Advances in Education 教育进展, 2026, 16(1), 585-592
Published Online January 2026 in Hans. https://www.hanspub.org/journal/ae
https://doi.org/10.12677/ae.2026.161082

Hans汉斯

生成式人工智能融入STEM课堂教学的国外实证研究综述

俞　快[1]，陈　斌[2*]，李亭萱[3*]

[1]上海交通大学媒体与传播学院，上海
[2]上海交通大学化学化工学院，上海
[3]上海交通大学教育学院，上海

收稿日期：2025年12月2日；录用日期：2025年12月31日；发布日期：2026年1月8日

摘　要

随着生成式人工智能的快速发展，大模型在课堂教学中的应用备受关注。本文以近三年国际期刊发表的29篇相关实证研究为分析对象，系统梳理大模型在STEM课堂教学中的应用研究现状。文献统计结果显示，当前该领域研究呈现明显的教育阶段聚焦特征：以高等教育场景为主，相关文献占比达62%；基础教育场景的研究相对薄弱，占比为38%。值得注意的是，截至目前，尚未检索到涉及学前教育或特殊教育领域的相关实证研究成果。从学科分布来看，基础教育的文献主要关注数学和物理，鲜少关注化学、生物、工程等学科。此外，在STEM课堂教学中，常用的3种大模型增强策略包括提示语工程，模型微调和检索增强生成。其中，提示语工程最为常用。最后，大模型在STEM课堂教学中的应用主要涵盖至教师教学(教)、学生学习(学)、学业评价(评)三个维度中。其中，学习评价相关的文献数量较多。总体而言，现有文献表明，大模型能够提升课堂互动性、促进深度学习和增强学生学习动机。未来研究应进一步关注大模型在具体学科中的优化实践，以推动STEM课堂教学向智能化、个性化与高质量方向发展。

关键词

STEM教育，课堂教学，文献研究，生成式人工智能

A Review of Empirical Studies in the International Journals on Integrating Generative AI into STEM Instruction

Kuai Yu[1], Bin Chen[2*], Tingxuan Li[3*]

[1]School of Media and Communication, Shanghai Jiao Tong University, Shanghai
[2]School of Chemistry and Chemical Engineering, Shanghai Jiao Tong University, Shanghai
[3]School of Education, Shanghai Jiao Tong University, Shanghai

[*]通讯作者。

文章引用: 俞快, 陈斌, 李亭萱. 生成式人工智能融入 STEM 课堂教学的国外实证研究综述[J]. 教育进展, 2026, 16(1): 585-592. DOI: 10.12677/ae.2026.161082

Received: December 2, 2025; accepted: December 31, 2025; published: January 8, 2026

Abstract

With the rapid development of generative artificial intelligence, the integration of large language models (LLMs) into STEM education has become an emerging research trend. This review synthesizes 29 empirical studies published in international journals for the past three years. The results of this review paper indicate that LLMs are mainly applied in higher education (62%), with comparatively fewer studies in K-12 contexts (38%), and no empirical studies were found in the context of preschool or special education. At the disciplinary level, K-12 studies tend to focus on mathematics and physics, whereas chemistry, biology, and engineering remain underexplored. Across these reviewed articles, three optimization methods in LLMs are often used: prompt engineering, model fine-tuning, and retrieval-augmented generation (RAG). The most often used method is prompt engineering. The integrations of LLMs into STEM education align with teaching, learning, and assessment in instructional practice. The assessment gains the greatest research attention. Overall, current empirical studies reveal that LLMs can increase instructional effectiveness, promote deeper cognitive engagement, and strengthen students' motivation. Future research should prioritize discipline-specific instructional design to advance more customized and higher-quality STEM instruction.

Keywords

STEM Education, Instructional Practice, Literature Review, Generative AI

1. 问题提出

《教育强国建设规划纲要(2024~2035 年)》明确提出，“促进人工智能助力教育变革”，并将探索“AI + 教育”应用场景新范式，打造人工智能教育大模型等作为未来教育目标[1]。在全球科技竞争日趋激烈的背景下，加强 STEM 教育成为各国抢占科技竞争和未来发展制高点的重要战略举措。STEM 教育不仅是培养未来科技人才的关键，更是推动国家创新能力和竞争力的重要支柱[2]。在此背景下，如何借助人工智能促进 STEM 课程的高质量建设发展，已成为教育研究与一线教学实践共同关注的议题。

ChatGPT 是由 OpenAI 于 2022 年 11 月发布的生成式人工智能平台。近年来，以 ChatGPT 为代表的大模型(大语言模型或视觉语言模型)，凭借其强大的数据处理能力和广泛的应用场景，不仅为教师的教学提供了前所未有的便捷，也为学生的学习方式开辟了崭新的可能性。大模型通过即时互动、个性化推荐等多种功能，丰富了教–学–评的方法与策略，因此受到了广泛的关注[3]。

从国际研究来看，大模型在 STEM 课堂教学中的应用处于快速探索阶段，现有文献在研究对象、方法路径与应用层面上呈现较大差异。因此，本文主要对国际上的相关实证研究进行综述，梳理前沿研究特征与经验，以期促进我国在该领域研究的发展，为大模型融入至我国 STEM 课堂教学实践提供可借鉴经验。基于此，本文围绕以下四个具体问题展开综述：

(1) 大模型的应用主要分布在哪些学段与学科？

(2) 在 STEM 课堂教学中，常用的大模型增强技术是什么？

(3) 在 STEM 课堂中，大模型在“教–学–评”中的侧重点是什么？

(4) 现有文献主要采用什么研究方法？

2. 文章检索

以“large language model”(或“LLM”“ChatGPT”“vision language model”“VLM”)，“students”(或“education”“teaching”)和“STEM education”(或“STEM classroom”“science education”“biology education”“chemistry education”“math education”“physics education”“engineering education”)为关键词，在 Web of Science 以及 EBSCO 数据库上检索文献。检索时间范围限定为文献发表于 2023 年 1 月至 2025 年 6 月。

为保证纳入研究的质量和相关性，具体筛选标准如下：(1) 发表在通过同行评审的学术期刊上的论文；(2) 属于实证研究；(3) 研究内容涉及切实在课堂教学中使用大模型而不是仅探讨师生对于大模型的满意度等；(4) 不包括学位论文或会议论文；(5) 语种为英语。通过浏览题目和摘要进行筛选，共得到 29 篇符合要求的文章。

3. 研究结果

3.1. 学段与学科分布

学段分布显示，目前大模型在 STEM 课堂教学的应用覆盖了基础教育至高等教育，但尚未发现涉及学前教育或特殊教育的研究。学段和学科的具体分布如图 1 所示。从学段分布来看，高等教育研究的文献数量较多，共 18 篇(62%)，基础教育文献数量为 11 篇(38%)。从学科分布来看，基础教育的研究文献未涉及化学、生物、工程等学科。总体而言，现有研究关注的问题与各阶段的教育目标高度契合。小学阶段侧重基础认知理解能力，初高中阶段聚焦于学科概念本质的理解，高等教育阶段则注重专业能力与高阶思维能力的培养。

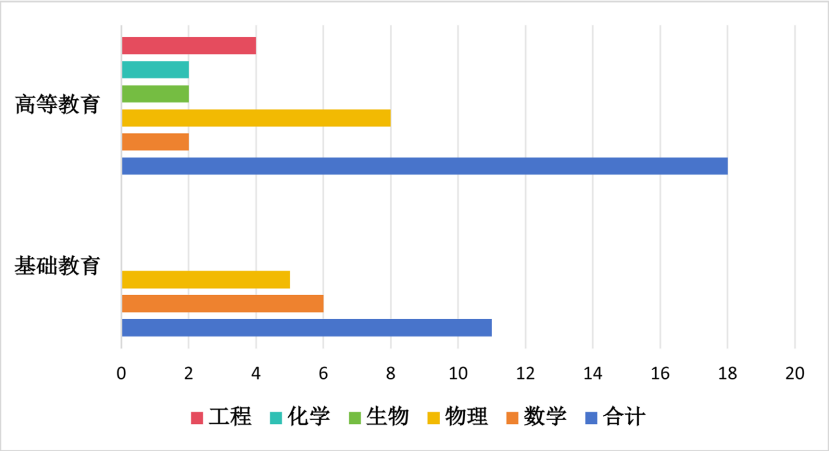

Figure 1. Distribution of disciplines and educational levels in the integration of LLMs into STEM classrooms
图 1. 大模型融入至 STEM 课堂教学的学科与学段分布

小学阶段研究以数学课程为主，重点关注学生对数学基本概念的理解。由于小学生认知水平处于发展的初始阶段，研究多围绕如何降低外在认知负荷，以帮助学生更好地理解题目。Hwang 等(2024)的研究开发了小学数学应用题生成系统，通过 TensorFlow 实现物体识别、ARCore 实现物体测量，获取真实场景中的几何信息，再结合 ChatGPT-3.5 模型，生成三个难度等级不同的几何应用题。研究结果显示，这些

题目在场景相关性、多样性和质量上均表现良好，能够有效适配不同认知水平学生[4]。Patel 等(2022)的研究使用 ChatGPT-3 模型对三到五年级课堂上常用的 250 道应用题进行简化，提高可读性，发现大模型能有效帮助教师将题目进行有意义的简化[5]。初中阶段研究侧重于开发高质量的教学支持系统。初中阶段的知识与概念逐渐从直观到抽象、从单一到综合，呈现出系统化与模块化的特点。基于此，Malik 等(2024)在关于数学课堂上的支架设计研究中，使用大模型生成与教学目标对齐的热身任务，助力教师自然且高效地引出“抽象性”这个核心概念[6]。

高中阶段以物理为主要学科，聚焦如何帮助学生理解复杂概念以及如何评价学生的理解程度。高中物理较多关注如何解决综合问题，因此，多数研究通过大模型辅助实现解题指导、评分反馈方面的支持。Bitzenbauer (2023)的研究聚焦量子物理课程，引导 12 年级学生使用 ChatGPT 辅助解释光子的概念并生成关于波粒二象性的 3 道概念测试题，并要求学生借助课本、文献等其他资源对大模型生成的内容进行批判性分析，从而提升学生在学习过程中的人机互动素养[7]。

高等教育是较为核心的学段。此学段的研究注重培养学生高阶思维能力，如逻辑推理、科学论证等能力的综合发展。Wu 等(2024)的研究结合同伴评估周期(Peer-Assessment Cycle)，开发出 PA-GPT 工具，即 AI 以虚拟同伴身份参与大一工科学生的学习过程，并与使用传统通用大模型即 ChatGPT 的对照组进行比较，发现 PA-GPT 在学生的知识构建、高阶思维能力方面的提升效果显著优于对照组学生[8]。Reddy (2024)以本科高年级的生物化学课程为研究对象，指导学生使用大模型辅助完成专业论文写作，并根据人工查阅文献结果和同伴反馈进行修订。研究发现，在大模型生成的文本中，知识概念较为准确，但仍存在引用错乱和部分内容不准确等问题，需要学生批判性地修改与甄别[9]。

3.2. 大模型增强技术

现有研究主要以大模型为主，传统机器学习模型为辅。在大模型的应用中，GPT 系列因强大的自然语言处理能力与使用便利性占据主导地位，其使用效果也在实验中得到了验证。其他大模型也各有适配场景，如 BLOOM、YOU 被用来检测小学数学回答连贯性[10]。在大模型的使用过程中，常用的 3 种增强策略是提示语工程(Prompt engineering)、模型微调(Fine-tuning)、以及检索增强生成(RAG)。

提示语工程无需改变通用大模型本身，而是通过优化输入指令引导模型生成更高质量的输出。此方法成本较低，但仍受限于模型原有知识。在 Chen 等(2025)的实验中，研究者采用 GPT-3.5 与 GPT-4o 批阅本科生物理作业，并运用提示语工程作为增强策略。通过调整评分标准的详细程度，以及是否要求模型进行推理，构建了四种不同的生成评分标准的提示词类型：简单思维链(COT)、详细思维链、详细对比与强制对比。思维链(COT)的推理模式要求模型展示逐步思考过程，“对比”的推理模式则要求模型逐项比较学生回答与评分标准。实验结果显示，针对大语言模型的评分弱点，补充关于“可接受表达形式”的详细规则，是使 GPT-4o 达到人工评分水平的关键。在提示词中明确规定回答格式也能小幅提升评分准确性。然而，过度结构化的推理要求(如强制对比)反而会导致准确率下降[11]。Tsai 等(2023)在大二的化工课上，以蒸汽轮机效率计算为内容载体，引导学生使用 ChatGPT 的提示词功能完成课程项目作业中的具体的问题解决任务[12]。

模型微调是特定专业数据集额外训练通用大模型，调整模型参数，从而增强大模型在解决专业问题上的能力。如 Yang 等(2025)的研究使用学生的中文文本和人工评分数据微调 ChatGPT-3.5-turbo，使用微调后的模型评估初中与高中学生对科学现象解释的回答内容，最终所有任务的评分准确率均超过 75%，能够较好地满足科学教育评价的专业场景需求[13]。

检索增强生成是在模型回答问题之前，先从外部库(如数据库、文档)实时检索相关信息，并将其作为上下文提供给模型。它不需要重新训练模型，而是为模型动态地补充新知识，实现即时查资料。Long 等

(2024)的课堂对话分析研究采用定制化的 GPT-4 模型，结合检索增强生成(RAG)技术，将课堂对话编码框架向量化后输入至大模型，使其掌握编码规则，再通过提示词引导模型按分类标准对对话轮次进行编码，最终实现对初中数学课堂师生对话的自动编码。这一机器编码效率是人工编码的 30 倍，且与人工编码的一致性达到 90%以上[14]。

此外，传统机器学习模型有时作为对比参照工具出现，用于比较不同模型的优势与劣势。Urrutia 等(2024)在研究中，将 XGBoost、BETO-mt 等传统 AI 模型与 GPT-3、BLOOM 等大模型的表现进行比较，结果显示最优传统模型 BETO-mt 的 F1 分数为 79.15%，甚至高于大模型的表现。这表明在特定文本分类任务中，传统模型仍具备一定竞争力[10]。Fussell (2025)的研究发现，大模型能够通过物理实验笔记推断学生的定量比较思维与实验技能的变化。此外，词袋模型、BERT 与 LLaMA 模型被用来作比较，其中，BERT 在资源消耗和性能之间达到较好平衡，为资源有限的研究场景提供了选择，体现了传统模型的独特价值[15]。

3.3. 教－学－评侧重点

大模型在 STEM 课堂教学中的应用研究在教师教学(教)、学生学习(学)、学业评价(评)三个维度各有侧重。其中，学习评价相关的文献数量较多，部分文献是涉及教-学-评中两个及以上维度的综合性研究，如图 2 所示。侧重教师教学的研究以支持教师教学设计、提供优质教学资源为核心，旨在减轻教师工作负担、提升教学效率。比如，Yu 等(2024)的研究提出并验证了混合智能反馈(HIF)系统的有效性，HIF 系统通过整合同伴反馈与机器智能，显著改善职前教师教学反思中出现的表面化、零散化问题。这一系统也具备在集体备课、教研等场景中帮助教师快速定位核心问题的能力[16]。

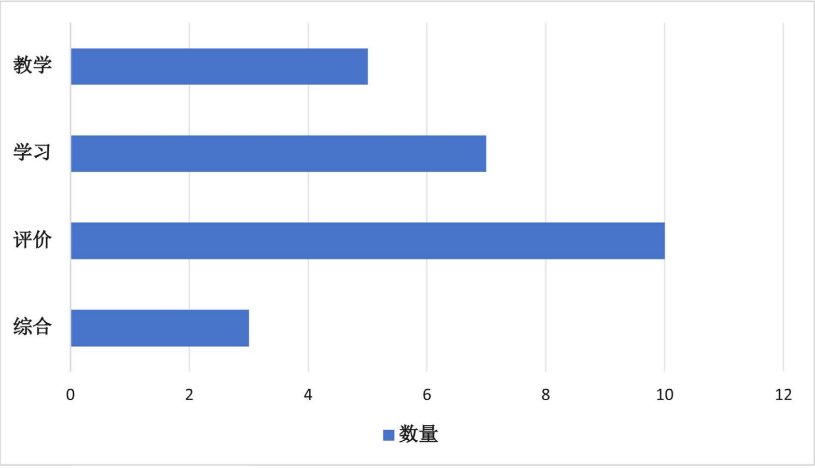

Figure 2. Distribution of LLM applications in STEM instruction: Teaching, learning, and assessment
图 2. 大模型在 STEM 课堂中的“教－学－评”分布

聚焦学生学习的研究主要探讨大模型如何促进学生知识掌握、能力发展以及学习态度的改进。Ng 等(2024)的研究以 74 名初中学生为研究对象，对比基于 ChatGPT 的 AI 聊天机器人(SRLbot)与传统 AI 聊天机器人(Nemobot)在物理“力与运动”主题学习中对学生自我调节学习(SRL)的影响。研究发现 AI 聊天机器人(SRLbot)在提升学生 SRL 能力、科学知识与动机方面，显著优于 Nemobot，其灵活性与个性化是关键优势。此外，研究还验证了人机交互次数是影响 SRL 效果的核心因素[17]。

在专业化程度较高的高等教育领域，大模型的辅助作用同样重要。Coban(2025)的研究以本科物理课程中的量子密码学实验为场景，结合 AR 可视化和 ChatGPT 个性化互动的教学能显著提升学生的概念理解和

视觉注意力分配[18]。在写作类任务中，大模型同样能够为学生学习提供有效支持。Behrens 等(2025)聚焦本科遗传学原理课程，借助大模型完成写作任务。学生认可大模型在生成写作大纲、提供学科知识的背景信息、语法检查与表述优化的辅助作用，89%的学生表示大模型帮助他们更深入思考遗传学伦理问题[19]。

侧重学习评价的研究致力于提升评价效率、优化评价质量，为教育决策提供科学依据。Wan (2024)的研究聚焦物理基础概念题，并使用大模型为学生提供反馈。学生认为 ChatGPT 反馈与人工反馈在正确性上无显著差异，且 ChatGPT 反馈更详细、更具针对性[20]。Xu 等(2025)关于高中物理简答题评阅的研究则使用大模型与人工分别对 80 份学生答卷进行评分并比较评分结果，发现大模型评分结果的一致性高于人工评分，但准确性低于人工评分。这一发现也为后续优化 AI 评价工具、明确人机协作评分等提供了研究方向[21]。在手写作业评价方面，大模型同样能发挥作用。Kortemeyer (2023)的研究结果表明，大模型可以较好地识别学生在电阻–电容电路问题上的作答情况，且大模型的评分与人工评分的一致性较好。这一研究发现适用于形成性评价与反馈，进一步拓展了大模型在作业评价中的应用范围[22]。Kieser (2023)在研究中发现，在物理教育的概念解释测试中，ChatGPT 能较好地模拟学生的常见错误。研究表明大模型能够帮助教师生成具有教学实践意义的数据且为试题开发等提供低成本、高效的先导性测试工具[23]。

在综合性研究中，Krupp 等(2025)聚焦使用大模型生成提示从而辅助学生进行量子计算。研究结果显示，GPT-4 生成的提示语不仅能引导学生解答量子计算中的选择题，还能通过灵活适配学生个性化疑问、对答题过程进行分步拆解，从而实现为学生提供即时讲解的功能。作者认为这是大模型在学生学习与作业评价方面的综合应用[24]。Martin (2023)的研究中使用了 BERT base uncased 模型，用于批阅本科有机化学课程中的科学论证作业。研究结果表明，人机评分结果相近，机器评分的准确率达 87%。在无监督机器学习辅助下，从学生作答数据中提取出 22 个论证模式，发现学生的推理模式以“描述性”、“关联性”为主，“线性因果”推理较少，这些结果被教师用来备课从而制定有针对性地教学设计。由此，作者认为这一研究发现是大模型在作业评价以及教师教学中的综合应用[25]。

3.4. 主要研究方法

通过梳理文献发现，现有文献使用的研究方法分为定量研究、质性研究以及混合研究方法。定量研究侧重验证大模型的有效性与可靠性，而质性研究则旨在深入解析人机交互过程中的行为特征编码。定量研究方法，通过收集结构化数据与统计分析，将大模型下的教学效果以量化形式呈现。其中干预研究最为常见，即通过设置实验组与对照组比较教学效果。Stadler 等(2024)的研究将 91 名无纳米技术先验知识的大学生随机分为实验组(使用 ChatGPT 3.5 检索信息)与对照组(使用谷歌检索信息)，两组完成相同的纳米颗粒防晒霜科学论证任务，通过标准化量表收集学生的作答数据后，采用协方差分析、中介效应检验等统计方法来进行比较。结果显示，实验组在外在负荷、内在负荷、关联负荷三类认知负荷上均显著低于对照组，凸显了大模型在减轻认知负荷上的优势[26]。

质性研究方法聚焦人机交互的过程性分析，深入探究人机协同下特有的模式。质性内容分析通过对聊天记录、文本记录等非结构化数据的多维度编码，解析用户(教师、学生)与大模型的交互模式。比如，在 Dilling (2024)研究中，通过对教师与大模型的对话交互记录的编码，探查基础教育职前数学教师与 ChatGPT 交互时使用的提示词类型与特征。结果表明，多数教师的交互行为较为单一，即仅包含单条提示词，鲜少使用追问对话功能[27]。此外，还有专家认定法、认知访谈法等分别从专业判断、用户感知等角度补充研究深度。

混合研究方法将定量与质性两种方法结合，整合二者优势。比如，用质性研究补充定量研究的结果，或用定量研究为前期质性研究结果做进一步阐释，从而使研究结果在整体上更具说服力。在 Küchemann (2023)研究中，26 名职前教师使用 ChatGPT 3.5 为十年级学生设计两组力学概念题，并与教科书习题的对

照组相比较。通过对生成的题目质量进行量化分析，发现两组生成的题目在准确度、难度等方面无显著差异；但在清晰度和情境适切性方面，ChatGPT 组存在一定局限性。作者进一步使用文本分析梳理题目内容后发现，这些局限性体现在 ChatGPT 组生成的题目在一定程度上缺乏抽象性[28]。

4. 结论与建议

通过系统梳理大模型在 STEM 课堂教学中的应用研究可以看出，这些研究成果为我国构建本土化的“大模型赋能 STEM 教育”模式提供了参考，也为教师提供了可操作的实践指导。总体而言，大模型为传统 STEM 课堂带来了新的活力。在具体教学实践中，大模型最初主要充当“工具”，协助教师简化题项，使知识呈现更加直观、简洁。随着教学目标从基础知识理解向高阶认知发展，大模型的角色逐渐扩展为学伴或助教。学生在与大模型的互动中反思和改进，体现出以促进学生深度理解为导向的学习过程。此外，基础教育的研究文献未涉及化学、生物和工程设计等教学。究其原因，化学和生物学很多教学依赖实验、试剂和安全操作。把大模型用作实验指导或替代教师指导可能会引发安全与责任问题，比如，错误的提示词可能导致危险。因此研究者与学校对把大模型直接用于学生实验持谨慎态度，导致相关实证研究较少。未来研究可探索如何利用大模型生成用于虚拟实验的操作指令或模拟生态系统的动态演变，同时，也可探索如何将工程设计嵌入至虚拟实验中。

未来，教师应根据 STEM 教育中不同的学科特点及学习目标设计相应的课堂活动，以充分发挥大模型在个性化教学中的潜力。尽管大模型在信息处理和即时反馈方面具有优势，但无法替代教师在专业判断与课堂管理等方面的核心作用。未来研究应致力于构建人类智慧与人工智能互补的课堂，从而提升整体教学质量。

基金项目

该研究受到国家自然科学基金面上项目“超快可视化研究液体水中手性等离激元异质纳米结构的光诱导输运动力学及机制”(编号：22573064)资助。

参考文献

[1] 中华人民共和国教育部. 中共中央 国务院印发《教育强国建设规划纲要(2024-2035 年)》[EB/OL]. 2025-01-19. http://www.moe.gov.cn/jyb_xxgk/moe_1777/moe_1778/202501/t20250119_1176193.html?zbb=true, 2025-11-20.

[2] 林成华, 张维佳. 世界主要发达国家 STEM 战略布局与借鉴建议[J]. 中国高等教育, 2024(7): 59-64.

[3] 王志军, 龙帅, 张吉. 人机协同智能课堂教学评价层级模型构建研究[J]. 远程教育杂志, 2025, 43(5): 32-40.

[4] Hwang, W. and Utami, I.Q. (2024) Using GPT and Authentic Contextual Recognition to Generate Math Word Problems with Difficulty Levels. *Education and Information Technologies*, **29**, 1-29. https://doi.org/10.1007/s10639-024-12537-x

[5] Patel, N., Nagpal, P., Shah, T., Sharma, A., Malvi, S. and Lomas, D. (2023) Improving Mathematics Assessment Readability: Do Large Language Models Help? *Journal of Computer Assisted Learning*, **39**, 804-822. https://doi.org/10.1111/jcal.12776

[6] Malik, R., Abdi, D., Wang, R. and Demszky, D. (2025) Scaffolding Middle School Mathematics Curricula with Large Language Models. *British Journal of Educational Technology*, **56**, 999-1027. https://doi.org/10.1111/bjet.13571

[7] Bitzenbauer, P. (2023) ChatGPT in Physics Education: A Pilot Study on Easy-to-Implement Activities. *Contemporary Educational Technology*, **15**, ep430. https://doi.org/10.30935/cedtech/13176

[8] Wu, T., Lee, H., Chen, P., Lin, C. and Huang, Y. (2025) Integrating Peer Assessment Cycle into ChatGPT for Stem Education: A Randomised Controlled Trial on Knowledge, Skills, and Attitudes Enhancement. *Journal of Computer Assisted Learning*, **41**, e13085. https://doi.org/10.1111/jcal.13085

[9] Reddy, M.R., Walter, N.G. and Sevryugina, Y.V. (2024) Implementation and Evaluation of a ChatGPT-Assisted Special Topics Writing Assignment in Biochemistry. *Journal of Chemical Education*, **101**, 2740-2748. https://doi.org/10.1021/acs.jchemed.4c00226

[10] Urrutia, F. and Araya, R. (2024) Who’s the Best Detective? Large Language Models vs. Traditional Machine Learning

in Detecting Incoherent Fourth Grade Math Answers. *Journal of Educational Computing Research*, **61**, 1723-1754. https://doi.org/10.1177/07356331231191174

[11] Chen, Z. and Wan, T. (2025) Grading Explanations of Problem-Solving Process and Generating Feedback Using Large Language Models at Human-Level Accuracy. *Physical Review Physics Education Research*, **21**, Article 010126. https://doi.org/10.1103/physrevphyseducres.21.010126

[12] Tsai, M., Ong, C.W. and Chen, C. (2023) Exploring the Use of Large Language Models (LLMS) in Chemical Engineering Education: Building Core Course Problem Models with Chat-GPT. *Education for Chemical Engineers*, **44**, 71-95. https://doi.org/10.1016/j.ece.2023.05.001

[13] Yang, J., Latif, E., He, Y. and Zhai, X. (2025) Fine-Tuning ChatGPT for Automatic Scoring of Written Scientific Explanations in Chinese. *Journal of Science Education and Technology*, **34**, 719-736.

[14] Long, Y., Luo, H. and Zhang, Y. (2024) Evaluating Large Language Models in Analysing Classroom Dialogue. *npj Science of Learning*, **9**, Article No. 60. https://doi.org/10.1038/s41539-024-00273-3

[15] Fussell, R.K., Flynn, M., Damle, A., Fox, M.F.J. and Holmes, N.G. (2025) Comparing Large Language Models for Supervised Analysis of Students' Lab Notes. *Physical Review Physics Education Research*, **21**, Article 010128. https://doi.org/10.1103/physrevphyseducres.21.010128

[16] Yu, J., Yu, S. and Chen, L. (2025) Using Hybrid Intelligence to Enhance Peer Feedback for Promoting Teacher Reflection in Video-Based Online Learning. *British Journal of Educational Technology*, **56**, 569-594. https://doi.org/10.1111/bjet.13559

[17] Ng, D.T.K., Tan, C.W. and Leung, J.K.L. (2024) Empowering Student Self-Regulated Learning and Science Education through ChatGPT: A Pioneering Pilot Study. *British Journal of Educational Technology*, **55**, 1328-1353. https://doi.org/10.1111/bjet.13454

[18] Coban, A., Dzsotjan, D., Küchemann, S., Durst, J., Kuhn, J. and Hoyer, C. (2025) AI Support Meets AR Visualization for Alice and Bob: Personalized Learning Based on Individual ChatGPT Feedback in an AR Quantum Cryptography Experiment for Physics Lab Courses. *EPJ Quantum Technology*, **12**, Article No. 15. https://doi.org/10.1140/epjqt/s40507-025-00310-z

[19] Behrens, K.A., Marbach-Ad, G. and Kocher, T.D. (2024) AI in the Genetics Classroom: A Useful Tool but Not a Replacement for Creative Writing. *Journal of Science Education and Technology*, **34**, 621-635. https://doi.org/10.1007/s10956-024-10160-6

[20] Wan, T. and Chen, Z. (2024) Exploring Generative AI Assisted Feedback Writing for Students' Written Responses to a Physics Conceptual Question with Prompt Engineering and Few-Shot Learning. *Physical Review Physics Education Research*, **20**, Article 010152. https://doi.org/10.1103/physrevphyseducres.20.010152

[21] Xu, Y., Liu, L., Xiong, J. and Zhu, G. (2025) Graders of the Future: Comparing the Consistency and Accuracy of GPT4 and Pre-Service Teachers in Physics Essay Question Assessments. *Journal of Baltic Science Education*, **24**, 187-207. https://doi.org/10.33225/jbse/25.24.187

[22] Kortemeyer, G. (2023) Toward AI Grading of Student Problem Solutions in Introductory Physics: A Feasibility Study. *Physical Review Physics Education Research*, **19**, Article 020163. https://doi.org/10.1103/physrevphyseducres.19.020163

[23] Kieser, F., Wulff, P., Kuhn, J. and Küchemann, S. (2023) Educational Data Augmentation in Physics Education Research Using ChatGPT. *Physical Review Physics Education Research*, **19**, Article 020150. https://doi.org/10.1103/physrevphyseducres.19.020150

[24] Krupp, L., Bley, J., Gobbi, I., Geng, A., Müller, S., Suh, S., *et al*. (2025) LLM-Generated Tips Rival Expert-Created Tips in Helping Students Answer Quantum-Computing Questions. *EPJ Quantum Technology*, **12**, Article No. 33. https://doi.org/10.1140/epjqt/s40507-025-00334-5

[25] Martin, P.P., Kranz, D., Wulff, P. and Graulich, N. (2024) Exploring New Depths: Applying Machine Learning for the Analysis of Student Argumentation in Chemistry. *Journal of Research in Science Teaching*, **61**, 1757-1792. https://doi.org/10.1002/tea.21903

[26] Stadler, M., Bannert, M. and Sailer, M. (2024) Cognitive Ease at a Cost: LLMS Reduce Mental Effort but Compromise Depth in Student Scientific Inquiry. *Computers in Human Behavior*, **160**, Article 108386. https://doi.org/10.1016/j.chb.2024.108386

[27] Dilling, F. and Herrmann, M. (2024) Using Large Language Models to Support Pre-Service Teachers Mathematical Reasoning—An Exploratory Study on ChatGPT as an Instrument for Creating Mathematical Proofs in Geometry. *Frontiers in Artificial Intelligence*, **7**, Article ID: 1460337. https://doi.org/10.3389/frai.2024.1460337

[28] Küchemann, S., Steinert, S., Revenga, N., Schweinberger, M., Dinc, Y., Avila, K.E., *et al*. (2023) Can ChatGPT Support Prospective Teachers in Physics Task Development? *Physical Review Physics Education Research*, **19**, Article 020128. https://doi.org/10.1103/physrevphyseducres.19.020128