

AI 让论文写得更好，也让学生学得更好吗？

将作品质量与独立论证能力分开，考察支持撤除后的迁移

ThoughtForge · 研究报告选集 03 · 核心章节节选

教育学 · 学习科学

研究提要

同一篇高分论文，可能来自更充分的学习，也可能来自更完整的外部代写；作品评分本身无法辨认两者。本节选提出以“支持撤除后的独立论证”作为主要结果，将直接提供文本与追问式反馈区分，并通过同等时间的对照、盲评及延迟任务检验学习是否迁移。重点不在判定学生用了多少 AI，而在辨认哪些认知工作仍由学生承担。

研究问题

生成式 AI 改善课程论文时，学生究竟提高了表达、完成效率，还是形成了可独立使用的论证能力？怎样的反馈方式更可能支持后者？

把“作品变好”与“论证能力增长”分开测量

课程论文评分同时受到论点、证据组织、语言表达与格式规范影响。AI 即使只改善表达，也可能提高总分；如果直接代为选择论点、组织反驳和修订推理，作品变好甚至可能与学生减少认知投入同时发生。因此，研究不应从总分变化反推出学习，也不宜把提示次数或生成字数当作学习投入的直接指标。

本报告将独立论证能力界定为：面对一个新问题，能够提出可辩护的主张，解释证据与主张之间的关系，识别竞争解释，并据反对意见修订。它可以通过不借助 AI 的短论证任务、对自身判断的说明与新材料迁移来观察。论文完成效率和主观满意度仍值得报告，但应作为不同结果，不与能力增长混为一谈。

已有证据提示测量分离，不能直接替代写作研究

Bastani 等关于高中数学的研究区分了使用 AI 时的练习表现与撤除工具后的独立表现，并比较不同工具设计。这提供了一个与本题相关的测量提醒：受支持的完成情况与独立能力可能不同。但研究情境是数学学习，不能直接外推为大学论文写作中的因果结论。

Roediger 与 Karpicke 对提取练习的研究则提示，眼前表现、自信与延迟保持之间需要分别检验。本报告据此提出一种设计推论：论文反馈也应安排学生独立解释与延迟迁移；这不是原研究已经证明追问式 AI 能提高写作能力。后续文献工作仍需核对写作反馈、论证评价与学习迁移研究中任务和量表的适配性。

本节参考：[1]、[2]

核心构想：把反馈设计成“需要自己作出判断”

直接改写型反馈可以给出一段更好的文本；追问式反馈则要求学生先说明证据为何支持主张、哪个反例最有威胁，以及打算如何修订，再提供有限的提示。待检验的机制是：后一种设计可能增加学生自己完成解释和判断的机会，使修订不只发生在文档上，也发生在其对问题的理解中。

这一设想需要限制适用条件。对尚未理解基本概念的学生，连续追问可能增加负担；给出示范再逐步撤除支持也可能更有效。因此，不应预设“越少给答案越好”。关键比较是在教学目标、时间与可用资料相近的条件下，哪些反馈把必要的判断留给学生，以及这些判断能否在新任务中再次出现。

拟议机制

反馈形式 → 本人解释与修订 → 支持撤除 → 独立论证迁移

拟议教学机制，不能由作品分数或交互次数直接推定。

对照设计：让工具撤除后的任务成为主要结果

在适用的伦理与教学许可下，先开展不计入高风险成绩的试点。可设置直接改写反馈、追问式反馈与结构化静态检查表三组，围绕同一论证任务与同一材料包学习，并按基线写作能力分层随机分配。记录实际时间、支持使用与退出情况；如果学生之间会分享反馈，需要在分配单位和实施安排中处理这种污染。

主要结果应预先指定为延迟、无 AI 的新材料论证任务，按主张清晰度、证据推理、反驳质量与修订合理性分别盲评。即时作品分数作为次要结果；可另设保留工具的任务研究工具使用能力，但不能将其与独立能力合成同一分数。样本量须依据评分一致性、组内相关、预期失访与最小有意义效应规划。

让每一个“提高”都有对应的测量

即时作品质量：建议任务：同一材料下完成短论文，分维度盲评；解释边界：只能说明受支持产物表现。

独立论证能力：建议任务：撤除 AI 后，用新材料回应相近问题；解释边界：仍受任务难度与先前知识影响。

自我判断准确性：建议任务：预测论证得分，再对照盲评分维度；解释边界：高自信本身不是能力提升。

策略迁移：建议任务：解释为何保留或推翻一个主张；解释边界：口头解释需防止仅测表达流利度。

最强反对意见：追问式反馈只是让学生练得更久

追问组若多花了时间，结果差异可能来自练习量，而非反馈方式。可以事先规定相近的学习时段，记录实际使用时间，并设置具有相似任务要求的静态检查表对照。主要分析按原始分配报告，实际投入用于解释实施情况；未经额外识别假设，不应通过控制政策后的投入时长就声称识别了纯机制效应。

另一个问题是评价的公平性。熟悉议论文表达的学生可能更容易在评分中获益，因此量表应区分表达润色与证据推理，使用多名不知道组别的评分者并检查一致性。AI 检测分数不能作为是否学习的指标。若新任务只换了题面却复用了相同论证模板，也可能高估迁移，需要增加材料或立场变化并说明迁移距离。

推进计划：先验证评价，再检验教学机制

第一步用既有、经授权的学生习作作检验量表，整理评分者分歧；第二步进行低风险试点，检查任务难度、时间与学生能否理解反馈；第三步才锁定干预、主要结果与分析方案，安排正式研究。保留模型版本、提示策略与交互记录的必要摘要，但不收集无关的个人信息，并明确材料如何保存和删除。

如果追问组只在即时论文中更好，而延迟独立任务没有相应变化，应将结论限定为任务支持；如果优势完全由表达维度驱动，则不能宣称论证能力提高。如果工具使用能力提高而独立能力不变，也可以形成有价值的教学判断：课程究竟希望培养哪种能力，需要在评价中说清楚，而不是让一项总分承担所有目标。

仍需回答的问题

对不同先前知识水平，示范与追问应如何组合？跨课程迁移与数周后的保持是否相同？学生对反馈的主动选择，是否会改变课堂真实实施的效果？这些问题需要后续研究，不能由一次短期试点作出全面判断。

参考文献与阅读范围

下列文献用于概念定位，未作系统文献综述；不代表本报告方案已获验证。

[1] Bastani, H., et al. (2025). Generative AI without guardrails can harm learning: Evidence from high school mathematics. PNAS, 122(26), e2422633122.

<https://doi.org/10.1073/pnas.2422633122>

阅读范围：公开摘要与书目信息；仅借鉴受支持表现与独立学习的区分，不外推数学研究的效应大小。

[2] Roediger, H. L., & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. Psychological Science, 17(3), 249 – 255.

<https://doi.org/10.1111/j.1467-9280.2006.01693.x>

阅读范围：期刊摘要与书目信息；用于即时表现和延迟保持的概念区分，不作为写作干预已有效的证据。

以上为报告节选，研究方案尚待实施。

本组选集为编写的研究方案展示稿，不是客户案例或系统运行记录。文献仅在上述阅读范围内使用；实际交付随课题、材料和处理预算展开。

本报告含 AI 辅助生成内容；观点不构成学术文献。