《心理科学进展》审稿意见与作者回应

题目：大学生 AI 素养与自我效能关系的贝叶斯元分析
作者：万千一 温斯情 马早明 胡 波

第一轮

审稿人 1 意见：

本文整体写作条理清晰，研究问题明确，元分析实施流程较为规范，结果呈现也较完整。但为进一步提升论文的学术严谨性与论证力度，建议作者重点从以下方面完善与修改。

意见 1：引言前四段对宏观背景的铺陈较长，与本文核心研究问题的直接关联尚不够紧密，导致研究切入点出现偏晚。此外，现有内容在逻辑上尚不足以推导出“必须采用元分析方法”来回答该主题的必要性。建议：适当压缩前四段背景性论述，尽早聚焦于“大学生 AI 素养与自我效能关系”的研究现状，凸显开展元分析的研究价值与必要性。

回应：感谢专家的宝贵建议。我们已对前言结构进行压缩与重组：删减原稿中较长的宏观背景铺陈（原稿前言以宏观变革与政策背景展开，导致切入点偏晚），并将前言开篇调整为“AI 进入高等教育—AI 素养的心理机制问题—聚焦 AI 素养与自我效能”的逻辑链条，使研究问题更早出现、与核心变量的关联更直接。具体修改如下（原文修改内容以标蓝）：

1. 压缩宏观背景、提前聚焦研究问题。修改稿前言第一段用更简洁的表述交代 AI 进入高等教育的背景，并直接引出“个体层面机制仍需更清晰证据”的问题，从而缩短铺陈、前置研究切入点。第二段进一步从心理学视角引入自我效能作为关键机制变量，并明确提出“检验大学生 AI 素养与自我效能关联强度及作用条件”的必要性，使核心研究主题在前言前两段即得到清晰呈现。

人工智能（AI）作为通用型技术正快速进入高等教育，改变知识获取、信息处理与问题解决方式（Hamamoto et al., 2024）。在 AI 能够高效生成与加工信息的情境下，高校人才培养目标加速由“知识导向”转向“素养导向”，强调学生在智能情境中的理解、判断与协作能力（Dong, 2024；OECD, 2023）。因此，AI 素养培养成为高校改革的重要议题，但其在个体层面如何产生有效影响的心理机制仍有待更清晰的证据支持（Curaj et al., 2025）。

从心理学视角看，AI 素养并非单一技术技能，而是包含知识理解、策略性使用与自我调控、风险与伦理判断及相关态度信念等复合结构。因此，AI 素养能否进一步转化为持续的学习投入与良好的适应结果，往往取决于个体动机与信念等关键心理因素的作用。自我效能感作为社会认知理论的核心变量，反映个体对完成特定任务能力的主观判断，能够影响目标设定、努力程度、坚持性与情绪体验，并进一步塑造个体在新技术学习与人机协作中的行为选择与表现（Atchley et al., 2024）。据此可推论，大学生 AI 素养越高，其在 AI 相关学习任务中的掌控感与能力预期越强，从而更可能形成较高自我效能。相反，AI 素养不足可能伴随不确定感与低掌控体验，并与较低自我效能水平相关。因此，检验大学生 AI 素养与自我效能之间的关联强度及其作用条件，对于揭示 AI 素养培养的心理机制与教育干预着力点具有重要意义。

2. 补强“为何需要元分析”的逻辑推导。我们在前言新增（强化）了对既有实证研究现

状的概括，指出该领域研究结论分散且不一致，并具体说明不一致可能来源于操作化与测量差异、样本与情境差异、研究设计与统计处理差异；在此基础上明确提出单项研究难以回答总体效应强度及边界条件，因此采用注册元分析整合证据、估计总体效应并检验调节因素，以提供更稳健的证据基础。

近年来相关实证研究不断增加，但关于大学生 AI 素养与自我效能关系的结论仍分散且不一致。不同研究对 AI 素养与自我效能的操作化与测量工具存在差异，削弱了结果可比性。样本来源与教学情境不一，可能带来效应差异。研究设计与统计处理方式不同亦会导致效应量波动。由此，单项研究难以回答二者总体相关程度有多强，以及在何种研究条件下更强或更弱等关键问题。基于此，本研究采用注册元分析整合既有研究，估计总体效应并检验调节因素，以期提供更稳健的证据基础。

意见 2：文中同时报告随机效应模型与贝叶斯随机效应模型，为什么同时采用两种方法的原因没有叙述清楚。从题目与研究定位来看，读者容易产生疑问：随机效应模型与贝叶斯模型分别承担何种分析角色？建议在方法部分明确说明两类模型并行使用的理由。

回应：非常感谢专家指出这一关键表述问题。我们已在方法部分补充说明两类模型并行使用的角色分工与理由：本研究以贝叶斯随机效应模型作为主要分析框架，用于估计总体效应及其不确定性并报告后验可信区间；同时报告频率学派随机效应模型，主要用于与既有元分析研究的常规报告口径对照，并作为对主要结论的敏感性/稳健性检验（两种估计相互参照，检验结论是否一致）。这一说明已明确回答“为何要同时报告”以及“两类模型分别承担什么功能”。

修改位置说明：

修改稿 2.6 统计分析：新增“贝叶斯随机效应为主要框架、频率学派随机效应用于对照与敏感性检验”的理由与分工说明。

本研究的元分析在 Python 环境下完成，涵盖数据整理、效应量换算、随机效应模型合并、异质性评估以及调节与亚组分析等步骤。因此 Fisher’s z 尺度下建立贝叶斯随机效应元分析模型，用于估计总体效应及研究间异质性。总体效应参数采用弱信息先验（Normal(0,1)），异质性参数采用半正态先验（HalfNormal(0.5)）。模型使用 MCMC 方法进行后验抽样，并通过 R-hat 与有效样本量（ESS）评估收敛性。为检验先验设定对结论的影响，进一步将异质性参数先验调整为 HalfNormal(1.0)进行敏感性分析；结果用于判断总体效应与异质性估计的稳健性。最终报告后验点估计及可信区间。

为便于与既有元分析研究对照，并作为对主要结论的补充稳健性检验，本研究同时报告频率学派随机效应模型的对应估计结果。效应量指标采用 Pearson 相关系数 *r*。合并分析前，将所有相关系数转换为 Fisher’s z 值以稳定方差并改善近似正态性；合并完成后再将 Fisher’s z 值反转换为 Pearson 相关系数 *r*，以便结果解释与呈现。”

修改稿 4.1 总体关系：在高异质性背景下采用贝叶斯随机效应作为分析以更充分呈现不确定性。“鉴于纳入研究之间存在显著异质性（例如 I2 较高），传统的 DerSimonian–Laird 随机效应模型在异质性参数的估计上可能受研究数量与异质性程度影响而不够稳定，且对不确定性的刻画主要依赖正态近似（DerSimonian & Laird, 1986；Röver & Friede, 2022）。因此，为检验结论的稳健性并补充呈现不确定性信息，本研究以贝叶斯随机效应荟萃分析为主。该方法将

每项研究的观测效应视为其真实效应的含误差估计，同时允许真实效应在总体平均效应周围合理波动，更契合心理学研究中效应可能随情境变化而呈现差异的特征（Röver, 2020c）。在模型设定中，总体效应采用弱信息先验 $\mu \sim \text{Normal}(0, 1)$，异质性参数采用 $\tau \sim \text{HalfNormal}(0.5)$。随后使用多链 MCMC（NUTS）进行采样（chains = 4，draws = 6000，tune = 1500，target_accept = 0.95，seed = 123），并以 R-hat 与有效样本量（ESS）评估收敛性。”

意见 3：文中多次出现“AI 素养如何有效提升自我效能”等因果取向表述，但纳入研究以横断研究为主，严格而言只能支持相关/关联。建议全面检查并将“提升、促进、增强”等措辞改为更审慎表达，并在讨论中明确指出因果推断仍需更多纵向或实验研究证据支持。

回应：非常感谢专家的重要提醒。我们完全同意：由于本研究纳入文献以横断面相关研究为主，元分析结论在严格意义上主要支持“相关、关联”层面的证据，而非因果结论。为此，我们已对全文进行系统性修订，重点完成以下三方面修改：

1. **统一弱化因果措辞，改为“相关/关联”表述**。我们对摘要、引言、讨论与结论中涉及“提升、促进、增强”等因果倾向措辞进行了逐句核查与替换，尽量采用“相关、关联、联系、可能相关”等更审慎表达。比如在摘要中已明确使用“稳定正向关联”来概括核心发现，避免因果化表述。

2. **在理论推导部分避免“AI 素养→自我效能”的确定性因果表述。**在社会认知理论相关段落，我们将原先更偏因果的措辞调整为“可能与……相关，并进一步与……相关”的关联性表述，以保证理论阐释与证据类型相匹配。

3. **在“局限与展望”中明确说明：**横断证据不足以支持因果推断，并提出未来研究方向。我们在“4.4 局限与展望”中补充并强化了因果推断边界：更重要的是，在横断设计占主导的情况下，元分析结果严格而言主要反映变量间的相关或关联强度，尚不足以支持严格的因果推断，也难以揭示二者关系随时间变化的动态过程与方向性（Maxwell & Cole, 2007）。未来研究需要更多纵向追踪与实验或准实验研究，以检验 AI 素养与自我效能的时间序列关系及其可能的因果路径。

意见 4：在给出 AI 素养的操作性定义后，建议进一步明确本元分析对“AI 素养”纳入的具体判定标准。同时，建议在剔除文献或纳入文献筛选阶段，说明概念内涵不一致或难以归类的研究大致数量与处理方式（如排除、单列、或用于敏感性分析）。

回应：感谢专家的重要建议。我们完全同意，AI 素养在现有研究中存在较大的概念界定与测量异质性，若不对构念边界进行清晰判定，将削弱效应量整合的可比性与结论的可解释性。为此，我们已对“AI 素养”的纳入判定规则与筛选过程进行补充说明。

修改位置说明：

引言 1.1 给出 AI 素养的操作性定义后，我们明确指出：“基于上述对 AI 素养内涵与核心维度的界定，需要进一步说明的是，在本研究的元分析框架中，并非所有涉及人工智能或相关技术的研究均被视为 AI 素养研究。为保证不同研究之间概念内涵的一致性与结果整合的可比性，本文在文献筛选与纳入阶段进一步制定了明确的 AI 素养判定标准，用以界定哪些变量可被纳入本元分析的“AI 素养”范畴。相关判定规则及其在文献筛选过程中的具体应用，将在方法部分的纳入与排除标准中作进一步说明。”，并说明本研究在 2.2 纳入标准和排除标准阶段制定了可操作的判定标准。具体而言，“本研究纳入荟萃分析的文献需满足以下条件：研究

对象为高等教育阶段学习者（本科生或研究生），样本来源不限国家、地区与专业；研究聚焦于人工智能素养（AI 素养）与自我效能之间的关系，且二者均被明确界定并作为主要变量进行分析；AI 素养与自我效能均采用具有信效度依据的量表或工具测量；研究报告了两变量之间的相关系数（如 Pearson's r 或 Spearman's ρ），或提供了可换算为相关系数的统计量（如 t、F、χ^2 等），确保效应量可提取；同时，文献需清晰说明样本量且样本量不少于 30，并为横断面相关研究或能够提取基线相关数据的纵向研究；文献语言限定为中文或英文，且可获取全文的学术研究成果（不包学位论文）。纳入 AI 素养的研究需明确以 AI literacy 为核心变量，且测量内容至少包含（知识理解、应用、评价、伦理）中两类维度；仅报告 AI 使用频率、态度、接受意向或一般数字素养且无法区分 AI 成分的研究予以排除。

在方法部分 2.3"选择程序"中，我们进一步报告了概念内涵不一致或指标难以归类为 AI 素养核心维度的研究数量及处理方式："在初步确定的 343 份研究中（见图 2），110 份重复研究被排除。在剩余的 233 份中，94 份根据标题和摘要被排除。其余 139 份研究在全文层面进行了筛选。在全文筛选阶段，部分研究虽涉及人工智能相关能力，但其 AI 素养的概念内涵或测量内容与本研究界定不一致，或相关指标难以归类为 AI 素养的核心维度。此类研究共 103 篇，主要为仅测 AI 使用或态度意向等与 AI 素养核心内涵不匹配，依据预先确定的纳入与排除标准予以排除。最终纳入 36 项研究进入荟萃分析。

意见 5：目前虽列出了关键词集合，但为满足系统综述/元分析的可复现要求，建议在在线附录中提供各数据库的完整检索式（包含布尔逻辑、检索字段、限定条件/时间范围等）、检索日期与更新日期，并说明去重与筛选流程细节，从而提高研究透明度与可核验性。

回应：感谢专家的重要建议。我们同意系统综述、元分析研究需尽可能提供可复现的检索与筛选信息，以提升研究透明度与可核验性。为回应该建议，我们已在在线补充材料中新增**附录 D"检索策略、去重与文献筛选流程"**，对原稿中仅呈现关键词集合的不足进行了补充说明，具体包括：在附录中分别给出了 Scopus（TITLE-ABS-KEY）、Web of Science（TS/Topic）与 CNKI（主题）的完整检索式，明确呈现布尔逻辑结构与检索字段设置，并同步报告各库的语言限定与时间范围（2023-01 至不限）。

补充首次检索日期与更新检索日期：我们在附录 S1 中报告了英文数据库的首次检索日期（2025-11-02）及更新检索日期（2025-11-06），并说明更新检索与首次检索采用一致的检索策略与限定条件，以确保纳入证据覆盖检索截止日前的最新研究。

补充去重与筛选流程细节：附录进一步说明了去重工具（Python/Excel/Zotero 6）及去重规则（优先按 DOI 匹配，对无 DOI 记录按"题名+第一作者+年份"匹配，并对自动去重结果进行人工复核）；同时详细描述了标题摘要初筛与全文复筛的两阶段筛选流程、分歧解决机制（两名研究者独立筛选—讨论一致—第三研究者裁决），并报告了编码一致性检验方法与结果（连续变量采用 ICC、分类变量采用 Cohen's Kappa），从而提高筛选与编码过程的可复现性与可靠性。

意见 6：建议在方法部分提前给出"自我效能类型划分"与"AI 素养测量工具分类"的明确标准与示例说明，以便读者在结果部分理解各分类术语的含义与来源，增强全文的连贯性与可读性。

回应：感谢专家的建设性建议。我们同意，结果部分涉及的“自我效能领域类型”与“AI 素养测量工具类型”等分类术语，如缺少方法部分的预先界定，会降低读者对分类含义及其来源的把握。为此，我们已在方法部分 2.4 数据提取中新增“2.4.1 测量工具标准与说明”，系统说明两类分类的判定依据（依据量表命名、条目情境与作者界定）、分类标准与典型示例，并补充多指标、边界个案的编码处理规则。上述补充使结果部分各分类术语能够在方法部分被清晰追溯，从而增强全文连贯性与可读性。

修改位置说明：方法部分 2.4 数据提取之后（新增 2.4.1 小节）。

2.4.1 测量工具分类规则

（1）自我效能领域类型划分

依据纳入研究对自我效能构念的命名、量表条目所指向的任务、情境以及作者对变量内涵的界定，本研究将自我效能划分为以下类型并用于后续亚组、调节分析：

一般自我效能（general self-efficacy）：测量个体在一般情境下的总体胜任信念（如一般自我效能量表）；学业、学习自我效能（academic、learning self-efficacy）：聚焦学习与学业任务完成信念（如学业自我效能、学习自我效能量表）；技术、AI 相关自我效能（technology- or AI-related self-efficacy）：自我效能明确指向技术任务或 AI 任务的学习、使用与完成（如 technology computer self-efficacy；AI self-efficacy、AI learning self-efficacy 等）。

（2）AI 素养测量工具类型分类

直接 AI 素养量表（direct AI literacy scales）：研究明确以“AI literacy、AI competence”等为核心构念，且量表内容覆盖 AI 素养关键维度（如知识理解、应用、评价、伦理等），多以自陈问卷形式测量；代理指标（proxy indicators）：研究未直接测量 AI 素养，而以相关概念或经验指标替代（如 AI readiness、AI capability、perceived AI competencies、或仅反映 AI、生成式 AI 使用经验、AI 教育、整合经历等）。

总体编码规则：若同一研究同时报告直接量表与代理指标，优先以与本研究定义更一致、且覆盖核心维度更充分的指标作为主要编码；边界个案由两名编码者回溯原文讨论达成一致，必要时由第三研究者裁定。

意见 7：文中提到在主效应合并前对同一研究的多个相关系数取平均以生成单一效应量，这一处理方式欠妥。通常应先将相关系数转换为 Fisher’s z 后进行平均，再转换回相关系数以用于合并分析。此外，若同一研究中多个效应量不独立，建议作者说明为何采用“取平均”策略，并考虑补充替代方案或敏感性分析，以降低依赖性处理带来的潜在偏差。

回应：感谢专家对同一研究内非独立效应量处理的提醒。我们同意：在同一研究存在多个相关系数时，不应直接在 r 尺度取平均。原稿中“取平均生成单一效应量”的表述确属笔误不严谨。我们已在修订稿中更正并明确：研究内合并时先将相关系数 r 转换为 Fisher’s z，在 z 尺度下进行平均，随后反转换为 r 用于合并分析。

同时，为降低研究内多效应量依赖性处理可能带来的潜在偏差，我们在方法部分补充说明了采用“独立样本为单位”的原因，并补充了稳健性分析：除主分析的处理方式外，可采用研究内聚合（within-study aggregation）方法处理；相关说明已写入修订稿相应位置，以提升

方法透明度与可核验性。

（1）修改位置： 2.4“数据提取与预设规则”（关于同一研究多效应量处理）。如图显示分析与研究内聚合分析非常接近。代表具有稳健性。具体原文修改如下：

为确保效应量的独立性，当同一研究报告多个相关系数（例如人工智能素养的不同维度与同一领域自我效能感的多项相关，或采用不同量表测量同一构念）时，主分析采用“核心效应量优先”策略：依据预先设定的优先级规则，仅保留一个与研究问题最匹配的核心效应量进入主效应合并，以避免同一研究被重复计权造成权重偏倚。为检验上述依赖性处理策略对总体结论的潜在影响，本研究进一步开展敏感性分析，并采用研究内聚合（within-study aggregation）：对同一研究内的多个相关系数先转换为 Fisher’s z，再在 z 尺度下按方差倒数进行加权平均生成单一聚合效应量，随后反转换为相关系数用于合并分析，并将该结果与主分析结果进行对照。其中，经过 python 数据统计分析结果发现（下图），此敏感性分析的结果与主分析结果基本一致，因此我们认为当同一研究报告多个相关系数的情况时，模型具有稳健性。

```
=== 数据预览（前10行）===
  study_id    effect_id    n outcome  ...     abs_r  is_core         z         v
0  Study_1   Study_1_E1  518    AI特定  ...  0.728741    False  0.926038  0.0019
2
1  Study_1   Study_1_E2  518    AI特定  ...  0.742826     True  0.956754  0.0019
2
2  Study_1   Study_1_E3  518    AI特定  ...  0.712012    False  0.891254  0.0019
2
3  Study_1   Study_1_E4  518    AI特定  ...  0.732696    False  0.934523  0.0019
2
4  Study_2   Study_2_E1  126    AI特定  ...  0.638841     True  0.756213  0.0081
0
5  Study_2   Study_2_E2  126     创造性  ...  0.640988    False  0.759849  0.008
30
6  Study_2   Study_2_E3  126     创造性  ...  0.643886    False  0.764783  0.008
30
7  Study_3   Study_3_E1  582    AI特定  ...  0.246740     True  0.251938  0.0017
7
8  Study_4   Study_4_E1  520    AI特定  ...  0.298203     True  0.307546  0.0019
4
9  Study_5   Study_5_E1  642     创造性  ...  0.145246    False  0.146280  0.0015
65

[10 rows x 10 columns]

每个研究效应量个数：
study_id
Study_1     4
Study_10    3
Study_11    3
Study_12    2
Study_2     3
Study_3     1
Study_4     1
Study_5     2
Study_6     2
Study_7     2
Study_8     2
Study_9     3
dtype: int64

=== 三种做法对照===
                                  model   k  ...       tau2         I2
0  WRONG: treat all effects independent  28  ...  0.072616  97.291953
1       MAIN: one core effect per study  12  ...  0.076327  97.497238
2        SA-1: within-study aggregation  12  ...  0.075595  98.844830

[3 rows x 7 columns]

提示：
- WRONG 一般会让某些研究因为报告了更多效应量而‘权重变大’（重复计权）。
- MAIN 与 SA-1 都能避免重复计权，但它们是两种不同的合理策略。
- 若 MAIN 与 SA-1 很接近，说明结论对依赖性处理策略‘稳健’。
```

意见 8：文中提及少数研究以标准化路径系数近似提取效应量，但路径系数往往为控制其他变量后的“条件效应”，与零阶相关并不等价，可能影响效应量可比性。建议在表 1 中明确标记此类研究，并开展敏感性分析（如剔除该类研究后重新估计总体效应），以检验主结论的稳健性。

回应：感谢专家指出路径系数与零阶相关在统计含义上的重要差异。我们同意：PLS/SEM 报告的标准化路径系数通常是在控制其他变量后的条件效应，严格而言不等同于零阶相关，可能影响效应量的可比性与合并解释。为提升透明度与可核验性，我们已在表 1 中对该类研究进行了明确标记（Duong, 2024；Soylu, 2025，均为 PLS）。同时，按建议开展敏感性分析：在主分析基础上剔除上述两项以路径系数近似提取效应量的研究后，重新进行随机效应合并。结果显示，剔除后总体效应量为 r = 0.533，95% CI [0.432, 0.594]（k = 34），与包含全部研究的主分析结果 r = 0.518，95% CI [0.505, 0.530]（k = 36）高度接近，结论方向与效应量大小均无实质性变化，表明本研究主结论对该类研究的纳入与否具有稳健性。

修改位置：（1）表 1 中对 PLS/路径系数近似提取的研究进行标注；（2）“剔除 PLS 研究后的敏感性分析”说明与数值截图：

```
============== RESTART: C:/Users/沐鱼/Desktop/元分析本研究/剔除两个 看看是否文件
.py ==============
=== 主分析（随机效应）===
{'k': 34, 'z': np.float64(0.5892664434909475), 'se': 0.058012752839236306, 'ci_z
': (np.float64(0.4755614479260443), np.float64(0.7029714390558506)), 'r': np.flo
at64(0.5293678208169748), 'ci_r': (np.float64(0.4426818907403638), np.float64(0.
6062504805461332)), 'tau2': np.float64(0.11238189152539874), 'Q': np.float64(264
3.297975628813), 'df': 33, 'I2': np.float64(98.75155959319532), 'pred_r': (np.fl
oat64(-0.0774030585057051), np.float64(0.8499828694410111))}

=== 子组随机效应结果 ===
  outcome   k         r    ci_low   ci_high         I2
0    AI特定  19  0.528824  0.402109  0.635662  99.017597
1   一般/学业   9  0.441209  0.300478  0.563137  97.189507
3   职业/创业   4  0.607502  0.295300  0.802481  98.879761
2     创造性   2  0.711744  0.483990  0.849189  97.657769

=== 结局类别调节检验 ===
{'Q_between': np.float64(215.91430512429451), 'df_between': 3, 'p_between': np.f
loat64(0.0)}

================= RESTART: C:\Users\沐鱼\Desktop\元分析本研究\敏感性分析.py ====
=============
=== 主分析（随机效应）===
{'k': 36, 'z': np.float64(0.5808305122446522), 'se': 0.05638201607632543, 'ci_z'
: (np.float64(0.4703217607350544), np.float64(0.69133926375425)), 'r': np.float6
4(0.5232688012499835), 'ci_r': (np.float64(0.4384592548499261), np.float64(0.598
8416780165027)), 'tau2': np.float64(0.11238768530017446), 'Q': np.float64(2750.9
978140619073), 'df': 35, 'I2': np.float64(98.72773435801747), 'pred_r': (np.floa
t64(-0.085265719088714 54), np.float64(0.8474780945094443))}

=== 子组随机效应结果 ===
  outcome   k         r    ci_low   ci_high         I2
0    AI特定  20  0.533812  0.412590  0.636455  98.972238
1   一般/学业   9  0.374433  0.240363  0.494485  96.258806
3   职业/创业   4  0.607502  0.295300  0.802481  98.879761
2     创造性   3  0.707454  0.593033  0.793868  95.836735

=== 结局类别调节检验 ===
{'Q_between': np.float64(372.6466220774905), 'df_between': 3, 'p_between': np.fl
oat64(0.0)}
```

且修改后的原文如下（3.研究结果 第一段）：

另外，有极少数研究未报告零阶相关，而是报告 PLS-SEM 的标准化路径系数；鉴于

路径系数为条件效应，我们已在表 1 中对该类研究进行标记，并在敏感性分析中剔除其后重新合并总体效应，结果显示，剔除后总体效应量为 $r = 0.533$，95% CI [0.432, 0.594]（$k = 34$），与包含全部研究的主分析结果高度接近，表示结论稳健。

意见 9：建议作者认真复核表 1 的变量列信息。当前表格中存在个别行“AI 素养量表”与“自我效能量表”列内容对调或填入错误的情况。

回应：感谢专家对表 1 信息准确性的提醒。我们已对表 1 中“AI 素养量表”与“自我效能量表”两列逐条回溯原文核对（包括量表名称、变量命名与测量描述），并对存在对调或误填的条目进行了更正与统一规范（如将误置于 AI 素养列的自我效能量表移回“自我效能量表”列，反之亦然；同时修正了个别拼写与命名不一致问题）。为避免类似错误再次出现，修订过程中由两名研究者分别独立复核表 1 信息并交叉校对，对有疑义的研究再次回溯全文确认后统一修订。修订后表 1 两列信息已与各纳入研究的实际测量工具一致，从而提升了表格的准确性与可核验性。

意见 10：目前贝叶斯模型缺少关键信息，建议补充报告先验设定、采样链数与迭代次数、收敛诊断指标（如 R-hat、有效样本量等）以及必要的模型检验/敏感性检验信息。此外，在贝叶斯框架下，建议提供与异质性相关的参数（如 τ 或 τ^2）及其 95% 可信区间，并进一步说明异质性比例相关指标的后验区间信息，以提高结果的可解释性与可复现性。

回应：感谢专家对贝叶斯部分可复现性与报告规范的专业提醒。我们同意原稿中关于贝叶斯模型的关键信息披露不足，已按建议对方法与结果部分进行了系统补充与规范报告。具体修改如下：

1. **补充先验设定与模型框架说明**：在统计分析方法中明确指出，在模型设定中，设定总体效应采用弱信息先验 $\mu \sim \text{Normal}(0, 1)$，异质性参数先验 $\tau \sim \text{HalfNormal}(0.5)$。同时补充说明选择弱信息先验的理由，以避免先验对后验结果施加过强约束并提高推断稳定性。

2. **补充采样设置与迭代信息**：在方法部分补充报告采样算法与关键参数：采用 NUTS 采样，chains = 4，draws = 6000，tune = 1500，target_accept = 0.95，并注明随机种子（seed = 123）。

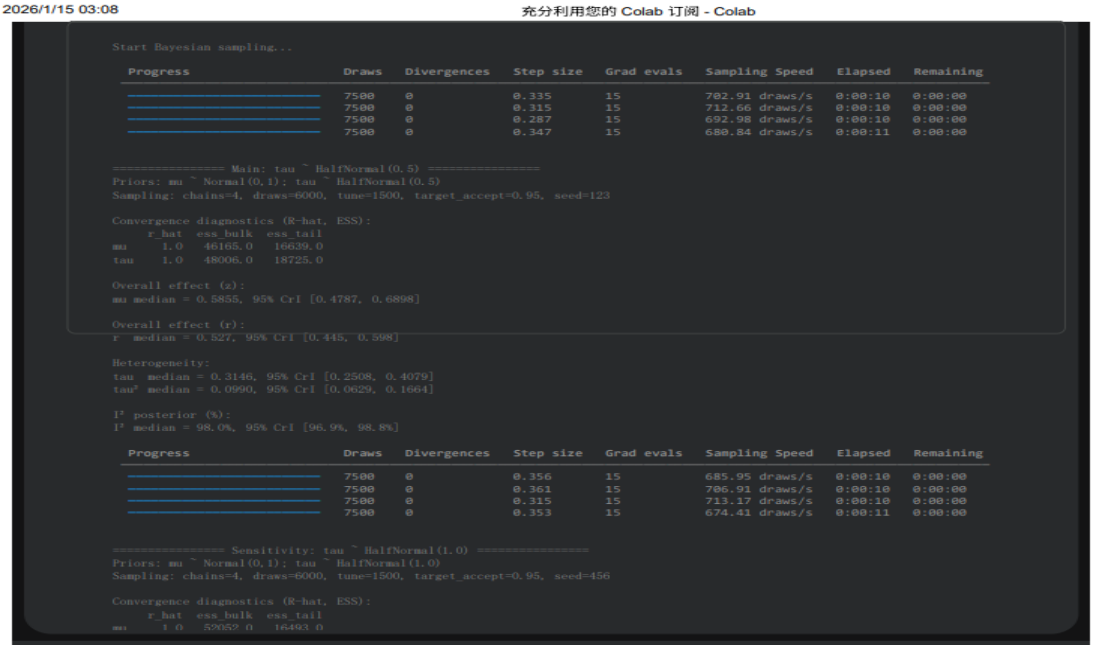
3. **补充收敛诊断指标**：在结果中补充 R-hat 与有效样本量（ESS）等收敛诊断信息。我们得到的关键参数均满足收敛要求（R-hat≈1.00），且 ESS 充足，同时未出现发散（divergences = 0），表明后验估计稳定可靠。

4. **补充模型检验与敏感性检验**：为检验结论对先验设定的稳健性，我们开展先验敏感性检验，将异质性先验由 $\tau \sim \text{HalfNormal}(0.5)$ 调整为 $\tau \sim \text{HalfNormal}(1.0)$（seed = 456），结果显示总体效应与异质性估计几乎不变，提示结论对先验设定具有稳健性。

5. **补充异质性参数与后验区间信息**：在结果部分新增报告 τ、τ^2 的后验中位数及其 95% 可信区间，并进一步给出异质性比例指标 I^2 的后验区间信息：τ 的后验中位数为 0.315（95% CrI [0.251, 0.408]），τ^2 为 0.099（95% CrI [0.063, 0.166]），I^2 的后验中位数为 91.73%（95% CrI [88.92%, 94.26%]）。

通过上述补充，我们完善了贝叶斯模型的先验设定、采样细节、收敛诊断、敏感性检验以及异质性参数与后验区间报告，使研究结果更具透明度、可解释性与可复现性。

修改位置：修订稿 2.6 统计分析（贝叶斯模型与采样信息）、结果部分 4.1（贝叶斯随机效应总体效应与异质性参数报告），以及在线附录、补充材料（收敛诊断、敏感性检验对照结果与必要的模型检验信息）。



修改原文：本研究首先依据各研究报告的效应量及其标准误进行经典固定效应合并（图4），以提供总体关联的基准估计。鉴于纳入研究之间存在显著异质性（例如 I^2 较高），传统的 DerSimonian–Laird 随机效应模型在异质性参数 τ^2 的估计上可能受研究数量与异质性程度影响而不够稳定，且对不确定性的刻画主要依赖正态近似（DerSimonian & Laird, 1986；Röver & Friede, 2022）。因此，为检验结论的稳健性并补充呈现不确定性信息，本研究以贝叶斯随机效应荟萃分析为主。该方法将每项研究的观测效应视为其真实效应的含误差估计，同时允许真实效应在总体平均效应周围合理波动，更契合心理学研究中效应可能随情境变化而呈现差异的特征（Röver, 2020c）。在模型设定中，总体效应采用弱信息先验 μ ~ Normal(0, 1)，异质性参数采用 τ~ HalfNormal(0.5)。随后使用多链采样（chains = 4，draws = 6000，tune = 1500，target_accept = 0.95，seed = 123），并以 R-hat 与有效样本量（ESS）评估收敛性。

贝叶斯随机效应模型结果显示，AI 素养与自我效能之间存在稳定的正相关：总体效应的后验中位数 r = 0.495，95%可信区间（CrI）[0.420, 0.570]。异质性较高：τ 的后验中位数为 0.315（95% CrI [0.251, 0.408]），τ^2 为 0.099（95% CrI [0.063, 0.166]），I^2 的后验中位数为 91.73%（95% CrI [88.92%, 94.26%]）。收敛诊断显示关键参数 R-hat≈1.00，ESS 充足，且未出现发散。为检验先验设定的影响，我们将 τ 的先验由 Half Normal(0.5) 改为 Half Normal(1.0)（seed =

456）进行敏感性检验，结果显示总体效应与异质性估计几乎不变，提示结论对先验设定具有稳健性。

意见 11：讨论中的部分解释超出了当前数据可直接支持的范围。例如，关于“地方执行差异、资源不均、教师培训”等社会文化因素的论述，在本研究所纳入的调节变量未直接测量相关情境指标的前提下，容易形成过度推断。建议将此类解释表述为“可能的原因……”，并强调其有待后续研究通过更细化的情境变量与研究设计加以检验，从而使讨论更加严谨。

回应：感谢专家对讨论部分推断边界的细致提醒。根据专家建议，我们认为主要集中在 4.3 研究意义：

我们同意原稿中关于“地方执行差异、资源不均、教师培训与社区支持”等社会文化因素的论述，若采用较为确定性的表述，可能超出本研究元分析数据能够直接支持的范围。根据建议，我们已对讨论中相关内容进行全面修订与规范化表达，主要修改包括：

1.措辞审慎化：将原先较为确定性的解释性表述统一调整为“可能、或许、一种可能解释是……”等概率性表达，避免将未被直接检验的情境因素作为确定结论。

2.补充证据边界声明：在讨论“研究意义”部分新增明确限定语，指出本研究纳入的调节变量并未直接测量资源投入、教师培训或政策执行差异等情境指标，因此相关论述仅为基于社会文化理论的可能解释，不能视为元分析可直接验证的证据。

3.强化未来研究方向：进一步补充说明后续研究可通过引入更细化的情境指标（如地区资源投入、学校数字化基础、教师培训强度与政策执行差异等），并结合纵向或多层研究设计，对上述解释路径进行实证检验，从而提高机制解释的证据强度与外推性。

修改原文：具体而言，尽管国家层面的 AI 素养政策在政策文件与顶层设计中往往提出了较为清晰的目标，且本研究发现 AI 素养与自我效能整体上呈显著正相关，但在不同地区的具体推进过程中仍可能出现落实方式与推进进度的差异。这些差异可能与资源投入、实施节奏有关，也可能与学校文化、教师专业能力以及社区支持等因素相关。社会文化理论强调，学习与发展发生在具体的社会文化情境中，个体的学习过程依赖“工具”以及与他人的互动支持。因此，一个可能的解释是：国家政策提供总体方向，但政策在不同地区转化为课堂实践的效果，可能会受到当地的资源条件、学校实践传统与支持系统的影响。

同时，不同地区在 AI 教育资源供给方面（如设备条件、培训机会与教学平台）存在差异，可能导致教师与学生可获得的学习支持不一致：部分地区已建立较成熟的数字化教学环境，而另一部分地区仍更多依赖传统教学方式。进一步而言，地方教育文化、学校管理理念与教师的技术认知差异，也可能影响政策落实的质量。从社会文化理论视角看，教师在 AI 学习中往往扮演关键的引导者角色（“更有经验者”，MKO），因此教师培训覆盖程度、教师对 AI 的理解与信念差异，可能会在一定程度上影响课堂层面 AI 素养培养的实际效果。

需要强调的是，本研究纳入的调节变量并未直接测量上述资源投入、教师培训或政策执行差异等情境指标，因此以上内容仅作为基于理论的可能解释，尚不能视为本次元分析可以直接验证的结论。未来研究可引入更细化的情境指标（如地区资源投入、学校数字化基础、教师培训强度与政策执行差异等），并结合纵向或多层研究设计，进一步检验这些情境因素在“AI 素养—自我效能”关系中的作用机制，从而为更有针对性的实践建议提供更直接的证据支持。

……

审稿人 2 意见：

本文以“大学生 AI 素养与自我效能的关系”为主题，采用 PROSPERO 预注册、PRISMA 2020 流程、频率学随机效应模型与贝叶斯随机效应模型相结合的方法，对近年全球 36 项实证研究（总样本量 N = 25,251）进行了系统整合。总体而言，论文选题具有较强的时代价值与学术意义，方法学规范性较高，数据整合规模在当前国内同类研究中处于明显领先水平，体现了作者良好的国际学术规范意识与较强的量化综述能力。

然而，作为主要刊登理论综述与进展性研究的《心理科学进展》，本文目前在理论整合深度、因果推断边界意识以及统计结论的谨慎性方面仍有显著提升空间。具体而言，论文整体更接近一篇方法学上非常完整的实证型元分析论文，而非以“理论争议—机制整合—未来研究方向”为核心的综述性论文。若能在不削弱方法学严谨性的前提下，进一步强化理论层面的整合、反思与命题化表达，本文具有较好的修改潜力。

关键问题与修改建议：

意见 1：调节变量分析的数量与统计可靠性问题。

作者在仅有 36 项独立研究的情况下，考察了多个调节变量（包括自我效能类型、AI 素养测量方式、国家/文化情境、研究设计、学生类别等），并报告了多项显著的组间差异。需要指出的是，在元分析研究中，是否适合开展多重调节分析，并不仅取决于总研究数，还高度依赖于各子组的研究数量（k）、测量同质性以及模型复杂度。从本文结果来看，部分关键结论建立在 k 极小的子组之上。在此条件下，即便观察到较大的相关系数，也难以排除抽样波动、情境特例或偶然放大的可能性。此外，本文同时进行了多项组间比较，但未对多重比较带来的 I 类错误风险进行系统控制或讨论，这在一定程度上削弱了显著性结果的解释强度。

建议作者在修订中：

1. 对调节分析结果进行可信度分级，避免对小样本子组结论作强解释；
2. 在方法或讨论中补充对子组研究数量不足、统计功效有限的明确说明；
3. 对多重比较问题作出方法学回应。

审稿意见 1.1：对调节分析结果进行可信度分级，避免对小样本子组结论作强解释

回应：（原文修改内容以标蓝）感谢审稿专家提出的这一重要方法学建议。根据该意见，我们对调节分析结果的呈现与解释进行了系统修订，在结果报告中引入了基于子组研究数量（k）的证据可信度分级（Evidence level），以区分不同调节效应的解释强度。具体而言，我们在调节变量分析结果表（表 3）中新增了“Evidence level”一列，并主要依据各子组所包含的独立效应量数量（k）对结果划分为较高可信度、中等可信度与探索性证据，同时在表注中对分级原则进行了明确说明。

此外，我们在正文第 4.2 节调节分析部分补充强调：鉴于调节分析对子组研究数量与估计稳定性的高度依赖，本研究在报告调节效应的同时，引入基于研究数量（k）的证据可信度分级（Evidence level），用于区分结果的解释强度而非统计显著性本身（Borenstein et al., 2009; Higgins et al., 2021）。具体而言，当某一子组包含的独立效应量较少时，即便观察到较大的相

关系数，其估计仍可能受到抽样波动或情境特例的影响，因此相关结果在解释时被视为探索性或趋势性证据。下文对各调节变量的解释，均结合其证据等级加以审慎展开。

本次回应所用到的参考文献：

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). Introduction to Meta-Analysis. *Introduction to Meta-Analysis*. https://doi.org/10.1002/9780470743386

Higgins, J. P. T., Thomas, J., Chandler, J., Cumpston, M., Li, T., Page, M. J., & Welch, V. A. (Eds.). (2019). *Cochrane Handbook for Systematic Reviews of Interventions*. Wiley. https://doi.org/10.1002/9781119536604

意见 1.2：在方法或讨论中补充对子组研究数量不足、统计功效有限的明确说明

回应：感谢审稿专家的建设性意见。根据该建议，我们在论文 4.4 的“局限与展望”部分对调节分析中子组研究数量不足及其统计功效限制进行了明确说明。具体而言，修订稿中指出，个别调节效应（尤其是国家与地区相关子组）所包含的效应量数量偏少，可能导致亚组比较与调节检验的统计功效不足，从而影响相关结论的稳健性（Cuijpers et al., 2021）。同时，修订稿进一步明确：对于基于效应量数量有限的子组所获得的调节结果，即便统计上达到显著水平，亦仅作为探索性或趋势性证据加以解释，而不作稳健的情境差异或机制性结论。通过上述补充，本研究在保留调节分析信息完整性的同时，对小样本子组结果的解释进行了必要限制，以回应子组样本量与统计功效不足可能带来的方法学风险。

本次回应所用到的参考文献：

Cuijpers, P., Griffin, J. W., & Furukawa, T. A. (2021). The lack of statistical power of subgroup analyses in meta-analyses: a cautionary note. *Epidemiology and Psychiatric Sciences*, *30*. https://doi.org/10.1017/s2045796021000664

意见 1.3：对多重比较问题作出方法学回应。

回应：感谢审稿专家对多重比较问题的关注。针对调节分析中同时考察多个潜在调节变量可能引入的 I 类错误累积风险，我们在修订稿中已从方法学解释层面作出明确回应。具体而言，与您的建议 1.1 一致。本研究在第 4.2 节调节分析部分，引入了基于子组研究数量（k）的证据可信度分级（Evidence level），用于区分调节结果的解释强度，而非仅依据统计显著性作判断（Borenstein et al., 2009；Higgins et al., 2021）。修订稿进一步明确指出：当某一子组包含的独立效应量较少时，即便观察到较大的相关系数，其估计仍可能受到抽样波动或情境特例的影响，因此相关结果在解释上被视为探索性或趋势性证据，并在后续讨论中予以审慎处理。通过该分级解释策略，本文在报告调节分析结果的同时，对多重比较可能带来的误判风险进行了有效控制，而未将单一显著性检验作为结论依据。

意见 2: AI 素养主观测量与“达克效应/自信偏差”的潜在问题。

本文纳入研究中，AI 素养与自我效能的测量均以自陈量表为主。在新一代的 AI 浪潮之下，哪怕是专业 AI 研究者，也很难自称具有较高素养。在整体发展呈现技术涌现、快速迭代、标准尚未充分明确的技术领域中，这一点需要引起高度重视。在新技术情境下，个体的“自评

能力”并不总能可靠反映其真实能力，反而可能受到过度自信、能力错判或校准偏差的影响。

在此背景下，AI 素养自评高分可能反映的并非真实的理解与能力，而是一般性的自信水平或积极自我概念，对 AI 能力与自身掌控程度的高估，或者与自我效能同源的主观信念系统。这将带来相关被系统性抬高以及心理机制解释被混淆的问题。

需要注意的是，作者的调节分析本身已呈现出一个值得警惕的模式：自编/整合型自陈量表的相关最高，而客观测验的相关最低。这一结果完全可能是方法效应的体现，而非构念关系本身的差异。

建议作者在修改时，注意在讨论部分明确区分“AI 素养的主观自评”与“AI 素养的客观表现/行为证据”，并承认当前证据主要支持前者；并且对 AI 素养自评可能受到过度自信或校准偏差影响进行正面讨论，而非仅作为局限性一笔带过，以及对结论措辞进行收缩，避免将“AI 素养自评与自我效能的相关”直接推演为“提升 AI 素养即可增强自我效能”。

回应：感谢审稿专家提出的这一重要理论与方法学问题。我们认同，在当前生成式 AI 技术快速演进、能力边界与评价标准尚未充分稳定的背景下，个体对自身 AI 素养的主观评估未必能够准确反映其真实理解与能力水平，且可能受到过度自信、能力错判或校准偏差等因素的影响。根据该建议，我们在修订稿讨论部分对这一问题进行了正面回应与明确区分。具体而言，修订稿中明确指出：本研究现有证据主要支持“AI 素养的主观自评”与自我效能之间的关联，而非 AI 素养客观能力或行为表现层面的关系（4.2）。同时，结合调节分析结果，我们进一步讨论了方法效应的可能性，即自编、整合型自陈量表呈现出更高相关，而客观测验类别的相关显著偏低，这一模式更可能反映同源测量与主观信念系统的影响，而非构念关系强弱本身的差异。此外，我们在讨论与结论部分对相关表述进行了收缩与限定，避免将“AI 素养自评与自我效能的相关”直接推演为提升 AI 素养即可增强自我效能的因果性或实践性结论，并明确指出未来研究有必要在主观自评之外，引入客观测验、行为指标或表现性任务，以更准确地区分能力信念、主观掌控感与真实 AI 能力之间的关系。**修改原文：**AI 素养的测量方式同样呈现显著调节效应。与成熟自陈量表和代理指标相比，自编、整合自陈量表得到的相关效应更大，而客观测验类别的效应较低。该模式可能反映了构念覆盖与方法效应的差异——在生成式 AI 技术快速演进、能力标准尚未充分稳定的情境下，个体对自身 AI 素养的主观评估还可能受到过度自信或能力校准偏差的影响，使其更多反应为主观能力感或与自我效能同源的信念系统，而非严格意义上的客观能力表现。因此，本研究的现有证据主要支持 AI 素养主观自评与自我效能之间的关联，而非 AI 素养客观能力或行为表现层面的关系。

意见 3：相关不代表因果

该问题与上面的问题相关，但值得作为一点单独提出。本文纳入研究以横断面相关研究为主，纵向与实验研究数量极少，且部分关键子组仅包含单一研究。在此条件下，论文在摘要、讨论及实践启示中多处使用接近因果含义的表述（如“提升 AI 素养有助于增强自我效能”），需要进一步谨慎。从现有证据来看，至少存在以下尚未排除的可能性：一、反向因果；二、第三变量解释。建议作者系统收缩因果性语言，明确本研究主要揭示的是关联结构而非因果机制；并在讨论或展望中专门增加一节，对未来因果研究路径进行具体说明。

回应：感谢审稿专家对因果推断边界的关键提醒。我们完全同意：鉴于本研究纳入文献以横断面相关研究为主，纵向与实验研究数量有限，且部分关键子组仅包含单一研究，本研究的元分

析结果在严格意义上主要揭示 AI 素养与自我效能之间的关联结构，而不足以支持明确的因果机制推断。在此条件下，反向因果关系及第三变量解释均尚未被排除。

根据该意见，我们对全文进行了系统性修订与补充。首先，在摘要、讨论及实践启示等部分，对涉及“提升、促进、增强”等可能引发因果解读的表述进行了逐一核查与收缩，统一改为“相关”“关联”或“联系”等更为审慎的表述，明确本研究结论所处的推断层级。其次，在理论讨论中，避免使用“AI 素养 → 自我效能”的确定性因果路径表述，而将相关关系界定为可能的关联方向或解释性假设框架。

此外，按照审稿专家建议，我们在讨论与展望部分新增 4.4.1“未来因果研究的可能路径”小节，提出未来研究可通过纵向追踪、实验或准实验设计，以及对潜在第三变量与中介机制的控制，进一步检验 AI 素养与自我效能之间的因果方向与作用机制（Maxwell & Cole, 2007）。通过上述修订，本文对自身证据类型与推断边界作出了明确限定，并为后续因果研究提供了可能的研究路径指引。

本次回应用到的参考文献：

Maxwell, S. E., & Cole, D. A. (2007). Bias in cross-sectional analyses of longitudinal mediation. *Psychological Methods*, *12*(1), 23–44. https://doi.org/10.1037/1082-989x.12.1.23

意见 4：建议在前言或理论部分增加对 AI 素养概念与测量分歧的系统梳理，明确不同操作化路径可能带来的心理学含义差异。特别是可能存在反向因果机制的理论可能性。

回应：感谢审稿专家提出的这一重要理论建议。根据该意见，我们在修订稿的前言、理论部分对 AI 素养的概念界定与测量分歧进行了系统梳理，并新增 2.4.1 测量工具分类规则这一章节进行具体说明，且明确指出现有研究中 AI 素养并非单一、稳定的构念，而是通过多种操作化路径加以界定，包括主观自评、代理指标以及客观测验等。修订稿进一步说明，不同测量方式可能对应不同的心理学含义：自评式 AI 素养更接近个体对自身能力与技术掌控感的主观判断，而代理指标与客观测验则更多反映学习经历或实际能力水平。同时，我们在理论讨论中正面提出了反向因果的理论可能性，即较高的自我效能或一般自信水平可能促使个体更积极地学习和使用 AI 技术，从而在主观测量中呈现更高的 AI 素养得分。通过上述补充，修订稿在理论层面区分了不同操作化路径的心理学内涵，并明确了 AI 素养与自我效能关系的潜在作用机制及其推断边界。

意见 5：方法部分可适度压缩技术实现细节（如具体软件库与采样算法），将重点放在方法选择的理论理由与适用边界。

回应：感谢审稿专家关于方法呈现的建设性建议。根据该意见，我们在修订稿中对方法部分的表述方式进行了调整，以提升可读性并突出方法选择与研究问题之间的理论关联。具体而言，在保证研究可复现性的前提下，我们适度压缩了关于具体计算环境、软件库与算法实现细节的描述，避免方法部分过度技术化；同时，将论述重点集中于分析策略本身的理论依据与适用边界。例如，修订稿明确说明在 Fisher’s z 尺度下采用随机效应模型与贝叶斯框架的理由，强调其用于整合相关系数并处理研究间异质性的理论优势，并指出相关结论应被理解为跨研究的平均趋势，而非确定性的因果机制。通过上述调整，方法部分在保持必要技术透明度的同时，更加突出分析方法的理论合理性及其解释边界，以回应审稿专家关于方法呈现取向的建议。

意见 6：调节分析结果需补充对小样本子组稳定性不足的说明，并对显著性结论进行分级解释。

回应：感谢审稿专家对调节分析解释稳健性的再次关注。针对小样本子组稳定性不足及显著性结果解释风险的问题，我们已按照此前您的建议 1.1 在修订稿中作出系统性修改。具体而言，修订稿在调节变量分析结果表（表 2、表 3、表 4）中引入了基于子组效应量数量（k）的证据

可信度分级（Evidence level）（Borenstein et al., 2009; Higgins et al., 2021），并在正文第 4.2 节中明确说明：对于 k 较小的子组，即便统计上达到显著水平，其结果亦仅作为探索性或趋势性证据加以解释，而不作强结论。相关分级原则与解释策略已在表注及正文中予以明确说明。

意见 7：讨论部分需显著加强对“主观 AI 素养—自我效能高相关是否可能源于自信偏差”的理论反思。

回应：感谢审稿专家对该理论问题的进一步提醒。根据这一建议，我们在修订稿讨论部分中显著加强了对“主观 AI 素养—自我效能高相关可能源于自信偏差或能力校准不足”的反思性讨论。修订稿指出（4.2），在生成式 AI 技术快速演进且能力标准尚未充分稳定的情境下，AI 素养的主观自评可能更多反映个体的一般性自信、技术掌控感或与自我效能同源的信念系统，而非真实能力水平。同时，我们结合调节分析结果进一步讨论了方法效应的可能性，即自编或整合型自陈量表呈现出更高相关，而客观测验类别的相关显著偏低，这一模式更可能反映同源测量与自信偏差的叠加效应，而非构念关系本身的差异。基于此，修订稿在讨论中对相关结论作出收缩性解释，强调现有证据主要支持主观层面的关联，并指出未来研究需通过客观测验、行为指标或纵向设计进一步区分真实能力、自信偏差与自我效能之间的关系。

意见 8：结论与实践启示中需避免直接的因果推断表述，并明确指出当前证据层级所能支持的推论范围。

回应：感谢审稿专家对结论与实践启示中因果表述边界的提醒。根据该意见，我们已对修订稿中的相关表述进行了系统性检查与调整。在结论与实践启示部分，已全面避免使用直接的因果性措辞（如“提升、促进、增强”等），转而采用“相关、关联、可能联系”等更为审慎的表述方式。同时，修订稿指出本研究所依据的证据主要来自横断面相关研究，当前结论应被理解为对 AI 素养与自我效能之间总体关联趋势的总结，而非因果机制的确认。相应地，实践启示被界定为方向性与启发性建议，其适用范围与解释力度需结合具体教育情境，并有赖于未来纵向或实验研究进一步验证。

第二轮

审稿人 1 意见

作者很好处理了上一轮审稿意见，对文章做出了比较全面的修改，修改后论文达到发表要求。建议作者进一步核查参考文献的格式及其引用规范。另外，经过上一轮修改，目前文中出现一些略显啰嗦和重复的内容，建议作者通读全文，在不改变文章表达内容的前提下，进一步凝练语言表述。建议修改后录用。

回应：尊敬的编辑老师、审稿专家：感谢您在第二轮审稿中对稿件提出的宝贵建议与肯定意见。我们已据此对全文进行复核与修订，重点完善参考文献格式及其文内引用规范，并在不改变研究结论与统计结果的前提下进一步精炼表述、减少重复。现逐条回应如下：1）参考文献格式与应用规范：已按期刊体例统一作者、年份、期刊名、卷(期)、页码与 DOI 等信息，并逐条核对文内引文与文后条目的一致性；如后续仍发现细节疏漏，我们将在校样阶段再次全面复核并及时更正。2）通读全文、凝练语言：已对多处表述偏长或重复的段落进行压缩与改写，主要涉及 2.1 搜索策略、2.4 数据提取、2.6 统计分析、3 研究结果（开头）、3.2 同质性检验、4.2 调节分析及 5 结论等（见稿件修订标注）。再次感谢您的审阅与指导。

审稿人 2 意见

作者已经回复了我的所有意见，我没有更进一步的问题了。

回应：感谢审稿专家对稿件前期修改的认真审阅与肯定。我们很高兴前期回复与修改已得到您的认可。再次感谢您对本稿完善所提出的宝贵意见。

第三轮

编委 1 意见：

作者针对专家意见进行了认真回复和修改，尚有以下修改建议：

意见 1. 全文无页码，不方便审稿人定位。

回应：感谢编委专家对此提出批评，由此带来的不便深表歉意，我们将此次认识养成习惯。现以在修回稿下方居中标注页码，我们也在后续的每一个修改回应条目中进行页码标注，方便专家查阅稿件与定位。

意见 2: 文中多处表述缺文献引用或依据，例如："从心理学视角看，AI 素养并非单一技术技能，而是包含知识理解、策略性使用与自我调控、风险与伦理判断及相关态度信念等复合结构。"

回应：针对"从心理学视角看，AI 素养并非单一技术技能，而是包含知识理解、策略性使用与自我调控、风险与伦理判断及相关态度信念等复合结构(尹开国，2024)。"添加尹开国的研究作为本句表述的参考文献，详见 **引言（p.42）**

意见 3: "社会认知理论（Social Cognitive Theory, Bandura）)认为，学习行为受"认知因素—行为因素—环境因素"三者的交互作用所影响。其中，个体对自身能力的判断，即自我效能（self-efficacy），是调节学习行为、策略选取、努力程度与坚持性的核心心理机制。"（文献引用？）

回应：针对 "社会认知理论(Social Cognitive Theory)认为，个体的认知因素、行为因素与环境因素之间存在持续的交互作用，学习行为及其结果正是在这种三元交互决定关系中形成和发展的(Bandura, 1986)。其中，自我效能作为个体对自身完成特定任务能力的主观判断，是影响其行为选择、策略使用、努力程度与坚持性的核心心理机制(Feng & Carolus 2026)。"同样，本段表述引用了 Bandura 和 Fang 等人的两份研究，详见：1.4 理论基础与研究问题（p.45）

意见 4: 本研究对自我效能划分类型缺对应文献引用。

回应：感谢编委专家关于本研究的部分段落缺乏文献引用与依据提出宝贵意见，据此我们对专家提出的三处缺失进行了检查。

针对"本研究自我效能划分类型缺乏对应文献引用"。我们对此深感歉意，是我们的疏忽。为此，我们在"**2.4.1 测量工具分类规则**"更新了本研究的"自我效能"与"AI 素养"划分类型的对应参考文献(Bandura, 1997；Schwarzer & Jerusalem, 1995；Midgley et al., 2000；Wang et al.202），在文内表述，并由表格综合展示。请见：**2.4.1 测量工具分类规则（p.48）**

本次回复所用参考文献：

尹开国. (2024). 人工智能素养：提出背景、概念界定与构成要素. *图书与情报*, (3), 60–68.

Bandura, A. (1986). *Social foundations of thought and action: A social cognitive theory*. Prentice-Hall.

Feng, S., & Carolus, A. (2026). Artificial intelligence literacy at school: A systematic review with a focus on psychological foundations. *Computers and Education: Artificial Intelligence, 10*, 100551. https://doi.org/10.1016/j.caeai.2026.100551

Bandura, A. (1997). *Self-Efficacy: The Exercise of Control*. Macmillan.

Schwarzer, R., & Jerusalem, M. (1995). Generalized self-efficacy scale. In J. Weinman, S. Wright, & M. Johnston (Eds.), *Measures in health psychology: A user's portfolio. Causal and control beliefs* (pp. 35–37). NFER-NELSON.

Midgley, C., Maehr, M. L., Hruda, L. Z., Anderman, E. M., Anderman, L. H., Freeman, K. E., Gheen, M., Kaplan, A., Kumar, R., Middleton, M. J., Nelson, J., Roeser, R., & Urdan, T. (2000). *Manual for the Patterns of Adaptive Learning Scales (PALS)*. University of Michigan.

Wang, X., & Li, P. (2024). Assessment of the relationship between music students' self-efficacy, academic performance and their artificial intelligence readiness. *European Journal of Education, 59*(4), e12761. https://doi.org/10.1111/ejed.12761

编委 2 意见：这篇稿件的写作随着修改有所进步，但存在如下问题：

意见 1：文章没有说清楚为什么必须探讨两个变量的关系，并做元分析，理论价值不足。一种能力和对能力的效能感有中等相关，这个有什么需要质疑的。另外这种相关，是特异性的吗，为什么一种领域的素养，要和一般效能感建立关系？而且有趣的是，AI 素养竟然不是和 AI 效能感相关最高，而是 AI 素养与创造性自我效能、职业或创业自我效能的关联相对更强，如何解释？作者没有做出有效的理论解释。

回应：感谢编委专家对本研究的理论贡献方面提出要求，我们将对围绕所提问题逐项对照、逐点回应，并对论文相应内容（主要涉及前言、结果和讨论部分）进行了针对性修改与完善。现将修改与完善的结果以论述的方式回应专家。详见本次回应稿件：**前言（p.42）；结果（p.52）；讨论（p.60）**

就“为何有必要探讨 AI 素养与自我效能的关系”而言，**一是**本研究并非仅关注两个变量之间是否存在一般相关，**而是试图回答一个更具理论意义的问题：作为智能时代的重要复合素养，**AI 素养能否进一步**转化为个体在学习与发展中的积极心理资源。二是**自我效能**作为社会认知理论中的核心变量，反映个体对自身能力的主观判断，并直接影响其目标设定、努力投入、坚持性及面对新技术任务时的行为选择。因此，考察 AI 素养与自我效能之间的关系，不仅有助于揭示 AI 素养能否由“知识—技能层面”进一步转化为“能力信念层面”**，从而为理解 AI 素养**作用于学习适应、持续投入**与**创新发展**的心理机制**提供依据**；二是也有助于回应国家推进“**人工智能+**”与**人工智能赋能教育变革的现实要求，为高校人才培养中如何将 AI 素养优势进一步转化为学生发展优势提供理论支持。

就“为何有必要采用元分析”而言，**一是**现有研究虽已积累了一定数量的经验结果，但在 AI 素养和自我效能的**概念界定、测量工具、样本来源、教学情境以及研究设计等方面存在明显差异**，导致研究结论呈现一定分散性，**单项研究难以稳定回答二者关系的总体强度及其边界条件**。甚至在相近研究领域中，**AI 素养与自我效能之间的关系也并非始终呈现单向增强趋势**。有研究发现，**AI 素养未必与 AI 自我效能表现出最强相关；相反**，较高水平的 AI 素养有时**会使个体更加清楚地意识到 AI 系统的能力边界、任务复杂性以及自身与技术之间的差距，从而对自身能力产生更为谨慎的判断，甚至增强其能力不确定感。并且，没有一项研究是针对本领**

域开展的跨国元分析，据此可以通过本研究来发现中国与其他国家（地区）的结构差异。其次，在心理学元分析领域，采用贝叶斯方法进行效应检验的研究仍较为有限。**与传统频率学方法相比**，贝叶斯方法在面对较高异质性时具有独特优势。尤其是在纳入研究中存在部分小样本的情况下，传统频率学方法在参数估计稳定性和不确定性表达方面可能受到一定限制，而贝叶斯方法能够结合后验分布、采样链特征与模型拟合结果，对总体效应及异质性参数进行更为充分的估计，从而以概率形式呈现真实效应可能落入的区间范围。

就“理论价值不足”而言。最后，就理论价值而言，与回答 1 相呼应，本研究的贡献并不止于再次验证 AI 素养与自我效能之间存在相关，而在于从更高层次澄清了这一关系的性质、边界及其理论含义。首先，本研究的元分析结果表明，AI 素养与自我效能之间存在稳定的正向关联，说明 AI 素养并非仅停留于知识理解与工具使用层面，而是能够进一步转化为个体对自身能力的积极判断，从而为理解 AI 素养如何作用于学习适应、持续投入与发展表现提供了心理机制层面的证据。其次，本研究发现，AI 素养与不同类型自我效能之间的关联强度并不一致，其与创造性自我效能、创业自我效能的关系强于 AI 特定自我效能和一般学业自我效能。这一结果表明，AI 素养的作用并不限于提升学生对 AI 任务本身的能力判断，而更可能通过增强创造生成、职业适应与未来发展中的掌控感和胜任预期，激活更高阶的效能信念，从而推动 AI 素养研究由“技术能力视角”进一步转向“心理资源与发展资本视角”。再次，本研究进一步发现，不同学生类别之间存在显著差异，整体上呈现出教育、传媒、语言类及综合类学生的相关强度高于理工、计算机类学生的趋势。这一发现提示，AI 素养的价值并不必然随着专业技术背景的增强而同步提高，理工背景也未自动转化为更强的效能关联。换言之，AI 素养的作用机制可能并不主要表现为对既有技术能力的简单强化，而更可能体现为对开放性任务、意义建构、创造生成与职业想象等高阶发展过程的支持。这一结果在一定程度上突破了“领域越接近、效能关联越强”的直观预期，为理解 AI 素养何以在不同专业情境中发挥差异化作用提供了新的解释空间。最后，本研究还发现国家样本类别并未表现出显著调节效应，而测量方式、研究设计与学生类别具有显著调节作用，这说明 AI 素养与自我效能之间的联系具有一定跨文化稳定性，但其作用程度又受到具体学习情境与实施条件影响。由此，本研究不仅回答了二者‘是否相关’，还进一步回答了‘在何种类型的自我效能、何种学生群体及何种研究条件下更强或更弱’，从而推进了该领域对作用边界与情境敏感性的理论认识。

意见 2：讨论中还在分析结果，为什么不放在结果部分呢。现在的讨论没有有效解释结果何以如此，算不上讨论。

回应：感谢编委专家的意见。针对“讨论部分仍停留于结果分析、缺乏对结果何以如此的有效解释”的问题，我们已对第三章“研究结果”与第四章“讨论”进行了进一步区分和调整。具体而言，在修订稿中，第三章主要呈现元分析的客观结果，包括总体效应、异质性、稳健性检验及调节效应分析；第四章则不再重复罗列统计结果，而是在结果基础上重点围绕其理论含义、形成原因及可能机制展开解释。详见：**结果（p.52）；讨论（p.60）**

意见 3：摘要说：调节分析显示，效能类型与 AI 素养测量方式显著影响效应大小。应该说明如何调节。

回应：感谢编委专家的提醒。根据该意见，我们已对摘要中“调节分析显示，效能类型与 AI 素养测量方式显著影响效应大小”的表述作进一步细化，不再仅笼统说明“存在显著调节”，而是明确交代了调节的具体方向和表现形式，即：AI 素养与创造性自我效能、创业自我效能的关联强于 AI 自我效能和一般学业自我效能；自编或整合式自陈量表所得效应高于直接 AI 素养量表、代理指标和客观测量；不同学生类别之间亦存在差异，其中教育、传媒、语言类学生

的关联相对较强，而理工、计算机和商管类相对较弱。通过上述修改，摘要已更清晰呈现调节效应“如何调节”的具体内容。详见： **摘要 （p.41），英文摘要（p.73）**

意见 4：还存在较多细节问题。如文中太多不必要的英文。参考文献格式混乱。

回应：感谢编委专家指出稿件中的细节问题。关于文中英文表述较多的问题，前稿中部分量表名称、测量指标及操作化表述直接沿用了国外文献原文，未及时转换为规范中文表述，确属我们的疏忽。根据意见，我们已对相关内容进行了系统检查与调整，现已尽可能统一改为规范中文表述。关于参考文献格式混乱的问题，前稿中确实存在部分文内引文字体未统一、括号格式不一致以及半角全角混用等情况。对此，我们已再次对照《心理科学进展》的著录要求，对全文文内引用和参考文献格式进行了逐项核查与统一，现已将文内引用括号统一为半角括号，并同步规范了字体、标点和格式细节。

主编意见：稿件经过多位专家的审阅，作者进行了认真的修改，达到了发表水平，同意发表。