LLAMA LIMA– A LIVING META-ANALYSIS ON THE EFFECTS OF GENERATIVE AI ON LEARNING MATHEMATICS

Anselm Strohmaier, Samira Bödefeld, Frank Reinhold

University of Education Freiburg

*The capabilities of generative AI in mathematics education are rapidly evolving, posing significant challenges for research to keep pace. Research syntheses remain scarce and risk being outdated by the time of publication. To address this issue, we present a Living Meta-Analysis (LIMA) on the effects of generative AI–based interventions for learning mathematics. Following PRISMA-LSR guidelines, we continuously update the literature base, apply a Bayesian multilevel meta-regression model to account for cumulative data, and publish updated versions on a preprint server at regular intervals. This paper reports results from the first version, including 13 studies. The analyses indicate a small positive effect (*g *= 0.28) with a wide credible interval [0.01, 0.59], reflecting the still limited evidence base.*

THEORETICAL BACKGROUND

With recent advances in large language models (LLMs) such as ChatGPT throughout the last couple of years, *generative artificial intelligence* (generative AI) has moved to the center of educational discourse. Generative AI refers to a class of AI systems, including LLMs, that generate seemingly novel content based on patterns learned from large-scale training data.

AI-based technologies have been a topic of educational research for several decades, for example in the form of intelligent tutoring systems, adaptive learning environments, and automated feedback systems (Pepin et al., 2025), but generative AI promises to offer completely new ways of supporting learning (Kasneci et al., 2023). However, due to their probabilistic and emergent nature, these systems enable flexible, language-based interactions but offer limited transparency and predictability with respect to specific outputs. Thus, the integration of generative AI into education has raised substantial challenges and concerns, including issues of accuracy, interpretability, lack of transparency, and bias. Consequently, the question of whether and how generative AI can be successfully integrated into education remains far from conclusively answered.

Additionally, while the general potential of generative AI for learning and teaching has been widely discussed and, in part, tested, analyses and empirical evidence regarding their effectiveness in specific content domains like mathematics are much more limited. In mathematics education, content-specific learning processes may interact with generative AI in distinctive ways. As one of the most recent overviews, the scoping review by Pepin et al. (2025) synthesizes existing research on the use of

2026. In H. Krzywacki, M. Hähkiöniemi, E. Haataja, M.S. Hannula & P. Hästö (Eds.). *Proceedings of the 49th Conference of the International Group for the Psychology of Mathematics Education: Research Reports*, pp. 1427-1434. PME.

ChatGPT in mathematics education and identifies mathematics-specific benefits and challenges across several dimensions, including mathematical content areas, teacher interaction, task and assessment design, student interaction and personalized learning, critical thinking, collaboration, and students' attitudes. Across all dimensions, the review highlights substantial potential alongside significant challenges, suggesting that the integration of generative AI in mathematics education is promising but far from unproblematic.

Experimental Research on Generative AI in Mathematics Education

The growing interest in generative AI has been accompanied by an increase in empirical research, also in mathematics education. However, as might be expected in a rapidly evolving field, much of the early work has been characterized by exploratory designs, short-term interventions, and highly specific instructional contexts (Pepin et al., 2025). While these studies provide valuable first insights, they often offer limited opportunities to understand causal effects on learning outcomes, making it difficult to gain a comprehensive overview of the field.

Recent research syntheses on AI in mathematics education indicate that the number of studies employing controlled experimental or quasi-experimental designs for generative AI remains small. The meta-analysis by Fock and Siller (2025) on the potential of generative AI in STEM education, based on a literature search conducted in November 2024, identified only three experimental studies in mathematics education that explicitly investigated the use of generative AI, pointing to a substantial gap between the attention the topic has received and the available empirical evidence.

At the same time, the rapid technological development of generative AI further complicates the availability and synthesis of existing findings. Syntheses on the use of AI in mathematics education largely pre-date the widespread availability of large language models (e.g., Tlili et al., 2025), which can roughly be dated to the publication of ChatGPT in November 2022. The more recent scoping review by Pepin et al. (2025), published in February 2025 and including studies up to May 2024, already illustrates how quickly assessments of affordances and limitations become outdated as model capabilities evolve. For example, they name limitations in ChatGPT's mathematical capabilities—particularly for complex tasks—as a key challenge in their usefulness for mathematics education. However, more recent work (Strohmaier et al., 2025) illustrates that the performance of LLMs has substantially improved even within 2025 and has, as of GPT-5, reached almost perfect performance for many school-relevant mathematical word problems. At the same time, Strohmaier et al. (2025) show that LLMs still exhibit substantial deficits in making sense of real-world contexts, particularly for problematic or non-sensical word problems that require genuine mathematical modelling and contextual reasoning. Consequently, empirical findings on the use of AI are often closely tied to specific models and versions and may no longer reflect the state of the technology at the time of their publication.

Beyond rapid technological change, studies on generative AI in mathematics education differ widely with respect to the type of AI system used (e.g., out-of-the box model versus enriched or specialized models), their pedagogical role (e.g., tutor, feedback provider, or task generator), the degree of learner autonomy, and the instructional integration of the technology. As with other educational innovations, the effectiveness of learning with generative AI is therefore likely to depend on a range of contextual and design-related factors.

Taken together, research on the effectiveness of generative AI in mathematics education is confronted with a methodological paradox. On the one hand, the pace of technological developments and heterogeneity of existing empirical research make systematic research syntheses highly necessary. On the other hand, conducting and publishing rigorous empirical studies—and summarizing them in conventional meta-analyses—requires substantial time (Huisman & Smits, 2017). By the time a meta-analysis becomes available, the underlying evidence base may already be outdated.

Living Meta-Analyses as a Response to Rapidly Evolving Fields

While rapid developments such as the emergence of generative AI or the new educational challenges during COVID-19 have posed an unknown urgency to educational research, other fields—most notably medicine—have long been confronted with the need to synthesize evidence under conditions of continual change and rapid evolution. In response, approaches known as *living evidence syntheses* have been developed (Elliot et al., 2014): *Living systematic reviews* (LSRs) and *living meta-analyses* (LIMAs) use several mechanisms to decrease the delay of conventional research syntheses:

- Most prominently, LSRs and LIMAs are not published as one-time products but are updated continuously, for example via preprint repositories. They therefore do not reflect a single point in time but the evolving state of the evidence.
- Accordingly, literature database searches are conducted repeatedly and often in an automated manner; newly published studies are continuously screened, coded, and added to the dataset with minimal delay.
- To ensure the inclusion of the most recent evidence, preprints are often included into the analysis, which is less common for conventional research syntheses.
- For meta-analyses, statistical methods are employed that allow for the cumulative analysis of data without inflating error rates. Bayesian meta-analytic models are particularly well suited for LIMAs, as they naturally accommodate the stepwise integration of new evidence.

THE PRESENT STUDY

Given the rapid development of generative AI technologies and the current fragmentation of empirical findings in mathematics education, a living meta-analysis offers a promising methodological framework for synthesizing evidence in this area.

LLAMA LIMA (*Large Language Models and generative AI in Mathematics – A Living Meta-Analysis*) therefore aims to provide an up-to-date and continuously expanding synthesis of experimental and quasi-experimental studies on the effects of generative AI on mathematics learning. As suggested by Elliot et al. (2014), frequent updates, regular literature searches, the inclusion of preprints, and a Bayesian multilevel meta-analytic approach, this project seeks to balance methodological rigor with the demands of a rapidly changing technological landscape. In doing so, it aims not only to estimate average effects but also to establish a foundation for future moderator analyses that can inform the pedagogical design and effective use of generative AI in mathematics education.

In this report, we address the following research question: What is the overall effect of instructional interventions using generative AI on mathematics learning outcomes?

As evidence grows over time, future versions of this study will also address which study-, intervention-, and context-related moderators may influence the effectiveness of generative AI-based interventions for learning mathematics. The following methods section already includes the analytic framework for these future analyses, but no moderator analyses are reported yet.

METHODS

Throughout the methods, we followed the PRISMA-LSR recommendations (Akl et al., 2024), which are an adapted form of the PRISMA statement for living research synthesis.

Literature Search

The literature search for this version of the analysis was conducted on December 1st, 2025, on SCOPUS for documents and preprints with the search string:

> ("generative AI" OR "generative Artificial Intelligence" OR genAI* OR "large language model*" OR LLM* OR "chatbot*" OR "AI tutor*" OR "AI assistant*" OR ChatGPT OR "Chat GPT") AND (educat* OR teach* OR instruct* OR pedagogy OR curriculum OR classroom OR student* OR school* OR "higher education" OR "K-12") AND (math* OR algebra OR geometry OR arithmetic OR calculus OR "word problem*" OR numeracy OR fraction* OR "quantitative reasoning") AND (intervention* OR experiment* OR "control group" OR RCT OR "randomized control*" OR treatment OR pretest OR posttest OR (pre* AND post*))

The database search identified 411 documents and 521 preprints. Abstracts of these 932 records were screened manually. Records were excluded if they: a) did not report an intervention study b) did not include a control group that did not use AI, c) did not

report mathematical performance as an outcome, d) did not use generative AI in the intervention or e) did not use human participants. For the remaining 45 records, full texts were retrieved (1 full text was not available) and assessed for eligibility. For 3 reports which were available both as preprints and published papers, the preprints were excluded. 14 studies reported insufficient data, and their authors did, as of this time, not react to requests for additional information. 14 studies were excluded due to the initial exclusion criteria. The remaining 13 studies were included in the analysis.

Moderator Coding

In the current version, the number of studies was still too small to allow for meaningful moderator analyses. However, we already developed a moderator framework following the guidelines by Lispey and Wilson (2000) and coded all studies accordingly. This framework comprises six categories: (A) metadata, (B) study methods and research design, (C) participant characteristics, (D) educational contextual factors, (E) intervention characteristics, and (F) outcome characteristics. At the current stage of the project, the number of studies is still insufficient for meaningful moderator analyses. Future versions will include meta-regressive models.

Effect Size Extraction

For each study, we calculated standardized mean differences (Hedges' g) in outcomes between the intervention and control group, using small-sample corrections. When pre-test scores were available and on the same scale as the post-test, we calculated g as the difference in gain, standardized by the pre-test standard deviation (Borenstein et al., 2009). If pre-test scores were not available but groups were reported to perform similar prior to the intervention, post-test differences, standardized by post-test standard deviations, were used. When means or standard deviations were unavailable, we transformed effect sizes from reported statistics. All effect sizes were coded such that positive values indicate a benefit of the intervention using generative AI. Sampling standard errors were estimated according to the procedures described by Borenstein et al. (2009) and corrected for within-study covariance.

Data Analysis

To estimate the overall effect, a Bayesian multilevel meta-analytic model was used (Harrer et al., 2022). We used the {brms} package (Bürkner, 2017) in *R*. The model specified effect sizes as nested within studies.

Prior Specification

Since Bayesian analyses require several arbitrary prior assumptions, we conducted a sensitivity analysis to examine the robustness of the pooled effect estimate with respect to different prior assumptions. Sensitivity analyses across a range of plausible prior specifications yielded similar posterior estimates for all estimated parameters, indicating that the conclusions are robust to reasonable variations in prior assumptions.

Model Estimation

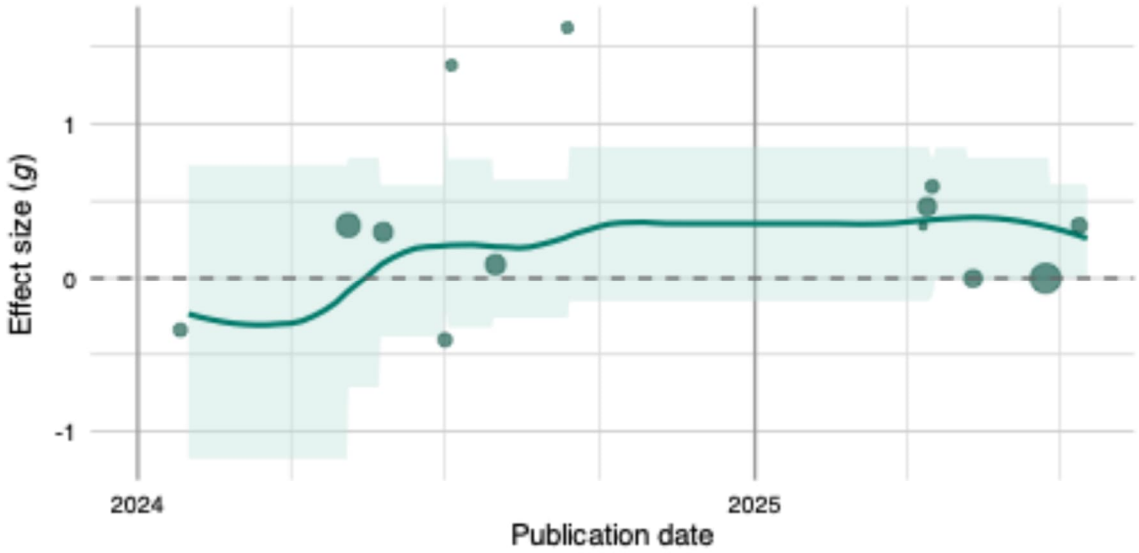
Model estimation was conducted using Monte Carlo sampling. Four independent Markov chains were run with sufficient iterations and warm-up samples to ensure stable posterior estimation. Posterior summaries are reported as posterior means along with 95% credible intervals, which represent the range within which the true effect lies with 95% posterior probability given the data and model assumptions.

RESULTS

Our analysis included 24 effect sizes nested within 13 studies. The posterior mean pooled effect was $g = 0.28$, with a 95% credible interval of [0.01, 0.59], indicating a small positive effect accompanied by substantial uncertainty. Sensitivity analyses for different priors did not result in substantial variability of these estimations.

Between-study heterogeneity, expressed as the posterior standard deviation of study-level effects, was estimated at 0.24 (95% CrI [0.01, 0.60]), suggesting considerable variability across studies with wide uncertainty. Additional variability among multiple effect sizes within studies was also observed ($SD = 0.46$, 95% CrI [0.27, 0.70]), supporting the use of a multilevel modelling approach.

Figure 1. Cumulative Bayesian meta-analysis over time. Study-level effect estimates (Hedges' g) are shown as points at their publication dates, with point size proportional to the effective sampling precision of each study, accounting for within-study dependence. The smoothed line and shaded region indicate the posterior median and 95% credible interval of the pooled effect as evidence accumulates.



DISCUSSION

The living meta-analysis LLAMA LIMA provides an ongoing synthesis of intervention studies on the effects of generative AI on mathematics learning. The initial analyses

reported here show a small positive average effect (g = 0.28) across 13 studies and 24 effect sizes, accompanied by wide credible intervals and substantial heterogeneity. Generative AI-based interventions may support mathematics learning, but the current evidence remains limited and does not yet permit robust or generalizable conclusions.

When compared to benchmarks such as Hattie's *hinge point* (d = 0.40; Hattie, 2008) or meta-analytic estimates for other interventions in mathematics education (e.g., digital media g = 0.55; Hillmayr et al., 2020; visualizations g = 0.50; Schoenherr et al., 2025), the observed effect appears modest. At the same time, the direction of the effect can be interpreted as cautiously optimistic given the early stage of empirical research in this field and the small number of available intervention studies.

The substantial heterogeneity across studies suggests that the effectiveness of generative AI in mathematics education is highly context-dependent and likely influenced by instructional design choices, pedagogical roles assigned to AI, and learner characteristics. The current number of studies did not yet justify meaningful moderator analyses and highlights the need for further rigorous experimental research. This situation also supports the choice of a living meta-analytic approach: on the one hand, reliable evidence is already needed to inform research and practice, while on the other hand, available evidence is still too limited to allow for more differentiated and stable conclusions about the effectiveness of generative AI in mathematics education.

Beyond its findings, this study makes a methodological contribution. While living meta-analyses have recently been proposed and exemplified in adjacent fields such as psychology (e.g., Howard & Slemp, 2025), their application in educational research remains rare. To our knowledge, LLAMA LIMA represents one of the first applications of a living meta-analysis in education, and the first focusing specifically on mathematics education. By enabling continuous updates of the evidence base, living meta-analyses offer a promising complement to conventional research syntheses in rapidly evolving fields. Future versions of LLAMA LIMA will include the expanding evidence base and allow for more differentiated analyses.

Notes

If you want to cite LLAMA LIMA, please consider checking for the latest published version on arXiv.org and use the permanent DOI associated with this publication. As a living meta-analysis, this study has the inherent limitation that it does not go through regular peer-review processes. We therefore appreciate any feedback, critique and suggestions that help to improve the study in the future.

References

Akl, E. A., Khabsa, J., Iannizzi, C., Piechotta, V., Kahale, L. A., Barker, J. M., McKenzie, J. E., Page, M. J., & Skoetz, N. (2024). Extension of the PRISMA 2020 statement for living systematic reviews (PRISMA-LSR): checklist and explanation. BMJ, 387, e079183.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). Introduction to Meta-Analysis. Chichester: Wiley.

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. Journal of Statistical Software, 80, 1-28.

Elliott, J. H., Turner, T., Clavisi, O., Thomas, J., Higgins, J. P., Mavergames, C., & Gruen, R. L. (2014). Living systematic reviews: an emerging opportunity to narrow the evidence-practice gap. PLoS Med, 11(2).

Fock, A., & Siller, H.-S. (2025). The Impact of Generative AI on student learning achievement in secondary STEM education: A meta-analysis. Proceedings of the 17th International Conference on Technology in Mathematics Teaching, London, UK.

Harrer, M., Cuijpers, P., Furukawa, T. A., & Ebert, D. D. (2021). Doing Meta-Analysis With R: A Hands-On Guide (1st ed.). Boca Raton, FL and London: Chapman & Hall/CRC Press.

Hattie, J. (2008). Visible learning (1st ed.). Routledge.

Hillmayr, D., Ziernwald, L., Reinhold, F., Hofer, S. I., & Reiss, K. M. (2020). The potential of digital tools to enhance mathematics and science learning in secondary schools: A context-specific meta-analysis. Computers & Education, 153, 103897.

Howard, J. L., & Slemp, G. R. (2025). Why we need a living meta-analysis of self-determination theory: an illustration examining temporal changes in need supportive education. Motivation and Emotion, 1-18.

Huisman, J., & Smits, J. (2017). Duration and quality of the peer review process: the author's perspective. Scientometrics, 113(1), 633–650.

Lipsey, M. W., & Wilson, D. B. (2000). Practical Meta-Analysis. Thousand Oaks, CA: Sage.

Pepin, B., Buchholtz, N., & Salinas-Hernández, U. (2025). A Scoping Survey of ChatGPT in Mathematics Education. Digital Experiences in Mathematics Education, 11(1), 9-41.

Schoenherr, J., Strohmaier, A. R., & Schukajlow, S. (2024). Learning with visualizations helps: A meta-analysis of visualization interventions in mathematics education. Educational Research Review, 45, 100639.

Strohmaier, A. R., Van Dooren, W., Seßler, K., Greer, B., & Verschaffel, L. (2025). Large Language Models Don't Make Sense of Word Problems. A Scoping Review from a Mathematics Education Perspective. arXiv preprint arXiv:2506.24006.

Tlili, A., Saqer, K., Salha, S. & Huang, R. (2025). Investigating the effect of artificial intelligence in education (AIEd) on learning achievement: A meta-analysis and research synthesis. Information Development, 41(3), 825–842.