arXiv:2512.21045v1 [math.HO] 24 Dec 2025

人工智能时代本科数学中的合成流畅性与认知卸载

Siyuan Wang[1†]、QingXia[1†]、QiongYe[1*†]

[1]温州肯恩大学科学、数学与技术学院，中国浙江省温州市，325060。

*通讯作者。电子邮箱：yeqiong@wku.edu.cn；其他作者：
siywang@kean.edu；qxia@kean.edu；
†这些作者对本研究的贡献相同。

摘要

生成式人工智能（AI）工具在高等教育中的迅速普及，正在改变学生参与本科数学学习的方式，也引发了人们对学习效果和评估效度的担忧。本研究考察人工智能可及性在跨两个学期、涵盖多门课程的数据集中的影响，课程包括商科微积分、线性代数和微积分 III。通过比较无监考作业与有监考考试的表现，我们分析学生在可使用人工智能的环境中的学习行为如何变化，尤其关注数学任务的认识论卸载。在社会认知框架的指导下，我们采用多种互补指标——成绩差距、作业—考试相关性和瓦瑟斯坦距离——来刻画练习与掌握程度之间的偏离。结果显示，随着课程内容从程序性数学转向概念性和空间密集型数学，诚信差距不断扩大。在商科微积分和线性代数中，作业形式（线上与手写、由助教评分）的差异并未导致实质性的表现模式差异，这表明纸笔作业并非天然更能抵御人工智能辅助的任务卸载。虽然作业在程序性课程中仍具有一定的预测效度，但高年级课程中作业与考试之间的一致性显著瓦解，这表明无监考评估越来越多地反映了合成式流利度，而非内化理解。这些发现凸显了在人工智能时代重新思考评估实践的必要性。

关键词： 生成式人工智能、本科数学教育、评估效度、作业—考试一致性、教学契约

1 引言

ChatGPT、Claude 和 Gemini 等大语言模型（LLM）的迅速发展，加上日益成熟的符号计算系统，从根本上改变了本科数学课堂。过去构成学生日常学习任务的内容，例如代数运算、矩阵计算、微分与积分，乃至常规的多元微积分运算，如今都能以极少的用户输入即时、准确地完成。随着这些工具日益普及，它们不再处于边缘位置，而是从结构上融入学生在课堂内外学习数学的方式。

这一技术变革在数学教育中引发了日益加剧的张力。一方面，人工智能可以充当支架工具Wood、Bruner和Ross(1976)，通过提供分步骤讲解、可视化内容和即时反馈来支持学习。谨慎使用此类工具，可能有助于降低认知门槛、减轻焦虑，并使学生能够专注于更高层次的推理。另一方面，人工智能也可能成为一种替代性工具，代替学生完成核心数学任务，并可能绕过传统上构建理解所需的过程。人工智能作为辅助工具与作为替代品这两种角色之间的界限，已变得越来越难以区分，在无人监督的学习环境中尤为如此。

这种张力触及大学数学教学目标这一长期争论的核心。传统课程体系一直在程序流畅性与概念理解、计算与建模、抽象和结构推理之间寻求平衡。如今，生成式人工智能和符号人工智能系统执行许多程序性任务的能力已达到与本科生相当或更高的水平Frieder等人 (2023)，这引出了一个根本问题：如果机器能够可靠地执行我们所教授的程序，那么对学生而言，什么才应被视为数学知识与能力?

关于高等教育中生成式人工智能的现有研究，主要考察了工具采用、学生态度、学术诚信问题或短期表现影响 Fuchs (2023);Sullivan、Kelly和McLaughlan(2023);Xie、Wu和Chakravarty (2023)。尽管这些研究为人工智能辅助作弊的风险以及基于检测的应对措施的局限性提供了重要见解，但它们在很大程度上将人工智能视为一种需要加以管理的外部干扰，而不是一种从根本上重塑课程作业和评估之认识论功能的力量。近期的概念性研究开始对这一观点提出挑战。Li和Wang Li和Wang(2023)认为，在生成式人工智能时代，传统作业已经失去了其教育意义，因为任务完成可以完全委托给人工智能系统，因此有必要重新设计作业和评价，使其转向人类–人工智能协作及过程导向评估。同样，Leaton Gray等人 Leaton Gray、Edsall和Parapadakis (2025)认为，人工智能不仅助长不当行为，还暴露出当代评估体系更深层的结构性弱点，因此应从监控驱动的诚信模式转向以效度为中心、以伦理为基础的教学法。然而，量化人工智能的可用性如何重塑不同类型数学课程中形成性评估与总结性评估之间一致性的实证证据仍然有限。

本研究通过一项为期两个学期、涵盖多门课程的实证调查来弥补这一空白，研究范围包括商科微积分、线性代数和微积分 III。这些课程经过有意选择，旨在呈现从程序性主导的数学内容，到概念性结构化且高度依赖空间思维的内容的递进过程。我们利用详细的评估数据，考察学生在人工智能工具随时可用的情况下，其表现如何演变，尤其关注无监考评估（作业）与有监考评估（考试）之间的差异。这一比较使我们能够探究诚信差距是否会随着时间推移而出现或扩大，从而反映学生学习行为的变化，而非个别的滥用事件。

本研究由以下研究问题引导：

研究问题1 人工智能工具的可用性如何影响不同类型本科数学课程中的学生表现？**研究问题2** 这些影响在侧重程序运算的课程与以概念或空间结构为主的课程之间是否有所不同？**研究问题 3** 作业表现与考试表现之间的差异如何反映学生学习策略及其对人工智能依赖程度的变化？**研究问题 4** 这些模式对课程设计、评估实践以及本科数学教育未来目标有何启示？

本研究将人工智能视为一种颠覆性的认识论力量，而非教学小工具，旨在为这一快速发展的讨论提供实证证据。了解人工智能如何重塑学生的所作所学，以及评估实际衡量的内容，对于重新思考人工智能时代的数学教育至关重要。

2 理论框架

为严谨分析生成式人工智能对学生学习和评估结果的影响，本研究采用社会认知视角，不仅将人工智能视为中性的技术辅助工具，更将其视为一种认知主体；其教育意义取决于学生如何将其纳入自身实践。我们综合运用工具生成理论、实用价值与认识论价值之分，以及教学契约，以阐释人工智能如何重塑数学活动。

2.1 人工制品与工具：生成与最近发展区

我们首先区分作为人工制品和作为工具的人工智能系统（例如 ChatGPT）。人工制品是供学习者使用的物质或符号工具；它本身并不能保证学习发生。根据Verillon 和 Rabardel (1995) 提出的工具生成理论，只有通过一个发展过程 —— 学习者将工具整合到自身的认知活动中——人工制品才会成为学习工具。

工具性生成包含两个相互交织的过程。工具化指人工制品影响使用者思维和行为的方式。例如，当人工智能自动提供基于坐标的解法、符号运算，

或完整的推导过程时，学生可能不再尝试对向量进行可视化、进行几何推理，或预测其定性行为。从这个意义上说，工具通过重新界定何为正当的数学行动，重塑了学生的认知图式。

相比之下，工具化是指学习者如何调整并塑造工具，使其服务于数学目标，例如学习如何限制提示词、要求提供符合形式定义的解释，或将人工智能的输出用作验证和批判的对象。如果要让人工智能发挥学习工具的作用，而非仅仅生成答案，工具化至关重要。

我们的核心主张是，在本研究中的许多学生身上，工具性生成仍未完成。人工智能并未成为数学工具，而是充当了义肢式人工制品，代替Risko 和 Gilbert (2016)所述的学生完成数学工作。在这些情况下，工具化占据主导地位，而工具化过程中的另一面——学习者对工具的塑造——则相对不足：人工智能取代了核心认知劳动，而非为其提供支架。因此，不能假定任务完成，尤其是在无监考环境下，能够反映学生已内化的数学理解。

这一现象与 Skemp 对工具性理解（遵循规则却不知其缘由）和关系性理解（既知道怎么做，也明白为什么）所作的区分密切相关Skemp (1976)。生成式人工智能促成了一种可称为合成流畅性的现象：学生无需参与传统上有助于发展关系性理解的推理过程，也能产出正确答案和润色完善的解答。这种合成流畅性威胁着核心数学实践，包括推理、表征和抽象。

传统上，最近发展区（ZPD）维果茨基 (1978) 表示独立完成任务的能力与在指导下完成任务的能力之间的差距。我们认为，人工智能可能会人为地扩大最近发展区。人工智能能够即时解决超出学生实际能力范围的问题，模仿出学生具备相应能力的假象，从而造成诚信差距。学生在无监考作业中看似能在最近发展区内完成任务，但在有监考考试中，其表现便会退回实际的独立能力水平，暴露出所谓的脚手架实际上只是一种假肢。

2.2 实用价值与认识论价值

为了说明这种影响为何因课程而异，我们采用 Artigue 对工具的 实用价值与认识论价值的区分 Artigue (2002)。实用价值是指工具帮助学生完成任务的效能，通常表现为高效地产生正确答案。认识论价值是指工具支持概念理解、意义建构以及数学知识构建的能力。

这种情况在不同课程中呈现出不同面貌：在商科微积分中（实用价值较高），任务通常具有算法式特征（例如，求解方程组）。人工智能能够高效完成繁琐的计算，因此具有很高的实用价值。虽然这有可能绕过程序性技能习得，但认知需求往往较低，学生仍能核验输出结果，并可能将人工智能用作提高效率的工具。

线性代数（混合机制）处于一种混合认识论机制之中，其特点是抽象程度高，且需要协调符号、几何和结构等多种表征。虽然人工智能系统能够高效执行矩阵计算和常规代数运算，但对于培养关系性理解的支持有限，而这种理解是推理向量空间、线性变换和证明所必需的。在这种情境下，人工智能通常充当转换或执行工具，而非概念支架；它能够生成正确的符号输出，却无法确保学生理解底层结构。因此，学生在线性代数中对人工智能的依赖极不稳定：对于计算导向任务，这种依赖可能无碍，但当评估转向概念性或结构性推理时，就会干扰学生的认识过程。

微积分 III（认识论黑箱）面临着一道独特的障碍：复杂代数与高阶三维空间推理相结合。与“视觉任务需要人工干预”这一假设相反，材料本身的难度往往超出学生独立认知能力的范围。因此，这门课程极易出现黑箱外包。当学生无法想象通量积分或进行曲面参数化时，他们不会采用部分支架，而是将整个问题交给人工智能。在这里，依赖人工智能并非出于便利，而是出于无奈，这也造成了假肢效应风险最高的情形——认识论价值被彻底消除。

这一区分直接引出了本研究的核心假设：无监考作业与有监考考试之间最大的差距，将出现在人工智能的实用价值很高、但认识论参与度很低的情况下。作业表现可能反映了对人工智能的成功实用性运用，而考试表现则揭示了学生并未内化概念图式。

2.3 教学契约的瓦解

最后，我们将这些动态置于 Brousseau 的教学契约 Brousseau(2002) 的框架中。传统上，布置一道题目隐含着这样的共识：解题需要运用其背后的数学知识。

生成式人工智能通过满足任务的可见条件——给出正确答案——却未实现其教育目的，打破了这一契约。这与 Brousseau 的 Topaze效应相符：答案是通过绕过预期学习过程的方式得出的。在人工智能的介入下，这一效应被系统化：学生给出了正确结果，但数学知识仍置身事外，完全存在于人工制品之中。

这一框架表明，人工智能充当着一种认知主体，重新界定了人类必须承担的工作。如果学生将程序性操作和概念建模都外包给人工智能，他们便放弃了自己在认知过程中的角色，使评估成绩衡量的成为工具的能力，而非学生的理解程度。

因此，评估产物，尤其是作业，可能不再是学习成效的可靠指标。在本研究中，作业与考试表现之间的偏离，因而不应仅被解释为学术诚信问题，而应被视为教学契约出现更深层断裂的证据：传统的能力信号已不再与实际的认知发展相对应。

3 方法论

本研究历时两个学期，涵盖三门本科数学课程，这些课程经过选择，代表认知与认识论谱系中不同的位置：商科微积分、线性代数和微积分 III。这些课程是经过有意选择的，旨在考察人工智能介导的问题求解的影响是否会随着所强调的数学知识与实践的性质而系统性变化。

这些课程构成了从以程序性内容为主的数学到对概念理解和表征能力要求较高的数学的递进，为考察人工智能依赖和学习成果的差异模式提供了天然的研究场景。

3.1 参与者

这三门课程的学生群体在族裔上高度同质，全部由中国学生组成，这减少了先前教育背景中的跨文化差异。然而，由于中国实行不同的高中数学分流，同一群体内部仍存在显著差异。在中国的中等教育体系中，学生分流至理科班和文科班项目，两者在数学学习的强度和严谨程度上存在明显差异。

修读商科微积分的学生，尤其是主修会计学和管理学的学生，主要来自文科班，通常在高等代数和微积分方面接受的正规训练较弱；而修读线性代数和微积分 III 的学生则更可能来自理科班，接受过显著更扎实的大学前数学训练。重要的是，这些背景差异在各学期之间保持稳定，因此无法解释高年级课程中观察到的作业—考试一致性骤降。

- 商科微积分课程的学生主要是大一本科生，专业集中在会计学、金融学、管理学及相关商科领域。每个班级通常约有 50 名学生。
- 线性代数课程的学生主要包括大一数据科学专业学生和大二计算机科学专业学生，另有少数会计学和金融学专业学生，以及几名重修该课程的高年级学生。班级人数通常在 25 至 30 人之间。
- 微积分 III 课程的学生主要是大二数据科学专业学生，另有少数重修课程的高年级学生。班级人数与线性代数课程相似，通常在 25 至 30 人之间。

研究所涉及的课程均由同一位授课教师教授，从而确保教学风格、评估设计和评分标准的一致性。

3.2 数据收集

评估数据通过两种主要方式收集，体现了标准的本科生评估实践。

- 自动化评估（WebWorK）Prat 和 Code（2021）；Roth、Ivanchenko 和 Record（2008）；Toth （2013）：三门课程都使用了 WebWorK 这一在线作业发布系统。该平台会为每道题生成随机化数值参数，确保所有学生使用的方法保持一致，但具体答案各不相同。这可以防止学生简单地互相分享答案，但无法阻止能够处理特定参数的人工智能解题工具。微积分 III 的作业评估完全依赖 WebWorK。

- 人工评估（人工评分）：商科微积分和线性代数课程包含人工评分的作业部分。学生需要提交书面解题过程，理论上这要求他们展示解题步骤。然而，现代人工智能工具（例如具备视觉能力的 GPT-5.2）如今能够生成逐步书面解答，因此这种形式容易受到基于誊抄的学术不端行为影响。

所有课程的期中考试和期末考试均在校内、于监考条件下进行，期间不得使用人工智能工具或其他外部计算辅助工具，明确允许的工具除外（例如获准使用时可使用基本计算器）。

3.3 变量

本研究的主要自变量区分在于无监考与监考评估环境。作业（线上和书面）被视为无监考，使用人工智能的可能性较高。虽然未禁止合作和使用外部资源，但既未明确鼓励使用人工智能，也未对其进行系统性限制。期中考试和期末考试被视为有监考，使用人工智能的机会很少或几乎没有。这些评估旨在衡量学生在时间限制下独立掌握数学的能力。主要因变量是在各门课程中学生作业和考试的标准化分数。

3.4 测量指标

在第2节的理论框架指导下，我们预先设定了一组互补的分析指标，旨在捕捉生成式人工智能出现后，练习与掌握程度之间的关系可能出现的不同断裂方式。我们没有依赖单一的汇总统计量，而是采用了一个在数据分析前定义、并在各课程和学期中统一应用的多维分歧框架。

这一方法基于两点考量。首先，工具生成理论预测，人工智能可能会对任务完成和概念掌握产生不同影响。其次，实践—认识论区分表明，干扰不仅可能表现为平均分数差异，还可能体现为形成性评估与总结性评估之间关联减弱、结构性碎片化。因此，预先定义了三个结果指标。

3.4.1 表现差值

第一个指标是原始表现差值，对每位学生定义如下：

$$\Delta = \text{Average Score}_{\text{Homework}} - \text{Score}_{\text{Exam}}$$

这一指标捕捉了无监考表现与有监考表现之间最直接的差异，反映学生在可使用人工智能的环境中取得的成功，在多大程度上能够转化为受控条件下的独立表现。

我们强调， Δ 被视为基线描述指标，而非学习损失的决定性指标。事先纳入该指标，是为了与现有评估比较保持连续性，并为下文所述的其他指标提供参照。

3.4.2 作业的预测效度

第二项指标通过计算每门课程、每个学期内学生作业平均分与考试成绩之间的斯皮尔曼秩相关系数（r），评估作业的预测效度。采用斯皮尔曼相关系数，是为了降低作业分数天花板效应带来的影响。

在传统教学环境中，作业预期发挥指导性练习的作用，从而与考试表现呈正相关。基于这一假设，我们将该相关关系的强度用作指标，以判断作业是否仍与独立掌握程度保持一致。

这种关联的减弱——本文称之为相关性崩塌——被解读为证据，表明作业表现已不再可靠地反映概念理解。重要的是，预先选定这一指标，是为了检测义肢式人工智能使用所预测的脱钩效应，而不受总体分数水平或分布形态的影响。

3.4.3 分布形态

为捕捉无监考表现与有监考表现之间超越集中趋势的结构性差异，我们预先设定了一项基于瓦瑟斯坦距离的分布形态指标。该指标量化了将一个分数分布转变为另一个分布所需搬移的“质量”，因此能够敏锐地反映离散程度、偏度和两极分化的变化。

令 F_H 和 F_E 分别表示给定课程和学期内作业成绩与考试成绩的经验累积分布函数（CDF）。鉴于评分量表的离散特性（$k \in \{0, 1, \ldots, 100\}$），我们计算瓦瑟斯坦距离Villani （2008） W_1 ，使用CDF的 L_1 范数：

$$W_1(H, E) = \sum_{k=0}^{100} |F_H(k) - F_E(k)| \,.$$

瓦瑟斯坦距离 W_1 衡量将作业分布转化为考试分布所需的代价。该指标对两极分化尤为敏感，

使我们能够区分难度普遍增加（整体平移）与学生群体内部出现分化（双峰性）：前者是难度均匀上升，后者则是一组学生仍保持掌握程度，另一组学生则表现崩溃。

选择这一指标是事先基于三个原因。首先，与均值差异不同，瓦瑟斯坦距离对作业分数中的天花板效应具有稳健性；在实用型人工智能使用率较高的情况下，预计会出现这种效应。其次，它能够敏感地反映异质性变化，包括离散程度增加或子群体的出现，且无需对单峰性或正态性作出假设。第三，它提供了一个易于解释的单一标量，用于概括分布差异，同时保留有关完整分数分布的信息。

在这一框架中，较小的瓦瑟斯坦距离表明作业表现与考试表现在分布上相吻合，这与作业能够有效帮助学生为独立掌握做好准备的情况一致。较大的距离则表明练习与评估之间存在显著错位，这与人工智能介导的无监考表现膨胀，以及监考条件下掌握程度碎片化的情况一致。

重要的是，瓦瑟斯坦距离作为一项补充诊断指标，与原始表现差值和作业—考试相关性并列使用。 Δ反映幅度，相关性反映预测一致性，而瓦瑟斯坦距离则反映不同评估体系之间的结构性差异。

通过在分析前定义这三项测量指标，并为每项指标奠定具体的理论预期，本研究将事后指标选择的风险降至最低。这些指标并非可互换的替代选项，而是正交诊断指标，共同刻画学生表现的幅度、一致性和结构。后续所有分析均在不同课程和学期中一致地应用这一框架。

4 数据分析与结果

本节展示对所研究的三门课程进行的实证分析结果。对于每门课程，我们使用方法论中定义的三项指标，分析无监考（作业）与有监考（考试）表现之间的差异：原始表现差值（Δ）、预测效度相关性（r）以及分布形态（W_1）。所有数据分析和图表生成均在 R 中完成。

4.1 商科微积分

商科微积分的数据采集自2025年春季学期。商科微积分的结果展现了一门课程的特点：人工智能使用似乎主要出于实用目的——用于提高效率，而非完全取代认知活动。

图1展示了商科微积分评估成绩的分布。A班的作业通过 WebWorK 平台布置，而 B班的作业则由助教手动批改。尽管两个班收到的习题集完全相同，WebWorK作业中的数值参数均经过随机化，以减少答案共享。

如表1所示，两个班的学生作业平均成绩都接近满分（A班为 98.5；B班为 96.6）。表中的 SD 表示标准

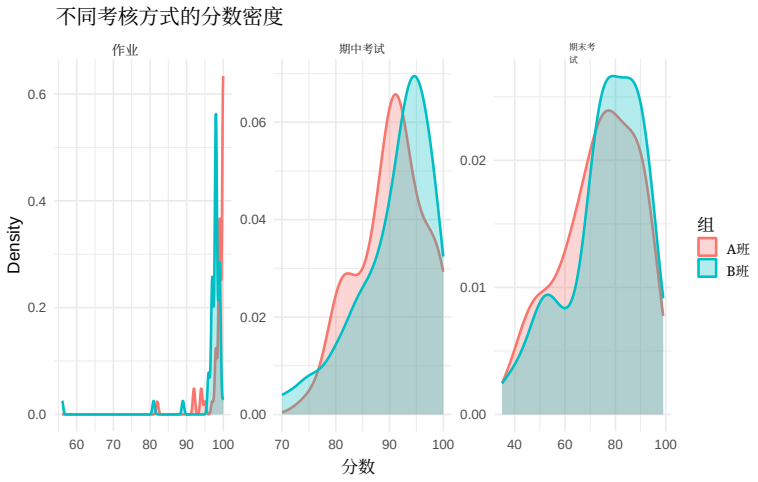

图1 商科微积分两组的密度图

表1 商科微积分的均值 ±标准差

组	作业	期中考试	期末考试
A班	98.5 ± 3.0	90.4 ± 6.3	73.4 ± 15.3
B班	96.6 ± 6.5	91.1 ± 7.0	75.7 ± 15.4

差。然而，在有监考的期末考试中，成绩显著下降（分别为 73.4 和 75.7）。A班有 $\Delta = 25.1$ 分，B班有 $\Delta = 20.9$分。虽然这一差距相当明显，但它反映出一种显著但相对一致的成绩下滑，这在学生使用工具绕过繁琐计算、但仍保留程序性认知的课程中十分常见。另一方面，期中考试的差距相对较小Δ's。

表2 商科微积分的斯皮尔曼相关系数

组	作业–期中考试	作业–期末考试	期中考试–期末考试
A班	0.508	0.410	0.604
B班	0.484	0.546	0.495

尽管原始分数有所下降，作业与考试表现之间的关系仍具有统计显著性。表 2 报告了作业成绩与期末考试成绩之间的斯皮尔曼相关系数，分别为 $r = 0.410$ （A班）和 $r = 0.546$ （B班）。这种中等程度的正相关表明，作业表现较好的学生通常在考试中也表现较好，这意味着作业练习仍具有一定的预测性练习价值。

表 3 商业微积分的瓦瑟斯坦距离 s

组	作业–期中考试	作业–期末考试	期中考试–期末考试
A班	8.1	25.2	17.1
B班	7.0	20.9	16.1

作业分布与期末考试分布之间的瓦瑟斯坦距离（W_1）在A班为25.2，在B班为20.9。如密度图（图 1）所示，考试分布向左偏移（分数较低），但仍保持相对单峰的形态。这种整体偏移表明，虽然没有工具时课程内容更难，但全班的表现相对一致，学生群体内部没有出现明显分化。

商科微积分结果还揭示了另一点：作业提交方式——在线提交还是纸笔作答——并不会导致实质上不同的表现模式。尽管提交形式不同，A班通过 WebWorK 完成作业，而 B班提交手写、由助教评分的解答，但两组的作业平均分都同样偏高，作业—考试相关性也相近，作业分数分布与考试分数分布之间的瓦瑟斯坦距离几乎相同。

这表明，在以程序性内容为主的课程（如商科微积分）中，评估媒介本身并不能显著减轻或加剧人工智能介导的成绩膨胀。即使要求学生提交书面作业，当代人工智能工具也能生成易于抄写的逐步解答，因此纸笔作业并不比在线平台更能抵御认知卸载。因此，商科微积分中观察到的诚信差距，似乎主要由任务的认知属性而非作业提交形式所驱动。

4.2 线性代数

线性代数的数据采集自2025年春季和2025年秋季学期。线性代数的结果呈现出一种混合模式，有效地弥合了商科微积分的务实稳定性与微积分 III 的辅助性崩溃之间的差距。这门课程既要求计算熟练度（矩阵运算），也要求抽象推理（向量空间、证明），无监考与有监考表现之间的一致性变化很大。在线性代数在本数据集中最能揭示诊断信息，对群体特征和评估任务的认知性质都表现出极高的敏感性。

线性代数是本研究核心假设的关键转折点——或压力测试。不同于主要以程序性内容为主的商科微积分，以及高度侧重概念与空间性的微积分 III，线性代数处于中间的认识论位置。其评估结果会因评估重点是计算熟练度，还是概念性与结构性推理，而出现显著波动。

这种分化在各学期中都清晰可见。在2025年春季，当期末考试侧重计算导向任务时，无监考作业与有监考考试之间的成绩差异仍与商科微积分相当，诚信差距 Δ 通常在20–36分之间。相比之下，在2025年秋季，当期末考试转向概念导向问题时，诚信差距显著扩大。对于E班、F班和G班， Δ 超过48分，与微积分 III中观察到的成绩崩塌相当，甚至更为严重。

这一明显的不连续性表明，线性代数的内容并非仅仅处于中等难度，而是在认识论层面上具有不稳定性：从程序性计算向概念抽象的细微偏移，就足以导致作业—考试一致性崩溃。因此，线性代数最清楚地证明，人工智能介导的诚信差距并非仅由课程层级驱动，而是由评估任务的认识论性质所决定。

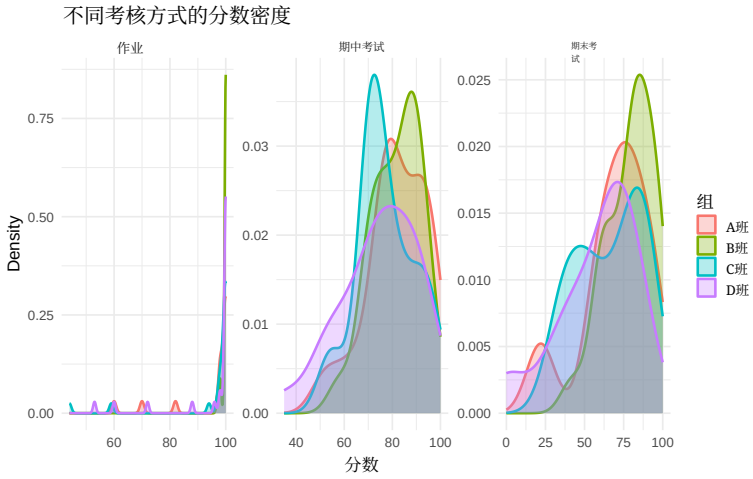

图2 线性代数（2025年春季）四个组的密度图

图2 和图3 分别展示了2025年春季和2025年秋季线性代数课程评估成绩的分布。B班和D班的作业通过 WebWorK 平台布置，而A班和C班的作业则由授课教师手工批改。A班和C班的习题集取自教材习题，而 WebWorK 作业则包含随机化的数值参数。2025年秋季，所有班级（E班、F班和G班）的作业均通过 WebWorK 平台布置。

每个学期的作业和期中考试题保持一致，学生在作业和期中考试中的表现也相对接近。然而，2025年春季的期末考试侧重计算，而2025年秋季的期末考试侧重概念，因此2025年秋季的期末成绩显著低于2025年春季（表4）。

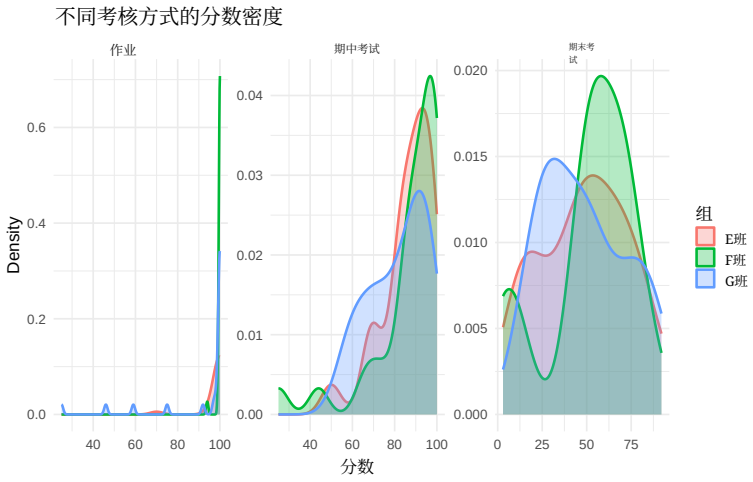

图3 线性代数三个组的密度图（2025年秋季）

表4 按学期划分的线性代数均值 ± 标准差 r

学期	组	作业	期中考试	期末考试
2025年春季	A班	94.3 ± 12.0	81.6 ± 12.5	69.6 ± 21.9
	B班	99.7 ± 0.80	81.7 ± 10.0	79.2 ± 15.1
	C班	94.8 ± 14.3	77.1 ± 12.1	65.7 ± 21.0
	D班	94.8 ± 12.7	74.6 ± 15.8	58.2 ± 25.0
2025年秋季	E班	98.3 ± 6.30	86.3 ± 12.3	48.7 ± 24.5
	F班	99.8 ± 1.15	86.7 ± 17.9	51.2 ± 24.7
	G班	91.3 ± 19.9	81.5 ± 13.7	48.8 ± 23.6

与其他课程相似，各组的作业平均分始终接近满分，范围为 91.3 至 99.8（表 4）。然而，诚信差距显著扩大，与商科微积分相比。2025年春季学生群体（A、B、C 和 D 班）的期末考试（以计算为主）平均分为 58.2 至 79.2，因此表现差值（Δ）约为 20–36 分，与.....相似在商科微积分的案例中。2025 年秋季队列（E班、F班和G班）中，这种差距变得极为严重：虽然作业分数仍然很高（98–99），但期末考试（侧重概念）平均分骤降至48–51，差距超过48分，这与微积分 III的情况相似。这种鲜明的反差表明，尽管有些学生群体设法弥合了差距，但另一些却遭遇了类似高年级微积分III学生所经历的认知支架失效。

线性代数中作业的预测能力最好可描述为不稳定。与相关性始终处于中等水平的商科微积分不同，线性代数的相关性呈现出极大的波动。2025年春季（表5 A、B、C、D班，侧重计算的期末考试）中，A班的作业与期末考试的相关性保持在中等水平（$r = 0.461$），但C班呈现负相关（$r = -0.176$），

表5 按学期划分的线性代数斯皮尔曼相关系数

学期	组	作业–期中考试	作业–期末考试	期中考试–期末考试
2025年春季	A班	0.718	0.461	0.649
	B班	0.209	0.372	0.553
	C班	−0.061	−0.176	0.722
	D班	0.581	0.516	0.613
2025年秋季	E班	0.641	0.471	0.577
	F班	0.305	0.252	0.420
	G班	0.582	0.283	0.256

这意味着，作业分数较高的学生在独立评估中的表现反而更差——这是Topaze效应的典型特征，即依赖工具反而会阻碍学习。从2025年秋季（表5 E班、F班和G班的概念导向期末考试）来看，相关性进一步减弱，F班和G班仅呈现较弱的关联（$r \approx 0.25$–0.28），表明评估的效度已出现问题。

表6 各学期线性代数的瓦瑟斯坦距离

学期	组	作业–期中考试	作业–期末考试	期中考试–期末考试
2025年春季	A班	12.6	24.6	12.0
	B班	18.0	20.5	4.6
	C班	18.5	29.0	11.7
	D班	20.2	36.6	16.3
2025年秋季	E班	12.1	49.7	37.6
	F班	13.1	48.0	34.9
	G班	13.5	42.5	32.7

Wasserstein距离度量证实了这种过渡状态。2025年春季（表6 A班、B班、C班和D班），W_1 数值介于20.5至36.6之间，通常高于商科微积分，但低于微积分 III，更接近商科微积分。然而，2025年秋季的数据（表6 E班、F班和G班）显示，W_1 数值飙升至42.5–49.7。这一数值范围与微积分 III相当。正如密度图（图2 的模式与图1和图3 的模式更像图4）直观呈现的那样，作业分布仍在100分处形成尖峰，而考试分布则变得平缓并扩散至不及格区间。这表明线性代数处于一个临界点：根据学生群体和具体内容（例如计算型还是抽象型），这门课程可能从一个易于驾驭的工具辅助环境，转变为认知完全卸载的状态。

2025年春季和秋季期末考试的样题收录于附录A。

4.3 微积分 III

微积分 III 的数据采集自 2025年秋季学期，且作业仅通过 WebWorK 发布。与商科微积分相比，微积分 III 的结果表明，非监考评估与监考评估之间的关系发生了根本性崩解，这与假肢人工制品假说一致。

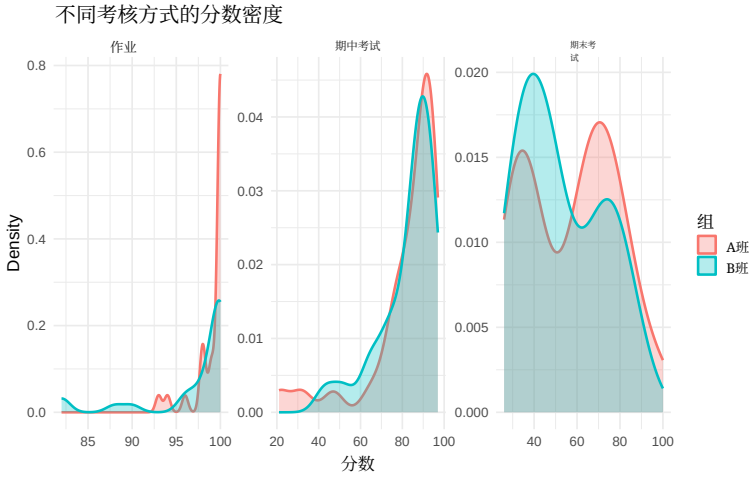

图 4 微积分 III 两个组的密度图

图 4 显示，微积分 III 中 A班和B班的分布均呈现出明显的双峰和两极化特征。一部分学生保持着中等考试表现，表明其仍具备独立能力；而相当一部分学生群体在监考条件下表现不佳。这种模式与异质化的人工智能参与模式相符，其范围从受限支持到大量认知卸载不等。

表7 微积分 III 的均值 ±标准差

组	作业	期中考试	期末考试
A班	99.1 ± 1.8	81.6 ± 18.4	57.3 ± 21.3
B班	97.1 ± 5.4	82.5 ± 13.7	52.7 ± 19.2

表7 显示，作业平均分仍处于顶端（A班为99.1；B班为97.1），与低年级课程相比在统计上无显著差异。然而，有监考期末考试的分数大幅下降至不及格的平均水平。A班平均分为57.3，得分为 $\Delta = 41.8$ 分。B班平均分为52.7，得分为 $\Delta = 44.4$ 分。这一分化程度几乎是商科微积分中所观察到的两倍，表明对于空间复杂材料而言，移除人工智能工具导致能力严重丧失。

表 8 微积分 III 的斯皮尔曼相关系数

组	作业–期中考试	作业–期末考试	期中考试–期末考试
A班	0.290	0.273	0.693
B班	0.188	0.391	0.535

作业的预测效度在本课程中实际上已消失。表 8报告了作业与期末考试的相关性：A班为 $r = 0.273$ ，B班为 $r = 0.391$ 。在A班，作业与期中考试之间的相关性低至0.290。这种微弱关联意味着，学生的作业成绩已不再是其对材料独立掌握程度的可靠指标。

表 9 微积分 III 的瓦瑟斯坦距离

组	作业–期中考试	作业–期末考试	期中考试–期末考试
A班	17.5	41.7	24.9
B班	14.6	44.4	29.8

这种结构性分化在瓦瑟斯坦距离上最为明显，其数值飙升至 41.7（A班）和 44.4（B班），在本研究观测到的数值中名列前茅。图 4 中的密度图直观地证实了这种断裂。作业分布（最左侧面板）在 100 分处形成尖峰，而期末考试分布（最右侧面板）则完全趋于平坦，扩散至整个分数区间，并在不及格区域（< 60）呈现显著密度。这种大规模的“概率质量”转移证实，高作业分数是人为造成的，掩盖了学生在监考环境下所需的概念理解方面的严重不足。

5 讨论

本研究结果为以下假设提供了实证支持：人工智能对学生学习的影响并非一概而论，而是受课程内容认识论价值的调节。通过对比商科微积分的实用性学习环境与微积分 III 的认识论要求，我们观察到两种截然不同的工具性生成模式。

5.1 假体效应与诚信差距

本小节通过考察人工智能相关的成绩差异如何在程序性导向的数学课程与概念要求较高的数学课程中呈现出不同形式，来探讨研究问题1和研究问题2。最直接的发现是无监考评估与有监考评估之间的成绩差异幅度。在商科微积分课程中，我们观察到成绩大幅下降，但仍在可控范围内：作业平均成绩为 97 分，

期末考试平均成绩为 74 分。虽然这表明学生依赖外部辅助工具，但作业与考试之间的相关性中等（$r \approx 0.50$），说明作业至少在一定程度上仍发挥了练习作用。在此，人工智能很可能发挥了务实的作用——自动完成计算，同时保留了部分程序性结构。

在线性代数这一谱系的中间地带，它发挥着过渡领域的作用。数据揭示出一个不稳定的局面：诚信差距波动很大。在一些班级（春季班 B）中，差距与商科微积分（$\Delta \approx 20$）中尚可控制的水平相当；而在另一些班级（秋季班 E、F、G）中，差距则扩大到与微积分 III（$\Delta > 48$）中灾难性的水平相当，甚至更高。这种差异支持了这一假设：线性代数既包含实用性任务（人工智能能够高效完成，且不会损害学习），也包含抽象型概念任务（人工智能在其中充当义肢）。如果课程过于偏重后者，又缺乏监考，那么支架便会崩塌。

相比之下，微积分 III 显现出严重的诚信差距。尽管作业平均分几乎保持完美（98），学生在期末考试中的独立表现却骤降至 55 分。这 40 分的差距表明，对于空间和概念上复杂的任务，人工智能发挥的不是支架作用，而是义肢式人工制品的作用。它取代了核心认知过程，而非助力其发展。这证实了我们的假设：需要高阶可视化和概念建模的课程，反而比依赖死记硬背的程序性课程更容易受到彻底外包的影响。

一个重要的跨课程发现是，作业形式在缓解人工智能介导的表现差异方面作用有限。在商科微积分和线性代数两门课程中，使用在线作业平台（WebWorK）的班级与要求提交手写、人工评分的作业的班级，其作业平均分相近，作业—考试相关性相似，分布差异指标也几乎没有区别。这一模式表明，纸笔作业通常被认为能够促使学生更多参与认识论活动，但在当代人工智能工具面前，并不能显著恢复评估效度。尤其是在传统上被认为手写解答对概念学习至关重要的线性代数课程中，尽管要求提交完整的书面作业，诚信差距依然存在。人工智能生成的逐步解答很容易被照抄，因此评估媒介与认知卸载这一根本问题在很大程度上彼此无关。因此，观察到的作业—考试一致性失效，与其说是由作业实施方式造成的，不如说更能归因于任务本身的认识论要求。

5.2 课外评估效度的削弱

本小节通过分析作业表现与考试表现之间的差异如何反映学生参与、学习策略以及对人工智能依赖程度的变化，来回答研究问题 3。最令人担忧的发现或许是高年级数据中观察到的相关性崩塌。在传统教学法中，作业表现是考试准备度的预测指标。我们的数据显示，在人工智能时代的概念性课程中，这种关联已经断裂。

在商科微积分中，作业与期中考试成绩之间的相关性仍具有统计学意义（$r \approx 0.50$）。在微积分 III 中，这种预测效度

消失了（$r \approx 0.19$–0.29）。这种缺乏相关性的现象表明，在微积分 III 中，学生的作业分数不再衡量其自身能力，而是衡量工具的能力。线性代数为这种效度衰减提供了“铁证”。C班中观察到的负相关（$r = -0.176$）尤其发人深省。在这种情况下，出色的作业表现不仅毫无参考价值，反而与掌握程度呈反向关系。这表明存在一种依赖陷阱：最勤于使用人工智能完成作业的学生，独立推理能力反而最弱。与契约依然有效的商科微积分、契约已经断裂的微积分 III 不同，线性代数显示契约正在瓦解——效度不再有保障，并且波动难以预测。

这代表了教学契约的根本性破裂。学生履行了契约的条款（提交正确答案），却没有履行其精神实质（参与认识过程），从而产生了Topaze效应：正确答案掩盖了完全不存在的学习。

5.3 分布断裂

从分布角度来看，这项分析进一步支持研究问题 3，表明人工智能介导的学习损失是结构性的，而非均匀发生的。通过瓦瑟斯坦距离对分布形态进行分析，进一步揭示了这种学习损失的结构性本质。微积分 III（$W_1 \approx 42$-44）中作业与期末考试分布之间的距离几乎是商科微积分（$W_1 \approx$21–25）的两倍。

线性代数的分布形态进一步展现了这一转变。2025 年秋季队列（42–49）中的 Wasserstein 距离（W_1）实际上超过了微积分 III，表明两极分化极为严重。这说明，即使在混合型课程中，学生群体也正在分裂为两个截然不同的组：一类学生利用工具进行学习（保持分布形状），另一类学生利用工具替代劳动（使分布趋于扁平）。

微积分 III 中如此高的瓦瑟斯坦距离反映了学生群体的两极分化。密度图显示，尽管作业分数仍集中在上限（合成流畅性），考试成绩却变得平缓，并大幅向下偏移。这表明部分理解这一中间状态正在消失；学生要么能够独立掌握材料，要么完全将认知劳动交给人工制品，导致他们在监考评估中束手无策。

5.4 对课程设计的影响

本小节直接回应研究问题 4，将观察到的评估错位转化为人工智能时代课程设计与评估实践方面的启示。数据表明，仅靠提高难度来让作业不受人工智能影响，是一种失败的策略。事实上，提高概念难度（如线性代数和微积分 III）似乎反而增加了对义肢式人工智能使用的依赖。对于程序性课程（商科微积分），重点应从计算转向阐释。由于人工智能擅长算法式执行，评估应检验学生能否结合具体情境理解人工智能的输出。对于概念性课程（线性代数和微积分 III），

传统的课后问题集实际上已不再适合作为总结性评估指标。如果保留无监考作业，就应根据过程（例如视频讲解、口头答辩）而非成果评分。

6 结论

本研究旨在追踪生成式人工智能在认知谱系各层面对本科数学教育的影响。研究结果表明，人工智能已不再是边缘工具，而是成为了核心认知主体，从根本上改变了学生参与的性质。

在不同课程中（研究问题1），我们观察到无监考作业表现与有监考考试表现之间存在系统性分化，这表明人工智能的可及性从根本上改变了评估成绩所代表的含义。然而，这些影响并不一致。分化的程度、结构和教育意义因课程类型而显著不同（研究问题2）。在商科微积分等程序导向型课程中，人工智能的使用似乎主要出于实用目的：作业对考试仍具有中等预测效度，说明练习与掌握之间存在部分一致性。相比之下，在微积分 III 等对概念理解和空间思维要求较高的课程中，人工智能的使用主要变成了替代性支撑，导致作业分数近乎满分，而考试表现却严重崩盘。

作业表现与考试表现之间的这些差异（研究问题 3）揭示了学生参与和学习策略更深层次的失灵。高年级课程中作业—考试相关性的崩塌以及观察到的较大瓦瑟斯坦距离表明，无监考评估越来越多地衡量的是合成流畅性，而非内化理解。在这种情况下，作业表现反映的是工具的能力，而非学生的能力，这表明教学契约已然破裂。

综合来看，这些发现对人工智能时代的课程设计与评估具有重要启示（研究问题 4 ）。仅仅提高任务难度或恢复传统纸笔作业，并不能减轻对人工智能的依赖，在某些情况下反而可能加剧认知卸载。在商科微积分和线性代数这两门课程中，手写作业都未能维持练习与独立掌握之间的一致性，这表明，仅靠改变形式无法恢复评估效度。为维护数学评估的完整性，应将无监考作业重新归类为形成性练习，而非掌握程度的总结性证据，同时更加重视有监考且以过程为导向的评估，要求学生进行解释、推理，并对概念理解负责。

这些结果挑战了这样一种假设：人工智能在最近发展区中仅仅充当“导师”或“脚手架”。相反，对许多学生而言，人工智能会人为扩大最近发展区，营造出能力已经提高的假象，而一旦移除工具，这种假象便会立即消失。为维护数学教育的完整性，我们建议对评估策略进行结构性调整：

•降低无监考任务的权重：作业应被视为形成性练习（按完成情况评分或采用低风险评分），而非衡量掌握程度的标准。
• 提升监考环境的重要性：总结性评估必须回归受控环境，以确保衡量的是人类认知，而非合成流畅性。

•过程导向评估：在概念理解至关重要的课程中，必须评估学生解释结果为何成立的能力。与生成结果本身相比，这项任务目前仍更难被完全卸载。

归根结底，人工智能本身并不会对数学教育造成危害，但它确实带来了颠覆性影响。正如我们关于诚信差距的数据所表明的，在人工智能时代之后仍沿用人工智能出现前的评估模式，无法衡量学生学习，反而会越来越多地衡量机器的能力，而非学习者的能力。

7 局限性与未来工作

本研究存在若干局限性，需予以说明。首先，数据来自单一院校，且学生群体相对同质，这限制了研究发现对不同院校类型和文化背景的普适性。其次，本研究属于观察性研究，而非实验研究；尽管无监考与有监考评估之间的差异在理论上有据可依，在实证上也稳健可靠，但若不直接测量工具交互，就无法对个别学生的人工智能使用作出因果判断。具体而言，缺少直接的人工智能使用日志，使我们无法细致地将个体表现模式归因于特定的人工智能参与方式。第三，本研究以作业分数作为人工智能可及表现的代理指标。该指标能够反映总体影响，却无法在个体层面区分对人工智能的有效支架式使用与完全认知卸载。

因此，未来研究应采用混合方法设计，包括对照实验、人工智能使用记录、出声思维协议和学生访谈，以厘清人工智能参与的不同模式如何影响学习。还需要开展涵盖更多课程和院校的纵向研究，以确定随着学生和课程体系逐渐适应人工智能辅助的数学学习，观察到的诚信差距会趋于稳定、扩大还是缩小。

致谢。本研究部分由温州肯恩大学和浙江省教育厅资助，合同编号为JGBA2024540和JGBA2024546。

声明

利益冲突
作者声明不存在任何利益冲突（包括财务和非财务利益冲突）。

经费支持
本研究部分得到温州肯恩大学和浙江省教育厅的资助，资助合同编号为JGBA2024540和JGBA2024546。

伦理批准与参与同意
本研究未对人类参与者进行医学或心理学实验，因此无需正式的伦理批准。所有分析数据均在常规教育实践过程中生成，并在分析前进行了匿名化处理。

发表同意不适用。

数据可用性
当前研究中生成和分析的数据集，可在提出合理请求后向通讯作者获取。

代码可用性
数据分析所用代码可在提出合理请求后向通讯作者获取。

材料可用性不适用。

作者贡献
所有作者均参与了本研究的构思和设计。数据收集与分析由作者完成。所有作者均撰写并严格修订了稿件，并批准提交的最终版本。

附录A 线性代数期末考试样题

A.1 2025年春季（计算导向）

1. 假设 $A = \begin{bmatrix} x & -2 & 2 \\ 3 & y & -1 \end{bmatrix}$ 有一个特征向量$\mathbf{v} = \begin{bmatrix} 1 \\ -2 \\ 3 \end{bmatrix}$。求$x$、$y$的值，以及与$\mathbf{v}$对应的特征值。2. 假设矩阵 C 与对角矩阵 $B = \begin{bmatrix} 1 & & \\ & -1 & \\ & & 1 \end{bmatrix}$ 相似。求C^2。

A.2 2025年秋季（概念导向）

设 A 为一个 3×3 矩阵，并假设$\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3$ 是 A 的特征向量，分别对应特征值 0、1、2。设$\mathbf{u} = \mathbf{v}_1 + \mathbf{v}_2 + \mathbf{v}_3$。

1. 判断集合 $\{\mathbf{u},\ A\mathbf{u},\ A^2\mathbf{u}\}$ 是线性无关还是线性相关，并说明理由。

2. 求 a、b 和 c 的值，使得 A 与 $B = \begin{bmatrix} 0 & 0 & a \\ 1 & 0 & b \\ 0 & 1 & c \end{bmatrix}$ 相似。

参考文献

Artigue, M. (2002). 在计算机代数系统环境中学习数学：关于工具化的思考的形成，以及技术工作与概念工作之间的辩证关系。《国际数学学习计算机杂志》，*7*(3), 245–274,

Brousseau, G.（2002）。数学教学情境理论：数学教学法 ´，*1970–1990*。荷兰多德雷赫特：施普林格。Frieder, S., Pinchetti, L., Chevalier, A., Griffiths, R.-R., Salvatori, T., Lukasiewicz, T., . . . Berner, J.（2023）。ChatGPT 的数学能力。第三十七届神经信息处理系统大会数据集与基准赛道。Fuchs, K.（2023）。探究自然语言处理模型在高等教育中的机遇与挑战：ChatGPT 是福是祸？《教育前沿》（第8卷，第1166682页）。Leaton Gray, S., Edsall, D., Parapadakis, D.（2025）。大学中的人工智能驱动的数字作弊，以及采用新型伦理教学法的必要性。《学术伦理杂志》，1–18。Li, H., & Wang, W.（2023）。生成式人工智能时代学生作业的设计与评价（中文）。《开放教育研究》，*29*（3），，Prat, A., & Code, W.J.（2021）。WebWork日志文件：学生作业行为的丰富数据来源。《国际科学与技术数学教育杂志》，*52*（10），1540–1556。Risko, E.F., & Gilbert, S.J.（2016）。认知卸载。《认知科学趋势》，*20*（9），676–688。Roth, V., Ivanchenko, V., Record, N.（2008）。评价学生对WebWork这一基于网络的作业发布与评分系统的反应。《计算机与教育》，*50*（4），1462–1482,

Skemp, R.R.（1976）。关系性理解与工具性理解。数学教学，*77*（1），20–26。Sullivan, M., Kelly, A., McLaughlan, P.（2023）。高等教育中的ChatGPT：学术诚信与学生学习方面的考量。应用学习与教学期刊，*6*（1），31–40。Toth, P.F.（2013）。衡量在线数学教学的效率：使用WebWork的经验。*Procedia*-社会与行为科学，*89*，276–282。Verillon, P., & Rabardel, P.（1995）。认知与人工制品：对工具中介活动相关思维研究的贡献。欧洲教育心理学杂志，77–101。Villani, C.（2008）。最优传输：新旧理论（第338卷）。柏林，海德堡：施普林格。Vygotsky, L.S.（1978）。社会中的心智：高级心理过程的发展（第86卷）。马萨诸塞州剑桥：哈佛大学出版社。Wood, D., Bruner, J.S., Ross, G.（1976）。辅导在问题解决中的作用。《儿童心理学与精神病学杂志》，*17*（2），89–100。Xie, Y., Wu, S., Chakravarty, S.（2023）。人工智能遇上人工智能：人工智能与学术诚信——关于计算机教育中减轻人工智能辅助作弊的调查。信息技术教育第二十四届年度会议论文集（第79–83页）。