10.1088/1742-6596/947/1/012015>

Martina Ragusa (martinaragusa@cnr.it) is a Research Fellow at the Institute for Educational Technology (ITD) of the Italian National Research Council (CNR) in Genoa, Italy.

Flavio Manganello, PhD, (flavio.manganello@cnr.it) is a Senior Researcher at the Institute for Educational Technology (ITD) of the Italian National Research Council (CNR) in Genoa, Italy.