LLAMA LIMA：生成式人工智能对数学学习影响的动态荟萃分析

版本 2，03/26

Anselm Strohmaier、Samira Bödefeld、Oliver Straser、Frank Reinhold 弗赖堡教育大学数学教育研究所

摘要。
数学教育中的生成式人工智能能力正在迅速发展，这给研究跟上这一进展带来了重大挑战。研究

动态荟萃分析声明

LLAMA LIMA 以版本化动态荟萃分析的形式持续维护，每次更新均构成一个新版本。本文档为版本 2，03/26。该元分析的更新版本将在正式退役并永久发表之前，通过 arXiv 公开发布。建议读者在引用或解读研究结果时查阅最新版本。本出版物尚未经过正式同行评审。作者欢迎反馈、批评和建议。

引用指南：
请在引用中将版本标注（版本 2，03/26）直接置于标题之后。泛指 LLAMA LIMA（例如在背景介绍或方法论中）时，可以省略版本标注。所有版本请始终使用同一个 DOI。请按以下格式引用本版本：

Strohmaier, A., Bödefeld, S., Straser, O., & Reinhold, F.（2026年）。LLAMA LIMA：生成式人工智能对数学学习影响的动态荟萃分析（版本 2，03/26）。arXiv。https://doi.org/10.48550/arXiv.2601.18685

更新状态：
持续进行中。本版本最后一次文献检索日期为2026年2月2日。下一次文献检索计划于2026年4月进行。下一版本计划于2026年5月发布。

与上一版本相比的变更：
Oliver Straser 现已成为共同作者。本版本新增了 6 项研究和 11 个新效应。分析、结果、图和表均已更新。新增了使用 RoBMA 进行的发表偏倚分析。引言中补充了参考文献。理论背景部分将措辞角色改为目的，并将目的“对话伙伴”纳入目的“自适应评估与辅导”。

CC BY

1 引言

近年来，大型语言模型 （大语言模型）取得了长足进展，例如 ChatGPT，推动生成式人工智能（生成式人工智能）在教育实践和研究中的迅速普及（Kasneci 等，2023；Ng 等，2025）。生成式人工智能是指包括大语言模型在内、基于从大规模训练数据中学习到的模式生成新内容的人工智能系统。

基于人工智能的技术长期以来一直受到教育领域的研究关注（例如智能辅导系统、自适应学习、自动反馈；Chiu 等，2023；Holmes 等，2019），但生成式人工智能加速并改变了这些努力。与许多早期技术相比，生成式人工智能支持开放式的自然语言交互，从而实现了新形式的教学支持，例如即时生成解释、提供自适应反馈、支持问题解决，或促进个体学习与协作学习；与此同时，这些系统通常对输出的可预测性有限，且其生成过程不透明，由此引发了人们对准确性、可解释性、偏差和教学控制的担忧（Kasneci 等，2023）。

尽管越来越多的研究考察了基于生成式人工智能的教育应用，但关于学习成果的实证证据仍主要处于探索阶段，且尚无定论。早期的荟萃分析表明，生成式人工智能对学习的效应存在显著差异，且高度取决于教学设计和情境因素（例如 Deng 等，2025；Ma 和 Zhong，2025；Wang 和 Fan，2025；Zhu 等，2025），并且由于存在显著的发表偏倚，这些效应可能被过度乐观地估计（Bartoš、Martinková 等，2025）。这些发现还表明，生成式人工智能的益处不太可能在不同教育领域中保持一致。数学学习具有领域特定的认知过程和教学实践特征（Schneider 和 Stern，2010；Schoe nfeld，2016）。因此，在数学教育中实施生成式人工智能的潜在益处与挑战，可能不同于其他领域中观察到的情况（Pepin 等，2025；Walkington，2025）。

这种普遍缺乏汇总实证证据的情况同样存在于数学教育领域（Fock 和 Siller，2025a、2025b；Pepin 等人，2025；Turmuzi 等，2026；Walkington，2025）。因此，两个核心问题仍未得到解答：（a）基于生成式人工智能的教学干预能否有效促进数学学习；（b）哪些因素决定其效果。虽然这两个问题都可以通过荟萃分析加以研究，但生成式人工智能技术及其研究格局快速演变，意味着传统的研究综合可能难以跟上变化。因此，我们通过动态荟萃分析来探讨这些问题。在 LLAMA LIMA（数学教育中的大型语言模型与生成式人工智能：动态荟萃分析）中，我们持续扩充数据库、更新分析结果，并即时公开研究结果。据我们所知，这是教育研究领域首个基于出版物的动态荟萃分析。

1.1 数学教育中的生成式人工智能实验研究

尽管生成式人工智能有望通过多种方式支持数学学习（Fock 和 Siller，2025a、2025b；Pepin 等，2025；Turmuzi 等，2026），但现有实证研究的

异质性很高，涵盖不同的教学设计、人工智能的教学目的、结果测量和学习者群体。许多研究采用探索性设计、短期干预以及范围狭窄的教学情境，这限制了单项研究发现的普遍适用性。在这些情况下，尽管生成式人工智能在教育实践中的应用日益广泛，且研究人员、教育工作者、政策制定者乃至学习者都对此抱有浓厚兴趣，单项研究往往仍不足以对整体有效性相关问题给出有力回答。一般而言，系统性研究综合和荟萃分析有助于整合并概述不同研究的发现。

然而，生成式人工智能的快速技术发展给现有证据的综合带来了挑战。开展并发表严谨的实证研究——并通过传统元分析对其进行总结——需要大量时间（Huisman & Smits，2017年）。许多关于教育中的人工智能的现有综述早于大型语言模型的广泛普及（例如，Tlili 等人，2025年），这说明了这一点。大型语言模型的普及始于 2022 年底 ChatGPT 的发布。较新的综合研究表明，随着模型能力不断演进，对证据的评估很快就会过时：例如，Pepin 等人于2025年2月发表的范围综述基于截至2024年5月的研究，讨论了 ChatGPT 数学表现方面的局限性，而这些局限性大多已通过后续模型版本得到改善。同样，Fock 和 Siller（2025b）于2025年11月发表的元分析基于2024年11月开展的文献检索，仅发现三项明确研究中学数学教育中生成式人工智能的实验研究。最近，Liu 等人（2026年）的元分析纳入了截至2025年9月发表的研究，共发现22项符合条件的研究。然而，该分析也纳入了学位论文和其他未经同行评审的文献，而且在某些情况下，纳入了不符合其所述纳入标准、没有对照组的研究。

这些例子表明，研究综合在发布时可能已无法完全反映技术现状。这凸显了生成式人工智能的快速发展与及时且富有信息量的研究综合之间存在的重大冲突和方法论挑战。

1.2 动态荟萃分析：应对快速发展的领域

尽管这种紧迫感在教育科学领域相对少见，但其他领域对快速发展的证据已更加熟悉。对此，医学领域已经发展出称为 动态证据综合（Elliott 等，2014）的研究方法：动态系统综述（LSRs）和动态荟萃分析（LIMAs）旨在通过持续更新证据库，缩短传统研究综合所需的时间滞后。

动态证据综合与传统研究综合相比，在几个关键方面有所不同 s:

- 持续发布：动态证据综合并非一次性发布的成果，而是旨在定期更新，从而反映证据不断演变的状况。

- 重复文献检索：定期开展检索，通常由自动化程序提供支持；新发表的研究会经过筛选和编码，并在尽可能短的时间内纳入其中。

- 纳入近期证据：为减少发表滞后，动态证据综合可以在纳入同行评审研究的同时，纳入预印本和会议论文集。这种做法需要在快速纳入新证据与遵循传统出版质量筛选标准之间权衡。

- 累积统计建模：传统的频率学派统计方法不适合逐步整合新证据，因为反复更新可能会增加错误率。贝叶斯荟萃分析模型是动态荟萃分析的自然选择，因为它们为累积更新证据提供了连贯的框架（Elliott 等，2014）。在这些模型中，现有证据被视为先验分布，并随着新研究的出现而更新，从而能够直接用感兴趣参数的可信区间来表示不确定性。

这些特点使动态荟萃分析尤其适用于教育领域的生成式人工智能等快速发展的研究领域。

1.3 数学教育中生成式人工智能的目的

为了梳理生成式人工智能支持数学学习的各种目的，我们参考了三篇近期文献综述（Pepin 等，2025；Turmuzi 等，2026；Walkington，2025）以及 Wang 和 Fan（2025）的元分析。我们提出了五个类别，用以描述生成式人工智能支持学生数学学习的潜在目的。

生成式人工智能作为数学专家。 生成式人工智能系统能够针对广泛的与学校课程相关的数学任务生成正确答案和完整解答（例如，Getenet，2024；Plevris 等，2023；Strohmaier 等，2025；Udias 等，2024）。这种类似专家的能力可以让学生检查自己的解答并比较不同的解法，从而支持数学学习。在这一用途下，生成式人工智能可以像增强型计算器一样使用：它不仅能进行数值计算，还能支持解答生成和问题解决，但不提供教学指导。

生成式人工智能用于自适应评估与辅导。 生成式人工智能系统还可用于分析学生解答和学习过程，并提供自动评估、反馈和个性化学习路径（例如，Chiu 等，2023；Giannakos 等，2025；Kasneci 等，2023；Pepin 等，2025；Wardat 等，2023）。这可以从导师或对话伙伴的角度进行，也可以不赋予其社会角色。与 ChatGPT 等通用系统相比，经过专门调整或融入学生个人学习进度、目标和先前知识信息的生成式人工智能工具，可能更有效地支持自我调节学习，并促进学生持续投入数学内容（Turmuzi 等，2026）。

生成式人工智能作为指导者。生成式人工智能可用于提供非个性化教学。为此，系统可以提供解释、逐步指导，

或推荐学习路径（例如，Pepin 等，2025；Schorcht 等，2024；Turmuzi 等，2026；Wu 等，2024）。如果系统参考典型的学习轨迹、概念性想法和错误观念，以及有关数学内容和相关教育理论的信息，其教学内容就可能与课程目标、评分标准和评估要求高度一致（Turmuzi 等，2026）。

生成式人工智能作为协作学习的促进者。生成式人工智能还可以通过促进协作活动（如小组问题解决或项目式学习）进一步支持数学学习（Pepin 等，2025；Turmuzi 等，2026）。在这些情境中，可以利用人工智能系统帮助组织协作、支持围绕数学思想展开讨论，或为任务中的协同作业提供支架。只要这种促进作用能够提升同伴互动和数学交流的质量，就可能有助于取得积极的学习成果。支架式教学已被证明是协作学习的重要促进因素，在数学领域也是如此（Kollar 等，2014）。

生成式人工智能作为教师支持。最后，生成式人工智能还可以通过协助教师进行教学规划和专业学习，间接支持数学学习。Buchholtz 和 Huget（2024年）指出，ChatGPT 等系统可用于生成教案并为教师提供反馈。只要这些用途有助于提升教学设计质量，或更好地使任务、解释与学生的学习需求相匹配，就可能转化为学生学习成果的提升。与此同时，人工智能生成的教学材料质量参差不齐，也存在生成内容泛泛而谈或缺乏情境针对性的教学资源的风险（Giannakos 等，2025）。更先进的生成式人工智能系统或许可以根据课程材料或教学约束进行配置，从而提升其对教师支持的实用性。

1.4 生成式人工智能在数学学习中的有效性决定因素

前文综述指出，生成式人工智能可以出于截然不同的目的融入数学教育。尽管这些目的从理论上为预期基于生成式人工智能的干预措施可能影响数学学习结果提供了依据，但这也意味着，其效应可能因学习者特征、教育情境、人工智能实施方式以及目标成果而有很大差异。因此，我们有必要超越“生成式人工智能平均而言是否有效”这一问题，系统探究它在何种条件下能够支持——或无法支持——数学学习。下文将概述其中一些重要影响因素。

学习者特征。基于生成式人工智能的干预措施的效应可能因学习者特征而异。教育技术研究表明，学生年龄和教育阶段等情境因素会调节教学支持的有效性；自适应支架式数字工具在与学习者现有能力相匹配时效果最佳（Hillmayr 等，2020）。在数学教育中，学生因素尤为重要，因为这些因素可能影响学生对生成式人工智能所提供的解释、反馈和自我调节学习机会的投入度（Pepin 等，2025）。因此，学习者

特征被视为潜在的调节变量，用于考察不同学生群体之间的效应是否存在差异。

教育情境因素。除个体学习者之外，使用生成式人工智能的教育情境也可能调节其有效性。相关情境因素包括所涉及的数学内容：例如，理解和处理文字应用题可能是大语言模型更擅长的任务，而几何题则需要不同的技术能力和非语言数学概念，例如处理视觉表征；这可能使解题本身以及将其迁移到教育情境中都成为人工智能面临的挑战。此外，学习情境（如课堂、在线或在家）以及教学环境（如个人学习、协作学习或教师指导学习）可能会影响生成式人工智能支持学习的有效性。

人工智能实施。 采用生成式人工智能的干预措施，在设计和实施方面存在很大差异。相关特征包括人工智能的社会角色、预期用途（见第 1.3 节）、学习者自主性的程度、干预的持续时间和强度，以及生成式人工智能是作为现有教学的补充还是替代。这些设计特征是技术增强型学习理论的核心（例如，Reinhold 等，2024），也可能解释不同研究中观察到的效应差异。

可能调节生成式人工智能干预效果的另一个特征，是指导其设计的底层学习理论。在不同研究中，生成式人工智能可能被融入不同的教学范式，例如直接教学、基于问题的学习或探究式学习、合作学习或同伴学习、元认知支架，或概念转变方法；这些范式会影响学习者与系统的互动方式，以及系统所支持的认知过程类型（例如，Hattie 和 Yates，2014）。

结果特征。 最后，观察到的效应可能取决于数学学习成果的概念界定与测量方式。结果特征包括所评估的数学知识类型、评估任务的形式，以及干预与结果测量之间的一致性。远端结果测量可能反映更广泛的学习目标，而近端测量通常对短期教学效应更敏感（Miller 等，2025）。就知识类型而言，生成式人工智能或许特别有助于支持再现和应用技能，而促进概念理解与迁移则可能更具挑战性。

2 本研究

基于生成式人工智能在数学学习中的潜力、实证研究发现的碎片化，以及快速演进领域中传统静态研究综合的局限性，本研究采用动态荟萃分析方法来整合现有证据。LLAMA LIMA（数学领域的大语言模型与生成式人工智能：动态荟萃分析）旨在对考察生成式人工智能对数学学习结果影响的干预研究进行最新且持续扩展的综合。依照关于动态

研究综合的建议（Elliot 等人，2014年），我们采用频繁更新、定期文献检索、纳入预印本以及贝叶斯多层次元分析方法。我们遵循标准化报告规范（PRISMA-LSR；Akl 等人，2024）。除估计平均效应外，LLAMA LIMA 的后续版本还将纳入系统性调节变量分析，以考察研究特征、教学设计和情境因素如何影响基于生成式人工智能的干预措施的有效性；待可用研究数量足以支持此类分析后，我们便会开展这些分析。

因此，LLAMA LIMA 聚焦于以下研究问题：

RQ1：采用生成式人工智能的教学干预对数学学习结果的总体影响是什么？

RQ2：哪些与学习者、研究、干预和情境相关的调节变量会影响基于生成式人工智能的数学学习干预的有效性？

3 方法

3.1 动态模式参数

本动态荟萃分析计划每两个月更新一次文献检索（即下一次数据库检索安排在2026年4月），并在交替月份更新发表内容（即下一个版本安排在2026年5月）。根据新发表文献的频率及其对总体效应和调节变量分析可行性的影响，这些时间间隔未来可能会进行调整。

在后续版本中，先前版本中被排除的报告可能会被纳入，条件是：(a) 先前未能检索到的报告现已可获取，或 (b) 因报告的数据不足而被我们联系的作者向我们提供了额外数据。

从版本 3起，将纳入调节变量分析。

计划最终将该研究从动态模式中退役，并作为永久版本发布，但截至目前，尚未预先设定时间表。

3.2 文献检索

当前版本动态荟萃分析的文献检索于2026年2月2nd进行，使用 SCOPUS 检索文献和预印本。检索针对数学教育中使用生成式人工智能的实验研究和准实验研究，并采用预先设定的、涉及生成式人工智能、教育、数学和干预设计的术语组合。完整的检索式见附录A.2

本版本的检索共确定了175条新记录。在筛选标题和摘要后，获取了17条新记录进行全文评估。其中1条记录的文本无法获取。经过资格评估并删除重复报告后，6项新研究符合纳入标准，并被纳入元分析。有关所有版本的研究识别、筛选、

资格评估和纳入流程的详细信息，均以图1中的PRISMA-LSR流程图进行了总结。

在筛选过程中，纳入的研究需满足以下条件：a) 报告来自实验或准实验干预研究的原始数据；b) 在干预中使用生成式人工智能，而对照组不使用生成式人工智能；c) 涉及人类学习者；d) 将数学表现作为结果测量进行报告；以及 e) 使用英语撰写。我们纳入了发表于同行评审期刊、编著书籍章节和会议论文集中的研究，以及在筛选时正在接受同行评审的预印本。未纳入学位论文、博士论文、报告及其他未接受同行评审的未发表材料。

3.3 调节变量编码

在当前版本中，由于可用研究数量有限，尚未纳入调节变量分析。不过，我们已依据 Lipsey 和 Wilson（2000）的指南制定了调节变量框架，并据此对所有研究进行了编码。研究按照四个主要调节变量领域进行编码：参与者特征、教育情境因素、干预特征和结果特征。此外，还对研究元数据和方法学特征进行了编码。为确保准确性，本文的两位作者使用预先制定的编码手册，分别独立对所有研究进行编码。对于存在的差异，作者通过讨论达成共识并予以解决。

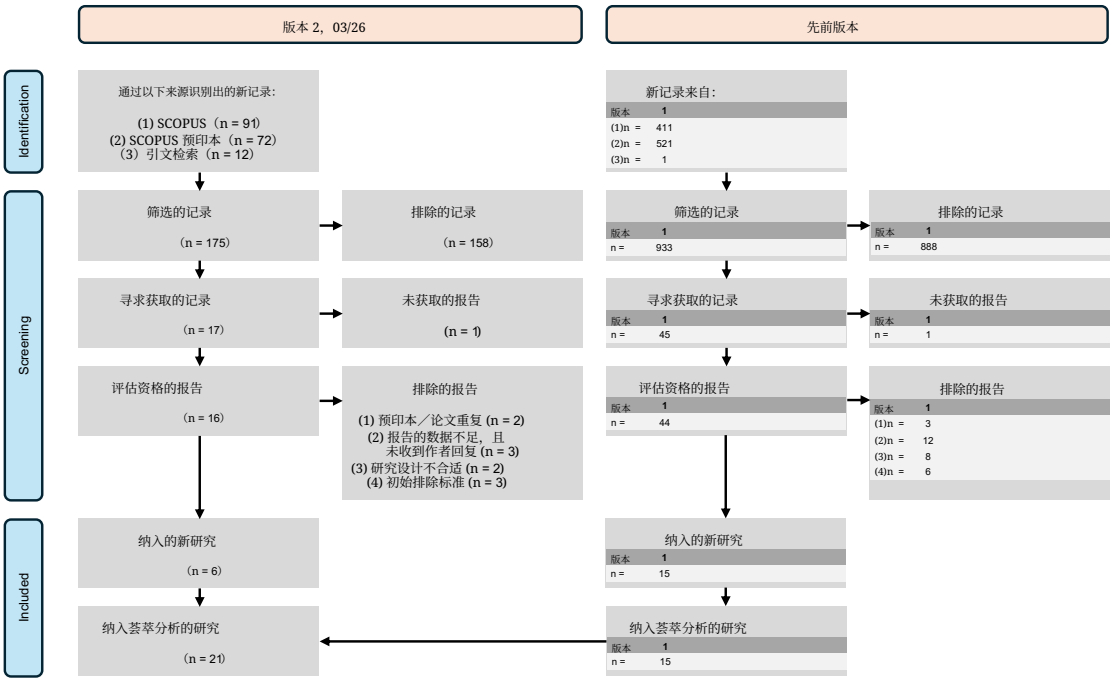


图1。记录该活体荟萃分析研究选择过程的 PRISMA-LSR 流程图。

参与者特征。参与者特征包括学习者依据《国际教育标准分类》（ISCED；联合国教科文组织，2012）划分的教育阶段、年龄、学校类型、地理位置，以及关于先前知识的信息（涉及数学、生成式人工智能和一般知识），还包括社会经济背景。

教育情境因素。 教育情境因素包括数学领域（按 PISA 研究采用的领域分类：数量、空间与图形、变化与关系、不确定性与数据；OECD，2006）、学习安排（如个人学习、合作学习、教师指导）、教学情境（如课堂、在线、实验室）以及评估情境（如低风险、高风险）。

干预特征。 干预特征包括生成式人工智能的预期用途（见第 1.3 节）、教师对生成式人工智能系统的自主权、学生对生成式人工智能的熟悉程度、熟悉该系统所需的时间，以及干预的持续时间和强度。此外，还记录了人工智能系统本身的特征，例如人工智能工具的类型、实施方式，以及其使用的是基础版本还是为该研究修改过的版本。

结果特征。 结果特征包括评估的数学知识类型（如程序性或概念性知识，以及回忆、应用或迁移）、评估任务的形式，以及干预与结果测量之间的一致性（近端与远端）。

元数据和方法学特征。 研究元数据包括发表日期、格式和科学领域等信息。方法学特征包括设计特点（如随机化）、样本量以及对照条件类型（如主动、被动或安慰剂）。

3.4 效应量提取

对于每项研究，我们使用小样本校正，计算干预组与对照组在结果指标上的标准化均值差（Hedges' *g*）。如果有前测分数，且其量表与后测相同，我们就将增益差异除以前测标准差，计算 *g*（Borenstein 等人，2009）。如果没有前测分数，但研究报告或设计隐含两组具有等同性，则使用后测均值差，并除以后测标准差进行标准化。如果未提供均值或标准差，则根据报告的统计数据转换效应量。所有效应量的编码方式均为：正值表示生成式人工智能干预带来益处。

根据 Borenstein 等人（2009）所述的程序，估算了每个效应的抽样标准误。为恰当地对研究内抽样误差的协方差进行建模，我们参照 Viechtbauer（2010）估算了抽样误差的完整方差—协方差矩阵，其中，同一组内各测量值之间的估计相关系数为 $\rho = .7$，不同时间点多个测量值之间的估计自相关系数为 $\varphi = .8$（采用 $0 \le \rho \le .9$ 和 $0 \le \varphi \le .9$ 进行的敏感性分析表明，这两个参数的取值对总体效应均无显著影响，$0.41 < g < 0.46$）。

3.5 数据分析

贝叶斯多层次荟萃分析。为估算基于生成式人工智能的教学干预对数学学习结果的总体效应，我们拟合了一个贝叶斯多层次元分析模型（Harrer 等人，2021）。所有分析均使用 R 和 *brms* 包（Bürkner，2017）完成。该模型将效应量视为嵌套于各项研究之中，以此处理数据的层级结构。

先验设定与敏感性分析。所有模型参数均设定了弱信息先验。合并平均效应采用以零为中心、方差为 1 的正态先验。研究间和研究内异质性参数采用限制为正值的指数先验。由于使用了预先设定的抽样误差方差—协方差矩阵，因此残差方差参数固定为 1。

为评估结果对这一先验选择的稳健性，我们针对替代先验设定进行了敏感性分析。具体而言，我们改变了合并效应先验的宽度（采用更窄和更宽的正态分布），也改变了异质性先验的分布形式和尺度（指数分布、半正态分布和半 t 分布）。此外，我们还采用了 Hattie（2012）报告的经验效应量分布作为探索性先验。在所有先验设定下，合并效应和异质性参数的后验估计都相近，范围为 $0.40 < g < 0.43$。

模型估计与推断。模型参数采用马尔可夫链蒙特卡罗抽样进行估计。运行了 4 条独立链，每条链包含 1000 次预热迭代和 3000 次抽样迭代。通过检查 R 和有效样本量来评估收敛情况，结果表明模型已充分收敛。通过目视检查密度叠加图来评估模型拟合情况，结果表明观测分布与模型推断分布相符。结果以后验均值及其 95% 可信区间报告。这些区间表示在给定观测数据和模型假设的情况下，参数以 95% 的后验概率落入的范围。

4 结果

4.1 描述性统计

该元分析纳入了 21 项研究，共提供 38 个效应量。这些研究共包含 4,071 名参与者。附录中的表 A.1 概述了纳入的研究及其基本研究特征。

4.2 合并效应量

我们的分析纳入了嵌套于 21 项研究中的 38 个效应量。后验平均合并效应为 $g = 0.42$，95% 可信区间为 [0.13, 0.72], ，表明存在正向效应，但同时伴有较大的不确定性

。针对不同先验的敏感性分析显示，这些估计值并未出现较大变异。

研究间异质性以研究层面效应的后验标准差表示，估计值为 0.28（95% 可信区间 [0.01, 0.67]），这表明各研究之间存在相当大的变异，且不确定性较高。研究内多个效应量之间也观察到额外变异（标准差 = 0.71，95% 可信区间 [0.52, 0.96]），支持采用多层建模方法。

4.3 调节变量分析

调节变量分析将在版本 3 中纳入。

4.4 发表偏倚

我们使用在 *RoBMA* 中实现的多层稳健贝叶斯模型平均元分析框架评估发表偏倚（Bartoš & Maier, 2020年；Bartoš、Maier 等，2025年）。该方法对纳入和未纳入发表偏倚校正的模型进行平均，并通过贝叶斯因子量化证据。偏倚成分的纳入贝叶斯因子表明，反对存在发表偏倚的证据较弱（BF = 0.65；后验概率 = .39）。模型平均效应估计值（g =0.29，95% 置信区间 [0.00，0.68]）略低于我们模型所得的估计值，这主要反映了对零效应模型进行模型平均，而非强有力的发表偏倚校正。由于 *RoBMA* 目前不支持纳入完整的抽样方差—协方差矩阵，因此本分析未对相关抽样误差进行建模，也未将这种方法用于所有分析。

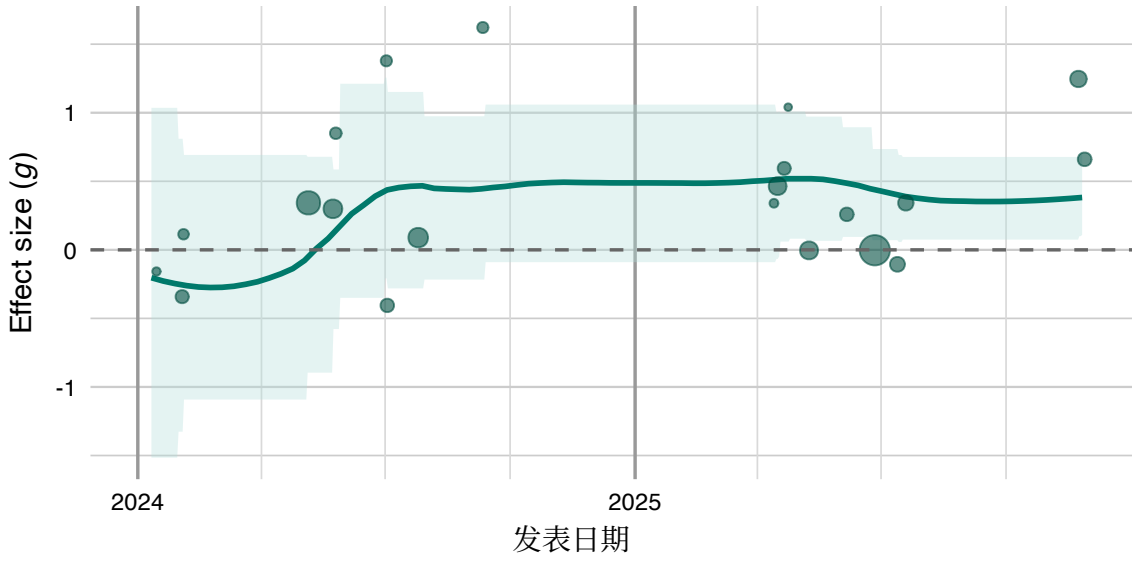


图 2。 随时间变化的累积贝叶斯元分析。研究层面的效应估计值（Hedges' g）以点表示，标示于各自的发表日期上；点的大小与各研究的有效抽样精度成正比，并已考虑研究内依赖性。平滑曲线及阴影区域表示随着证据不断累积，合并效应的后验中位数和 95% 可信区间。

4.5 结果的变化

与版本1相比，我们新增了6项研究，共贡献14个效应。后验平均合并效应增加了 $\Delta g = 0.11$。这不应被解读为一种趋势，因为部分新纳入的研究早在2024年就已发表。

5 讨论

LLAMA LIMA 持续综合了利用生成式人工智能支持数学学习的干预研究。我们的分析显示，在 21 项研究和 38 个效应量中，平均效应略呈正向（$g = 0.42$）。结合较宽的可信区间和显著的异质性，这表明基于生成式人工智能的干预可能有助于数学学习；但根据目前的证据，其效应尚不足以得出稳健或可推广的结论。为了解效应大小，可参考数学教育中其他干预措施的元分析估计值（例如，数字媒体 $g = 0.55$；Hillmayr 等，2020；可视化 $g = 0.50$；Schoenherr等，2025），这些估计值表明，目前该效应相对较小。关于非数学领域的研究结果，Wang 和 Fan（2025）报告称，使用 ChatGPT 对学习表现的平均效应显著较高，为 $g = 0.87$，但这一结果可能受到发表偏倚的显著影响（Bartoš、Martinková 等，2025）。Hattie 的关键点（$d = 0.40$；Hattie，2008）也可作为基准。不过，需要注意的是，这一效应量通常直接来自前后测比较。相比之下，在我们的元分析中，效应量是通过比较干预组与对照组的增益差异来确定的。

不同研究中的效应存在显著异质性，这表明生成式人工智能在数学教育中的有效性高度取决于情境因素。在当前版本中，可用研究数量有限，尚不足以通过调节变量分析系统地考察这些影响因素。尽管如此，现有证据和理论框架已表明，将生成式人工智能引入数学课堂时，需要仔细考虑教学设计、情境和学习者特征，不能指望它必然带来学习增益。与此同时，现有概念框架和既往研究综合指出，生成式人工智能有望通过多种途径支持数学学习。

此外，本研究展示了动态荟萃分析方法在快速发展的研究领域中综合证据的价值。虽然动态荟萃分析已在其他若干学科中得到应用，但据我们所知，教育领域唯一现有的实践是以持续更新的数据库形式提出的（Howard和Slemp，2025）。与之不同，本研究采用了版本化动态荟萃分析，与既有的出版实践相契合，展现了在教育领域实施持续更新的证据综合的另一种互补方式。

5.1 局限性

作为一项版本化动态荟萃分析，所有结果都是有条件的，并可能随着新研究的出现而变化。此外，生成式人工智能的功能、界面和易用性发展迅速，因此不同时期的效应可能无法直接比较。我们通过透明地公布更新内容和版本来应对这一局限，并遵循此类出版形式的相关建议（Akl 等人，2024）。纳入预印本可以缩短发表滞后，但也可能纳入未经严格同行评审的证据。不过，目前纳入的21项研究中，只有7项来自预印本服务器，且这些研究均曾作为会议论文发表，推测经过了一定程度的审查。与此同时，目前有14项研究仅因未报告所有实验组的均值和标准差而未被纳入，这或许表明该研究领域的方法学严谨性并非总是充分。

5.2 结论

我们的分析表明，生成式人工智能系统可能具有支持数学学习的潜力。与此同时，相较于文献和公共话语中所讨论的广泛预期与拟议应用，目前有关有效实施方案、其前提条件及情境条件的实证证据仍然有限。因此，除了持续开展系统性评估之外，还需要进一步进行严谨研究，以明确生成式人工智能能否切实促进数学学习，以及在何种条件下能够做到这一点。借助 LLAMA LIMA，我们持续观察并评估这一发展。

6 参考文献

Akl, E. A., Khabsa, J., Iannizzi, C., Piechotta, V., Kahale, L. A., Barker, J. M., McKenzie, J. E., Page, M. J., & Skoetz, N. (2024). 动态系统综述的 PRISMA 2020 声明扩展版（PRISMA-LSR）：核对清单与说明。《英国医学杂志》，387，e079183。https://doi.org/10.1136/bmj-2024-079183 Bartoš, F., & Maier, M. (2020). RoBMA：用于稳健贝叶斯荟萃分析的 R 软件包。载于 https://CRAN.R- project.org/package=RoBMA Bartoš, F., Maier, M., & Wagenmakers, E.-J. (2025). 稳健贝叶斯多层荟萃分析：在存在依赖效应量的情况下校正发表偏倚。PsyArXiv。https://doi.org/10.31234/osf.io/9tgp2_v1 Bartoš, F., Martinková, P., Wagenmakers, E.-J., & Bartoš, F. (2025). 校正发表偏倚后，未发现 ChatGPT 对学生学习表现、学习感知和高阶思维的影响证据。载于：PsyArXiv。Bastani, H., Bastani, O., Sungu, A., Ge, H., Kabakcı, Ö., & Mariman, R. (2025). 缺乏护栏的生成式人工智能可能损害学习：来自高中数学的证据。《美国国家科学院院刊》，122(26)，e2422633122。Bešlić, A., Bešlić, J., & Kamber Hamzić, D. (2024). 小学数学教育中的人工智能：分析对学生成就的影响。数字化转型与人工智能应用国际会议，Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). 《元分析导论》。Wiley。Buchholtz, N., & Huget, J. (2024). ChatGPT 作为反思工具：提升职前教师的课程规划能力。第17届 ERME 专题会议 MEDA4 会议论文集，2024年9月3日至6日，意大利巴里，Bürkner, P.-C. (2017). brms：使用 Stan 进行贝叶斯多层模型分析的 R 软件包。Journal of Statistical Software, 80(1)，1 - 28。https://doi.org/10.18637/jss.v080.i01 Canonigo, A. M. (2024). 运用人工智能提升学生对数学的概念理解和信心。Journal of computer assisted learning，40(6)，3215-3229。Cheng, L., Croteau, E., Baral, S., Heffernan, C., & Heffernan, N. (2024). 在在线数学学习平台中利用聊天机器人促进学生学习。Journal of Educational Computing Research，62(4)，687-717。Chiu, T. K., Xia, Q., Zhou, X., Chai, C. S., & Cheng, M. (2023). 关于教育中的人工智能所带来的机遇、挑战及未来研究建议的系统性文献综述。Computers and Education: Artificial Intelligence，4，100118。

Dasari, D., Hendriyanto, A., Sahara, S., Suryadi, D., Muhaimin, L. H., Chao, T., & Fitriana, L. (2024). 教学四面体中的 ChatGPT：它是否构成例外？数学教学与学习案例研究。Frontiers in Education，Deng, R., Jiang, M., Yu, X., Lu, Y., & Liu, S. (2025). ChatGPT 能否提升学生学习？实验研究的系统综述与荟萃分析。Computers & Education，227，105224。https://doi.org/10.1016/j.compedu.2024.105224 El-Shara, I. A., Tabieh, A. A., & Helu, S. Y. A. (2025). 使用 MatGPT 对本科生数学能力的影响。International Journal of Information and Education Technology，15(4)。Elliott, J. H., Turner, T., Clavisi, O., Thomas, J., Higgins, J. P., Mavergames, C., & Gruen, R. L. (2014). 动态系统综述：缩小证据与实践差距的新机遇。PLoS Med，11(2)。https://doi.org/10.1371/journal.pmed.1001603 Fardian, D., Suryadi, D., Prabawanto, S., & Jupri, A. (2025). 在课堂中融入 Chat-GPT：高等教育线性代数学习研究。International Journal of Information and Education Technology，15(4)。Fock, A., & Siller, H.-S. (2025a). 人类繁荣视角下中学 STEM 教育中的生成式人工智能：范围综述。Research Square（预印本）。https://doi.org/10.21203/rs.3.rs-6923010/v1 Fock, A., & Siller, H.-S. (2025b). 生成式人工智能对中学 STEM 教育学生学习成就的影响：荟萃分析。第17届数学教学技术国际会议论文集，英国伦敦。Getenet, S. (2024). 职前教师与 ChatGPT 在多策略问题解决中的应用：对小学数学教学的启示。International Electronic Journal of Mathematics Education，19(1)。Giannakos, M., Azevedo, R., Brusilovsky, P., Cukurova, M., Dimitriadis, Y., Hernandez-Leo, D., Järvelä, S., Mavrikis, M., & Rienties, B. (2025). 生成式人工智能在教育中的前景与挑战。Behaviour & Information Technology，44(11)，2518-2544。Harrer, M., Cuijpers, P., Furukawa, T. A., & Ebert, D. D. (2021).《使用 R 进行荟萃分析：实践指南》（第1版）。Chapman & Hall/CRC Press。https://www.routledge.com/Doing-Meta-Analysis-with-R-A-Hands-On-Guide/Harrer-Cuijpers-Furukawa-Ebert/p/book/9780367610074 Hattie, J. (2012).《教师可视化学习：最大化学习影响》。Routledge。Hattie, J., & Yates, G. C. R. (2014).《可视化学习与学习科学》。Routledge。Henkel, O., Horne-Robinson, H., Kozhakhmetova, N., & Lee, A. (2024). 有效且可扩展的数学支持：人工智能数学辅导工具在加纳的影响实验证据。人工智能教育国际会议，Hillmayr, D., Ziernwald, L., Reinhold, F., Hofer, S. I., & Reiss, K. M. (2020). 数字工具促进中学数学与科学学习的潜力：情境特定荟萃分析。Computers & Education，153，103897。https://doi.org/https://doi.org/10.1016/j.compedu.2020.103897 Holmes, W., Bialik, M., & Fadel, C. (2019). 教育中的人工智能：对教学与学习的承诺及影响。Center for Curriculum Redesign。Karaman, M. R. (2024). ChatGPT 制定的教案是否更有效？一项实验研究。International Journal of Technology in Education，7(1)，107-127。https://doi.org/10.46328/ijte.607 Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., . . . Kasneci, G. (2023). 善用 ChatGPT？大型语言模型在教育中的机遇与挑战。Learning and Individual Differences，103，102274。https://doi.org/10.1016/j.lindif.2023.102274 Kollar, I., Ufer, S., Reichersdorfer, E., Vogel, F., Fischer, F., & Reiss, K. (2014). 协作脚本和启发式例题对不同先前学业成绩师范生数学论证能力习得的影响。Learning and Instruction，32，22–36。https://doi.org/10.1016/j.learninstruc.2014.01.003 Kretzschmar, V., & Seitz, J. (2024). 人工智能学习助手：提升德国学校移民学生的数学学习。2024 IEEE人工智能会议（CAI），Lademann, J., Henze, J., & Becker-Genschow, S. (2025). 利用人工智能定制聊天机器人增强学习环境：对学习表现、认知负荷和情感变量的影响。Physical Review Physics Education Research，21(1)，010147。Lipsey, M. W., & Wilson, D. B. (2000).《实用荟萃分析》。Sage。Liu, B., Zhang, W., & Wang, F. (2026). 生成式人工智能能否有效提升学生数学学习成果？——对2023至2025年实证研究的荟萃分析。Education Sciences，16(1)，140。Liu, Y., Zha, S., Zhang, Y., Wang, Y., Zhang, Y., Xin, Q., Nie, L. Y., Zhang, C., & Xu, Y. (2025). BrickSmart：利用生成式人工智能支持儿童在家庭积木游戏中的空间语言学习。2025年CHI计算系统人为因素会议论文集，Ma, N., & Zhong, Z. (2025). 生成式人工智能对学习成果影响的荟萃分析。Journal of Computer Assisted Learning，41(5)，e70117。Miller, A. H., Espinas, D. R., McNeish, D., & Barnes, M. A. (2025). 针对有数学学习障碍或有患病风险的小学低年级学生开展强化数学干预的剂量反应。Educational Psychology Review，37(4)，91。Ng, D. T. K., Chan, E. K. C., & Lo, C. K. (2025). 将生成式人工智能融入教育的机遇、挑战与学校策略。Computers and Education: Artificial Intelligence，100373。OECD. (2006). 科学、阅读与数学素养评估：PISA 2006 框架。PISA，OECD Publishing。https://doi.org/10.1787/9789264026407-en

Pardos, Z. A., & Bhandari, S. (2024). ChatGPT生成的帮助在数学技能方面带来的学习增益，与人工导师撰写的帮助相当。PLoS ONE, 19(5), e0304013. Pepin, B., Buchholtz, N., & Salinas-Hernández, U. (2025). 数学教育中ChatGPT应用的范围调查。Digital Experiences in Mathematics Education, 11(1), 9-41. https://doi.org/10.1007/s40751-025-00172-1 Plevris, V., Papazafeiropoulos, G., & Jiménez Rios, A. (2023). 将聊天机器人置于数学与逻辑问题的考验中：比较并评估ChatGPT-3.5、ChatGPT-4和Google Bard。Ai, 4(4), 949-969. Reinhold, F., Leuders, T., Loibl, K., Nückles, M., Beege, M., & Boelmann, J. M. (2024). 解释教育情境中使用数字工具学习的学习机制：认知过程框架。Educational Psychology Review, 36(1), 14. Rücker, C. R., & Becker-Genschow, S. (2025). 借助人工智能提升对STEM教育的热情：将特定领域聊天机器人用作个性化学习助手。Computers and Education Open, 100315. https://doi.org/10.1016/j.caeo.2025.100315 Schneider, M., & Stern, E. (2010). 学习的认知视角：十项基石性发现。载于 O. f. E. C.-O. a. D. (OECD) (编), 学习的本质：运用研究启发实践 (第69-90页)。OECD. Schoenfeld, A. H. (2016). 学会数学思考：数学中的问题解决、元认知与意义建构（再版）。Journal of education, 196(2), 1-38. Schorcht, S., Buchholtz, N., & Baumanns, L. (2024). 提示问题——通过比较提示技术，探究人工智能辅助问题解决的数学教育质量。Frontiers in Education, Serrano Heredia, A., del Pilar García-Chitiva, M., Camacho-Zúñiga, C., Mirabal, L. F. M., & Vázquez-Villegas, P. (2025). 通过基于人工智能的自主学习教育干预，弥补工程专业学生的数学知识不足。2025 IEEE全球工程教育大会（EDUCON），Steinbach, M., Bhandari, S., Meyer, J., & Pardos, Z. A. (2025). 大语言模型产生幻觉时：考察数学辅导系统中错误反馈的效应。第十二届ACM Learning@ Scale会议论文集，Strohmaier, A. R., Van Dooren, W., Seßler, K., Greer, B., & Verschaffel, L. (2025). 大型语言模型无法理解文字应用题：从数学教育视角开展的范围综述。arXiv预印本 arXiv:2506.24006. Tlili, A., Saqer, K., Salha, S., & Huang, R. (2025). 探究教育中的人工智能（AIEd）对学习成就的影响：元分析与研究综合。Information Development, 02666669241304407. Turmuzi, M., Azmi, S., & Kertiyani, N. M. I. (2026). 学校数学教育中的ChatGPT：机遇、挑战与教学启示的系统综述。Teaching and Teacher Education, 170, 105286. Udias, A., Alonso-Ayuso, A., Alfaro, C., Algar, M. J., Cuesta, M., Fernández-Isabel, A., Gómez, J., Lancho, C., Cano, E. L., & de Diego, I. M. (2024). ChatGPT在大学数学入学考试中的表现。International Electronic Journal of Mathematics Education, 19(4), em0795. 联合国教科文组织。(2012). 国际教育标准分类（ISCED）2011。UNESCO. Utami, I. Q., Hwang, W.-Y., & Hariyanti, U. (2024). 运用生成式预训练变换器，在真实情境中生成情境化、个性化的数学文字题及其对几何学习的影响。Journal of Educational Computing Research, 62(6), 1384-1419. https://doi.org/10.1177/07356331241249225 Viechtbauer, W. (2010). 使用metafor软件包在R中开展荟萃分析。Journal of Statistical Software, 36(3), 1 - 48. https://doi.org/10.18637/jss.v036.i03 Wahba, F., Ajlouni, A. O., & Abumosa, M. A. (2024). 基于ChatGPT的统计学学习对本科生统计推理和统计学态度的影响。Eurasia Journal of Mathematics, Science and Technology Education, 20(7), em2468. Walkington, C. (2025). 生成式人工智能对数学教育的影响。School Science and Mathematics. Wang, J., & Fan, W. (2025). ChatGPT对学生学习表现、学习感知和高阶思维的影响：元分析的见解。Humanities and Social Sciences Communications, 12(1), 1-21. https://doi.org/10.1057/s41599-025-04787-y Wardat, Y., Tashtoush, M. A., AlAli, R., & Jarrah, A. M. (2023). ChatGPT：一种用于数学教学与学习的革命性工具。Eurasia Journal of Mathematics, Science and Technology Education, 19(7), em2286. Wu, T.-T., Lee, H.-Y., Li, P.-H., Huang, C.-N., & Huang, Y.-M. (2024). 通过基于ChatGPT的学习辅助工具，在混合式学习中促进自我调节发展与知识建构。Journal of Educational Computing Research, 61(8), 1539-1567. Xie, Y., Pan, D., Wang, L., Xie, Y., & Zeng, X. (2025). 小学数学人机协同精准讲评课模式的构建研究。国际教育技术会议，Xing, W., Song, Y., Li, C., Liu, Z., Zhu, W., & Oh, H. (2025). 开发用于初中数学学习的生成式人工智能驱动的可教学智能体：一项基于设计的研究。British Journal of Educational Technology. Xuan, S. H., Nguyen, A. T., Nguyen, T., Nguyen, L., Nguyen, H., Pham, N., Phung, T., Ngo, B., Nguyen, V., & Nguyen, M. (2025). 评估生成式人工智能对数学教育的影响：越南高中的一项比较研究。Human Behavior and Emerging Technologies, 2025(1), 8886206. https://doi.org/10.1155/hbe2/8886206 Zhu, Y., Liu, Q., & Zhao, L. (2025). 探究生成式人工智能对学生学习成果的影响：一项元分析。Education and Information Technologies, 1-29.

附录表 A.1。纳入研究及其部分研究特征概览。

作者	出版 date	出版 形式	数量 参与者	ISCED 等级	内容 area	人工智能用途	AI 功能	人工智能系统 修改	数量 效应	合并研究 效应（g）
Bastani 等（2025）	2025年6月25日	期刊	943	2; 3	C; Q; U	1;2;3;4	sup	1; 2	8	0.00
Bešlić 等（2024）	2024年7月3日	会议录	86	2	Q	3	sup	2	1	-0.40
Canonigo（2024）	2024年9月13日	期刊	60	3	C	3	rep	1	1	1.62
Cheng 等（2024）	2024年1月31日	期刊	79	2; 3	C; Q; S	2;3	rep	3	1	-0.34
Dasari 等（2024）	**2024年1月11日**	**期刊**	**20**	**6**	**C**	**1;3**	**重复；上标**	**1**	**2**	**-0.16**
El-Shara 等（2025年）	2025年4月18日	期刊	94	6	C	2;3	rep	1	2	0.59
Fardian 等（2025年）	2025年4月15日	期刊	30	6	C	2;3	重复；补充	1	2	0.34
Henkel 等（2024）	2024年5月5日	会议录	477	1; 2	C; Q	2;3	rep	3	1	0.34
Karaman（2024）	**2024年2月4日**	**期刊**	**29**	**1**	**S;Q**	**6**	**sup**	**2**	**1**	**0.11**
Kretzschmar和Seitz（2024）	2024年7月30日	会议录	275	3; 2	C	3	rep	3	2	0.09
Lademann等（2025）	2025年5月7日	期刊	214	2	C	3	rep	2	1	0.00
Liu等（2025）	2025年4月25日	会议录	90	1	S	3	sup	2	1	1.04
Pardos 和 Bhandari（2024）	2024年5月24日	期刊	274	成年人	C; U	3	支持；代表	1	2	0.30
Rücker 和 Becker-Genschow（2025）	**2025年11月24日**	**期刊**	**195**	**2**	**C**	**3**	**sup**	**3**	**1**	**1.25**
Serrano Heredia 等（2025）	2025年6月3日	会议录	550	6	C	3	rep	3	2	0.26
Steinbach等人（2025）	2025年7月17日	会议录	131	成年人	U	3	sup	2	1	0.34
Utami等人（2024）	**2024年5月29日**	**期刊**	**51**	**1**	**S**	**2**	**sup**	**3**	**3**	**0.85**
Wahba 等人（2024）	2024年7月1日	期刊	56	6	U	2;3	rep	1	1	1.38
Xie 等人（2025）	**2025年11月26日**	**会议录**	**85**	**1**	**S**	**2**	**sup**	**3**	**2**	**0.66**
Xing 等（2025）	2025年4月17日	期刊	212	2	C; Q; S	2;3	sup	2	1	0.46
Xuan 等（2025）	**2025年7月14日**	**期刊**	**120**	**3**	**C**	**2;3**	**sup**	**1**	**2**	**-0.11**

注。新纳入的研究以**粗体**显示。出版形式：期刊 = 期刊；会议录 = 会议论文集；ISCED 等级：0 = 幼儿教育；1 = 小学教育；2 = 初中教育；3 = 高中教育；4 = 高中后非高等教育；5 = 短周期高等教育；6 = 学士；7 = 硕士；8 = 博士。内容领域（PISA）：C = 变化与关系；Q = 数量；S = 空间与图形；U = 不确定性与数据。人工智能用途：1 = 数学专家；2 = 评估与反馈；3 = 指导者；4 = 对话伙伴；5 = 协作学习促进者；6 = 教师支持。人工智能角色：sup =补充；rep = 替代

附录A.2

搜索字符串：

(“生成式人工智能” OR “生成式人工智能” OR 生成式人工智能* OR “大型语言模型*” OR 大语言模型* OR “聊天机器人*” OR “人工智能辅导员*” OR “人工智能助手*” OR ChatGPT OR “Chat GPT”) AND (教育* OR 教学* OR 指导* OR 教育学 OR 课程 OR 课堂 OR 学生* OR 学校* OR “高等教育” OR “K-12 教育”) AND (数学* OR 代数 OR 几何 OR 算术 OR 微积分 OR “文字应用题*” OR 数理能力 OR 分数* OR “定量推理”) AND (干预* OR 实验* OR “对照组” OR 随机对照试验 OR “随机对照*” OR 处理 OR 前测 OR 后测 OR (前* AND 后*))