arXiv:2603.00895v1 [cs.LG] 1 Mar 2026

真实世界手写大学数学作业中的人工智能评分评估：迈向基准的大规模研究

Zhiqi Yu
加州大学欧文分校 欧文，美国
zhiqiy10@uci.edu

Xingpping Liu独立研究者 上海，中国
x
pliu203@gmail.com

Haobin Mao加州大学伯克利分校 伯克利，美国
haobinmao@berkeley.edu

Mingshuo Liu加州大学欧文分校 欧文，美国
mingsh15@uci.edu

Long Chen加州大学欧文分校，欧文，美国
lchen7@uci.edu

Jack Xin
加州大学欧文分校，欧文，美国
jack.xin@uci.edu

Yifeng Yu*加州大学欧文分校，欧文，美国
yifengy@uci.edu

关键词： AI评分；光学字符识别 +大语言模型；数学推理；评分细则引导的评估；形成性反馈；真实世界基准；人类与人工智能的一致性。

摘要

大型本科 STEM 课程的评分往往因教学工作量繁重而只能提供有限的反馈。我们针对加州大学欧文分校真实的手写单变量微积分作业，开展了一项大规模实证研究，评估人工智能评分。我们采用经光学字符识别结果条件化、并结合结构化评分细则引导提示的大语言模型，为近 800 名学生提交的数千份自由作答测验作业评分并提供形成性反馈。在不存在单一真实标签的情况下，我们将模型表现与助教官方评分、学生调查和独立人工审查进行比较，发现其评分与助教评分高度一致，且在各次测验中，绝大多数人工智能生成的反馈均被评为正确或可接受。

除了微积分，这一场景还凸显了 OCR条件数学推理和部分得分评估中的核心挑战。我们分析了主要失效模式，提出实用的评分标准与提示设计原则，并引入多视角评估协议，以支持可靠的真实课程部署。基于本文开发的数据集和评估框架，我们提出建立手写数学 AI评分标准化基准的构想，以支持可复现比较和未来研究。

*通讯作者。

1 引言与研究动机

大规模评分仍是本科教育中最持久的挑战之一。在微积分等选课人数众多的 STEM课程中，助教（TA）必须在紧迫的截止时间内处理数千份开放式作答提交内容。由此产生的工作负担往往会导致两个系统性问题：（1）评分质量和一致性下降；（2）个性化反馈寥寥无几，甚至完全缺失——尽管有意义且内容充实的反馈对学习至关重要。这些问题直接损害基础数学教学中的公平性、透明度和学生参与度。微积分是全美规模最大、资源投入最密集的 STEM教育入门课程之一，因此，可扩展的评分方案有望惠及每年数千个班级和数万名学生。

为此，我们开发并部署了一套大规模人工智能评分系统，将光学字符识别（OCR）与大语言模型（LLM）相结合，用于评估手写解答并生成与评分标准对齐的反馈。人工智能生成的分数仅供参考；助教评分仍是正式课程成绩。据我们所知，这是首批针对大学数学课程中基于大语言模型的评分开展的大规模真实课程研究之一。

我们在加州大学欧文分校一门大规模单变量微积分课程序列的三个学期中部署了该系统，涉及20次测验和1000多名学生。该课程序列由两门连续课程组成：Math 2A和Math 2B。

1. **2025年冬季**。 我们在Math 2B的两次测验中测试了该系统，每次测验包含三道自由作答题，班级学生人数超过500人。

2. **2025年春季**。 我们将该系统应用于Math 2A的六次测验（每次两道题，一道自由作答题，另一道为选择题或填空题，约400名学生），以及Math 2B的六次测验（每次包含两道或三道自由作答题，约360名学生）。2025年春季的数据集是改进系统的主要依据；该学期的五次测验用于与助教评分、学生调查和独立人工审查进行详细比较。

3. **2025年秋季**。 我们还在 Math 2B 的另外六次测验中进一步验证了该系统，每次测验包含两到三道自由作答题，涉及约 200 名学生。

本文所有定量评估均仅针对 2025 年春季的自由作答题，因为选择题和填空题无法体现基于人工智能的手写内容识别、依据评分标准的部分分数评定以及生成清晰且有建设性反馈的核心挑战。

我们通过助教评分的一致性、学生调查和独立人工审查（评审员人数超过 20 人）来评估人工智能辅助评分。人工智能给出的分数与人工评分高度一致；独立评审员普遍认为生成的反馈正确或可以接受，剩余错误主要集中在光学字符识别失败以及少数评分标准的边缘情况。这些结果表明，该系统能够在为学生提供更详细反馈的同时减少助教工作量；完整的定量结果请见第 6 节。

从技术角度来看，使用大语言模型为手写数学评分时，一个关键方面是提示词和评分标准设计。由于许多部分正确的解答并没有唯一的真实分数，系统必须将自由形式的学生手写作业转化为一致且细致的判断，同时确保这些判断符合课程评分标准。因此，除了应对嘈杂的光学字符识别结果和多样的笔迹之外，我们还开发并迭代优化了一个结构化提示词框架，将系统级指令、针对具体题目的参考解答以及机器可读的

评分细则纳入其中。第4节将介绍这一框架；它对于稳定分数并为数千份各不相同的提交内容生成有用的反馈至关重要。

近期关于人工智能辅助评分手写STEM解答的研究 [1, 2, 3, 4, 5, 6, 7],等，随着大语言模型的进展而迅速增长，但通常更重视评分可靠性，而非对反馈质量进行课程规模评估。我们研究了在数千份真实手写微积分作业上的课程规模部署，并利用助教评分、学生调查和独立人工审查（20+ 名评审员）对分数和反馈进行评估。

主要贡献

我们关注一个尚未得到充分探索的机器学习问题：在没有真实标注、输入含噪、采用部分得分评分且受到高风险约束的情况下，如何可靠地评估和部署基于大语言模型的推理系统。

- **大规模真实场景部署**。 我们报告了一个端到端的光学字符识别+大语言模型评分系统（分数和反馈）的大规模部署，该系统应用于一门高选课人数、计学分的大学课程中的自由格式手写微积分解答，并使用了数千份真实测验作业。

- **提示词与评分细则评分流水线**。 我们开发了一套实用的处理流程，将基于光学字符识别的推理与精心设计的系统提示词和多评分细则策略相结合（通常包括一个灵活评分细则和一个固定评分标准）。我们通过失败案例分析反复改进提示词和评分细则，以提升部分得分评分和形成性反馈的质量。

- **多视角人在回路评估**。 我们提出了一套结构化评估方案，结合助教评分一致性、学生反馈以及二十多位外部评审人员的独立评审，用于评估评分可靠性以及人工智能生成的反馈的准确性和实用性；在此场景中，不存在真实标签。

- **未来基准的基础**。 我们已汇集大量匿名化手写解答，并收集了相应的OCR输出、调优后的评分细则、人工智能生成的分数与反馈，以及独立人工审阅结果。作为未来工作，我们计划分两个轨道发布基准子集。*A* 轨（干净）将扫描作业与校正后的转录文本及经教师/助教批准的参考评分配对，从而支持对OCR与基于评分细则的评分进行受控评估。*B* 轨（含噪）将保留原始光学字符识别输出和审阅者标签，以评估鲁棒性和分流能力——也就是说，在输入含噪时，系统何时应转交人工审阅，以及它能否可靠地标记出可能由OCR导致的错误。

尽管本研究聚焦于单变量微积分，但所探讨的挑战——可扩展性、公平性和反馈质量——同样存在于许多大型STEM课程中，因此基础数学自然适合作为更广泛的人工智能辅助评估的试验平台。

2 人工智能评分处理流程概述

我们的处理流程由三个模块化阶段组成（图 1）：区域分割、OCR 到 LaTeX 的转录，以及基于评分细则的大语言模型评分与反馈。我们采用标准化的答题纸格式，将每道题分为解答区域和最终答案区域，使处理流程能够通过独立处理各个区域来应对题目数量的变化。详见附录 A.2 节。我们在区域级别而非整页输入上进行 OCR 和评分，以减少跨题干扰和幻觉；自由格式的解答区域仍然

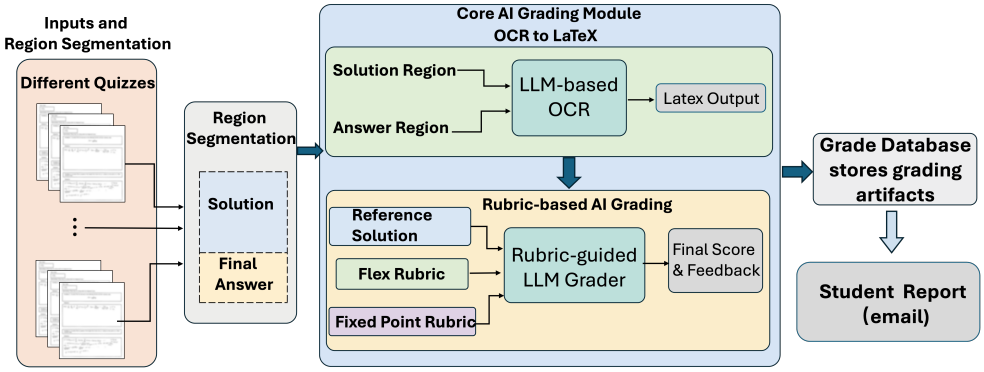


图1：所提出的人工智能评分流水线概览：核心模块将手写解答转换为 LaTeX，并依据参考解答和结构化评分细则对其进行评估，生成详细的分数和反馈。

是最具挑战性的部分。最终的分数和反馈会通过电子邮件发送给学生。见图 1。

1. 光学字符识别与转录文本。通过自动光学字符识别将学生手写作业数字化，并转换为 LaTeX，以保留数学结构。在我们评估的工具中，ChatGPT 4.1 mini 配合设计良好的提示词，在准确性、可扩展性和成本之间实现了较为理想的权衡；详情请参见第3节。

2. 评分细则引导的提示。我们采用两级提示词设计：系统消息提供全局评分原则和输出约束，问题专属评分细则则编码为结构化提示词，规定评分和反馈标准、输出格式以及示例解答。我们在样例提交内容上迭代优化提示模板，使大语言模型输出更好地符合评分细则的意图（第4节）。

3. 大语言模型评估。转录文本和评分细则提示词由大语言模型处理，主要使用 GPT-4.1-mini；如有需要，还会使用 o3-mini 进行补充评估，从而生成数值评分和解释性反馈。由于评分缺乏绝对的真实标注，我们通过三项互补验证来评估其可靠性：与助教评分的一致性、学生反馈以及独立人工审查（第6节）。

模型选择与能力演进。除非另有说明，我们在光学字符识别和评分环节均使用 GPT-4.1-mini，在准确性、延迟和成本之间取得平衡，以满足课程规模部署的需求。此外，对于少量代数内容较多或其他较难的案例，我们使用 o3-mini，因为这些案例需要更强的推理能力。比较结果请参见第 5 节。我们不会盲目追逐每一次新模型发布，而是保持处理流程稳定，并允许随着模型能力提升而替换后端模型；系统性的跨模型比较和路由策略留待未来工作。

3 光学字符识别组件

对于手写数学作业的转录，我们在2024年秋季/2025年冬季使用课程数据评估了多种光学字符识别选项，并选定 Mathpix 作为当时最可靠的选择。2025年4月 GPT-4.1/4.1-mini 发布后，我们重新进行了光学字符识别评估，发现采用针对光学字符识别的提示时，GPT-4.1-mini 在我们的数据上生成的转录文本质量更高。在数千份手写提交作业中，约 88% 的转录文本是可接受的，也就是说，它们要么 (i) 完全正确 （字符级 LaTeX 忠实度），要么 (ii) 足以用于基于评分细则的

评分。与主要面向逐字转录的 Mathpix 不同，GPT-4.1-mini 可以利用上下文推理来消除歧义，从而提高对潦草且形式多样的学生手写体的鲁棒性。在一个固定的高难度子集（即**171 份解答子集**）上进行直接比较时，该子集由 59 份测验组成，每份测验包含相同的三个问题（共 171 份解答），GPT-4.1-mini 的可接受转录文本比例达到 84%，而 Mathpix 为 55%。代表性示例见附录 A.3 节。

GPT-4.1-mini 的光学字符识别性能对提示词设计较为敏感；我们采用具备上下文感知能力的约束来提高转录文本的稳定性。下面是我们的基础光学字符识别提示词示例。

OCR 提示词

本题要求学生使用左黎曼和估算定积分：

$$\int_1^5 (x^2 - x - 1)\,dx$$

使用**四个子区间**和**左端点**，无需计算精确的极限。

1. 使用光学字符识别识别图像中的所有文本，并将其转换为 LaTeX 代码 2. 不要更正任何数学或逻辑错误，也不要更正拼写错误。

为了更好地利用 GPT-4.1-mini 光学字符识别的上下文推理能力，在提示词中陈述问题非常重要。另外两种针对性的提示词策略，旨在解决两个常见且可通过提示词处理的转录文本不稳定来源。

- **幻觉**。当答案框空白或只写了少量内容时，GPT-4.1-mini 可能会生成完整解答。上述提示词 1，以及预先在答案框内印上“解答”和“最终答案”这两个词，都能稳定输出并抑制幻觉（见附录 A.2 节）。

- **自动纠错**。GPT-4.1-mini 偶尔会在转录时纠正学生的数学错误。明确指示模型不要纠错（提示词 2）有助于减少这种行为，但仍有少数情况（在我们的样本中不足 2%）。附录 A.5 节提供了一个典型示例。

仍有两项主要的光学字符识别挑战：(1) 几何图示，GPT-4.1-mini 在这方面的能力有限，而且在我们的实验中，Mathpix 无法从学生作业中重建任何图示；(2) 可靠地处理划掉的文字或大量潦草涂写的文本。后一种情况下，如果划掉的区域在视觉上与相邻墨迹融为一体，光学字符识别系统可能会转录已擦除的内容，或漏掉附近有效的符号。我们尝试了几种提示词改进方案来抑制已擦除内容的识别，但这通常也会删除附近的有效内容。

除这些可通过提示词解决的问题外，我们还分析了**171 份解答子集中**全部 27 个有问题的光学字符识别案例。GPT-4.1-mini 的失败大多与布局有关，其中最主要的失败模式涉及复杂除法表达式。具体而言，当嵌套分数书写时留白不足，主分数线的范围就会变得模糊，导致项被合并、误解析或遗漏。其他类型的错误并不常见。有关这种与分数相关的失败模式示例，请参见附录 A.4 节。

4 评分与结构化提示工程

在光学字符识别的可靠性已基本得到验证后，AI评分面临的一项核心挑战在于提示词设计，具体而言，是构建能够确保大规模一致性、公平性和评分标准

的结构化提示词。我们的实验表明，精心设计系统消息工程和评分细则制定，是实现人类水平的评分准确率的决定性因素。

4.1 系统消息设计

我们评分系统中的提示词由两部分组成：定义评分行为的系统消息，以及规定题目级标准的评分细则部分。我们发现，基于 API 的评分可能不如在交互式网页界面中进行的评分全面。虽然外部并未公开确切原因，但这一发现促使我们在系统消息中明确列出所有评分指示，以便基于 API 的部署能够生成一致且高质量的输出。

为减少随机性，我们将 GPT-4.1-mini 的温度设为接近零的值（0.0–0.1），以提升响应稳定性和可复现性。对于 o3-mini，温度并非可公开设置的 API 参数，因此无法调整。我们通过反复迭代完善系统消息，着重强调逻辑一致性、事实正确性和类人反馈。摘录如下：

系统消息摘录

你是一位专业的数学评分员。 评估时，请：

1. 区分数学上正确的陈述与错误的推理。
2. 避免反馈中出现矛盾。
3. 如果错误尝试后来得到了纠正，不要因此扣分。
4. 针对扣分给出简洁且有建设性的解释。
5. 使用清晰、自然且适合学生阅读的语言撰写反馈。

这些结构化指令有助于减少逻辑不一致，并避免对无害的早期错误进行过度扣分——这些都是简单提示下常见的失效模式。附录 A.6 节提供了有代表性的示例。

4.2 评分标准与提示框架

我们的评分细则框架基于数千份 Math 2A 和 B 的提交内容制定，旨在兼顾可解释性与评分精确度。每道题都可以关联多份评分细则，以适应不同的有效解题策略。对于大多数问题，我们采用由以下部分组成的双评分细则设计：

1. **简洁的参考解答**，概述一种或多种典型正确解法。
2. **两份互补的评分细则**：
   - **灵活评分标准** ——模拟人工评判，并能包容多样的推理方式；支持自由形式评估，为部分正确的作答以及少见但有效的解法给分。
   - **固定评分标准** ——细致且按分值计分，包含针对具体问题的检查项，用于核查标准解答中要求的步骤和常见错误。

   关于两种评分标准及其 JSON 输出的对比例子，请参见附录 A.7。该例说明，对于非典型但正确的解答，灵活评分标准如何补充清单式评分细则。重要的是，灵活性并不会系统性地抬高分数：

在我们的实验中，如果局部正确的步骤在整体上不一致或数值错误，灵活评分标准给出的分数可能低于固定评分标准。

3. **选择规则**。为减少模型混淆，每份提交内容都会通过独立的 API 调用，分别依据两种评分标准进行评估；我们采用最大值规则：选取两种评分标准中较高的分数作为最终成绩，并采用与之对应的反馈。我们在 Math 2A 和 B 测验 5–6 上评估这一设计。这些测验包含难度更高的问题，且允许更广泛的有效解题路径和部分作答。

为评估评分结果与人工评分行为的一致性，我们将仅使用固定评分标准、仅使用灵活评分标准，以及使用最大值规则所得的成绩与助教评分进行比较（助教评分仅作参考，并非绝对的真实标注）。以平均绝对误差（MAE）为主要指标，最大值规则在两门课程和各项测验中均持续取得最低的 MAE（图2）。这些结果表明，最大值规则能够为有效推理赋分，而这些推理可能在单一评分标准下未得到充分认可，因此可产生更稳定、更符合人工评分的分数。

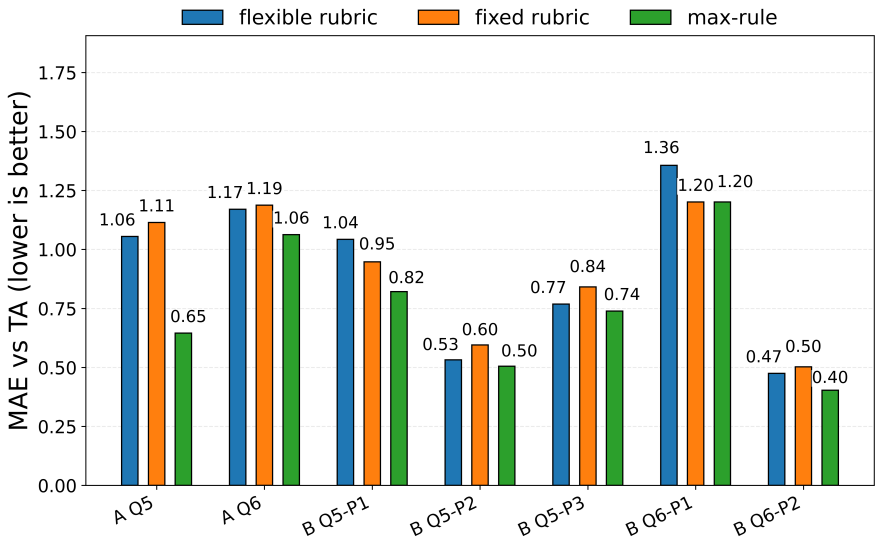


图2：每组中的柱形从左到右分别对应灵活评分标准、固定评分标准和最大值规则：A-Math 2A、B-Math 2B

范例引导的评分细则生成。我们利用示例帮助模型以一致的格式编写新的评分细则。我们主要使用2025年春季 Math 2A 和 B 的测验，按题型（例如定积分）建立了一个小型代表性题库，并为每种题型完善了一套高质量的评分细则。对于后续学期（例如2025年秋季）中同一类别或密切相关类别的新题目，我们会向大语言模型提供题目描述、参考解答和最接近的现有评分细则，并要求它生成一份详细程度相当的评分细则。生成的评分细则通常只需经过少量人工修改（例如补充新发现的误解或课程特定约定）即可投入使用。未来的工作将把这些评分细则整理到轻量级数据库中，并进一步自动化评分细则编写流程。

两种可能导致不合理扣分的情况。我们观察到两种反复出现的情况：即使数学内容本身正确，LLM评分器仍可能扣分。第一，模型可能会惩罚条理不清或不规范的书写形式，尽管它正确识别出了有效的解答。第二，我们的答题纸将作答区与最终答案框分开：上下文推理有助于提高解答区域的光学字符识别准确性，但最终答案框周围几乎没有上下文，因此更容易出现OCR错误。

因此，即使解答转录文本准确且数学内容有效，最终答案仍可能被误读（导致正确的作答看起来像是错误的）。

为缓解这些问题，我们计划在未来的部署中加入人工评审后备机制：当正确答案得到明确支持但未被判定为满分时，启用人工评审。在所报告的实验中，我们已经纳入了评分标准指导，避免基于风格进行扣分；当解答步骤明确支持正确数值时，也会容忍最终答案框中的轻微光学字符识别误差。我们实验中使用的代表性评分标准指导见附录 A.8 节。

5 o3-mini 与 GPT-4.1-mini 的评分方差和针对性准确性评估

我们使用代数密集型的 **171 份解答集**，比较 o3-mini 和 GPT-4.1-mini 在评分准确性和分数方差方面的表现。为隔离评分者的影响，我们将光学字符识别固定为 4.1-mini，仅改变评分模型（GPT-4.1-mini 与 o3-mini）。由于 o3-mini 强调多步推理，我们预计它在涉及复杂代数运算的题目上表现更好；两名独立评审员发现，两者的反馈总体上同样准确，但在出现差异时更倾向于 o3-mini。然而，o3-mini 的方差和成本更高，因此我们仅将其用于最具挑战性的题目，并采用简单的多次运行稳定化来提高可靠性。详情请参见附录 A.10。

降低方差。 对于每个题目，我们运行 o3-mini 三次，取平均分，并输出最接近该均值的一次运行结果（及其反馈）。受大语言模型推理中自洽性可作为稳定性信号的启发 [8], ，我们不采用多数投票——这种方法通常需要大量独立运行——而是采用这一最接近均值的规则进行连续部分给分评分。在多数情况下，两位独立人工评审员认为所选运行结果的反馈准确性更高。对于结果波动的情况，参见附录 A.10.3 节中与多数投票的比较。

人工审阅。 如果三次运行中恰有一次给出满分，而另外两次没有，我们就会将该题目标记为需要人工审阅。

6 与人工评分和反馈的评估

我们评估了人工智能辅助评分在两门单变量微积分课程（A 和 B）中的表现，涵盖近 800 名学生，并考察分数的一致性和反馈质量。一个核心挑战是缺乏真实基准：目前没有公开可用的微积分解答专家评分数据集，而且人工评分者之间也存在不可忽视的意见分歧，尤其是在评分边界或非常规答案上（附录 A.10.4）。这种差异使得难以界定一个普遍认可的“正确”分数，也使反馈评估不可避免地带有主观性。

鉴于这种固有的模糊性，我们采用一种务实的、多视角的评估框架，由三个互补的部分组成：

- **评估与助教评分的一致性**。我们将人工智能给出的分数与助教（TA）给出的分数进行比较。助教评分可能不一致，而且往往缺乏详细的个性化反馈，但它是大学评估中的实际参照，因此，与助教评分实践保持一致是部署的必要条件。由于助教 Gradescope 评分细则旨在提高人工评分效率，我们的人工智能评分标准经过调整，以便进行自动解读并提供反馈；因此，我们将分数比较视为对结果一致性的衡量，而非对评分标准严格等同的衡量。

- **直接收集学生反馈**。我们将人工智能生成的分数和反馈返还给学生（示例消息见附录 A.11），并收集 Math 2A 学生关于感知正确性和实用性的问卷回复。这有助于发现潜在的评分错误，并了解学生如何解读人工智能反馈，同时也考虑到，当人工智能给出的分数低于助教给出的分数时，学生可能会觉得人工智能过于严苛。

- **外部评估人员进行独立评审**。一组经过培训的研究生、本科生和高年级高中生使用与人工智能相同的评分细则，独立评估人工智能输出，评判其正确性和公平性，并给出自己的分数。这在与助教评分的一致性和学生反馈之外，提供了一个额外且部分独立的第三方视角。

由于实际、技术和行政方面的限制（例如分割失败、扫描质量问题、试卷代码不匹配以及评审人员的可用性），并非所有测验都能纳入每项比较；这些排除项在评估前确定，且与评分结果无关（即非选择性排除）。分类统计见附录 A.1 节。

此外，我们评估所有自由作答提交内容的一致性时，不按解答正确与否进行分层，因为即使底层数学内容正确，光学字符识别的可靠性和基于评分标准的推理也必须发挥作用；因此，无论答案正确与否，评分一致性都不是显而易见的。

6.1 与助教评分的一致性

我们评估人工智能评分器与人工助教对 Math 2A 和 B 的 3,945 条手写自由作答学生题目记录的评分一致性。Math 2B 的所有题目均为自由作答题；Math 2A 中非自由作答题（例如选择题或填空题）不纳入本分析。每条记录都包含一个人工智能评分、一个助教评分，以及差值 $g = \text{AI} - \text{TA}$。

我们从两个层面进行汇总：

1. 所有记录中分数差的总体分布；

2. 测验层级汇总指标：平均分数差，以及分差在1分以内的百分比（“1分以内”）；

6.1.1 总体一致性（全局）

总体分数差分布在 0 处呈现尖锐峰值，这意味着对于许多提交内容，人工智能和助教给出的分数相同或非常接近（图3）。直方图也有较短的尾部：大多数差异接近 0，但少数提交内容的分数差距较大。平均而言，人工智能评分比助教评分低 0.40 分，标准差为 1.12（约一分）。

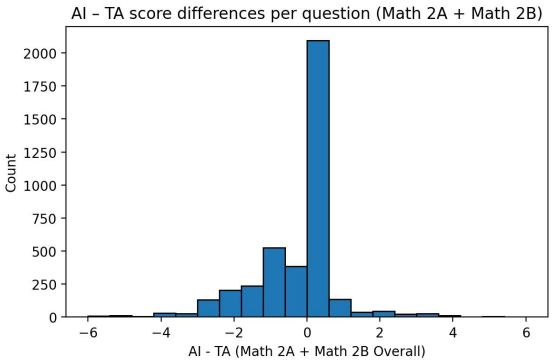


图3：人工智能与助教评分差的总体分布。

6.1.2 测验层级摘要

在各项测验中，平均绝对分数差（MAE）通常介于 0。5 至 1。06 分之间，误差在 1 分以内的比例为 68% 至 86%，表明人工智能评分与助教评分总体上高度一致。

course	quiz	n_questions	total_records	mean_gap	mean_abs_gap	pct_within1
Math 2A	Quiz2	2	540	-0.47	0.72	82%
Math 2A	Quiz3	1	265	+0.23	0.71	82%
Math 2A	Quiz4	2	308	-0.23	0.78	77%
Math 2A	Quiz5	1	255	-0.38	0.65	83%
Math 2A	Quiz6	1	88	-0.05	1.06	68%
Math 2B	quiz2	3	402	-0.48	0.71	81%
Math 2B	quiz3	3	618	-0.39	0.47	86%
Math 2B	quiz4	3	453	-0.71	0.83	72%
Math 2B	quiz5	3	624	-0.41	0.69	78%
Math 2B	quiz6	2	392	-0.59	0.80	77%

图4：测验层级摘要：平均分数差和1分以内的百分比。

对于与 $|g| \geq 1$ 存在较大差异的情况，负分数差（$g \leq -1$）主要归因于光学字符识别错误，或助教评分偶尔较宽松（即人工智能的评分相对更严格）。关于一个具有代表性的正向异常值（ $g > 1$），请参见附录 A.9。

6.2 学生调查结果

以下总结了关于我们的评分系统所提供反馈的自愿学生调查结果。调查在 Math 2A 的两个班级中进行（n $= 301$ 名受访者，约占 400 名选课学生）。Math 2B 未开展调查。两个独立班级在所有项目上的回答模式都高度一致。

课程	准确	清晰	更倾向助教	对人工智能持开放态度
2A 第1节（n=158）	60%	63%	52%	78%
2A 第2节（n=143）	61%	62%	53%	76%

表1：学生调查摘要。“准确” = 反馈非常/大体准确；“清晰” =反馈非常/大体清晰；“对人工智能持开放态度” =是，或在改进后也许可以。

总体而言，该调查表明学生对人工智能辅助评分持谨慎开放态度，许多学生认为其准确性和清晰度尚可。开放式回答最常提及手写识别（光学字符识别）、部分分评分严格度以及解释有限，并强调需要人工监督。这些意见促使我们采用稳健的光学字符识别，并建立自动触发复核机制，将低置信度案例升级为人工核查（第7节）。

6.3 人工智能与独立人工评审员的一致性

虽然人工智能与助教评分一致性已经很高，但仅仅匹配分数并不能保证评分的正确性或反馈质量。因此，我们邀请独立评审员审核了 3,851 条手写自由作答学生—题目记录中的 AI输出（Math 2 A：1,819 条；Math 2B：2,032 条），核验了：(i) 输入质量（光学字符识别结果的可辨读性）；(ii) 在相同评分量规下的分数一致性；以及 (iii) 评分标准忠实度和反馈准确性。由此，我们在原始的与助教评分的一致性之外，提供了一项独立且符合评分标准的评估。鉴于手写自由作答缺乏客观的真实分数，这项评估尤为重要。

输入质量（光学字符识别）。大多数答卷清晰可读：87.64% 被评为可接受，12.36% 被评为有问题（图5）。典型失效模式请参见光学字符识别第3节。

评分一致性。在 3851 份评分后的回答中，分数差（人工智能–评审员）高度集中于 0，且两侧尾部轻微、对称；超过 90% 的差值在 ± 1 分以内（图6），表明数值一致性较强。

评分标准忠实度与反馈准确性。独立评审发现，人工智能评估中有 79.79% 完全正确、9.55% 可以接受，10.67% 不正确（图 7）。除光学字符识别相关错误外，还出现了两种主要错误模式：（i）过度扣罚数学上正确、但评分标准中未明确列出的有效解题方法；（ii）对符合评分标准形式的解释过于宽容，这类解释遵循熟悉的步骤，却包含错误的推理。

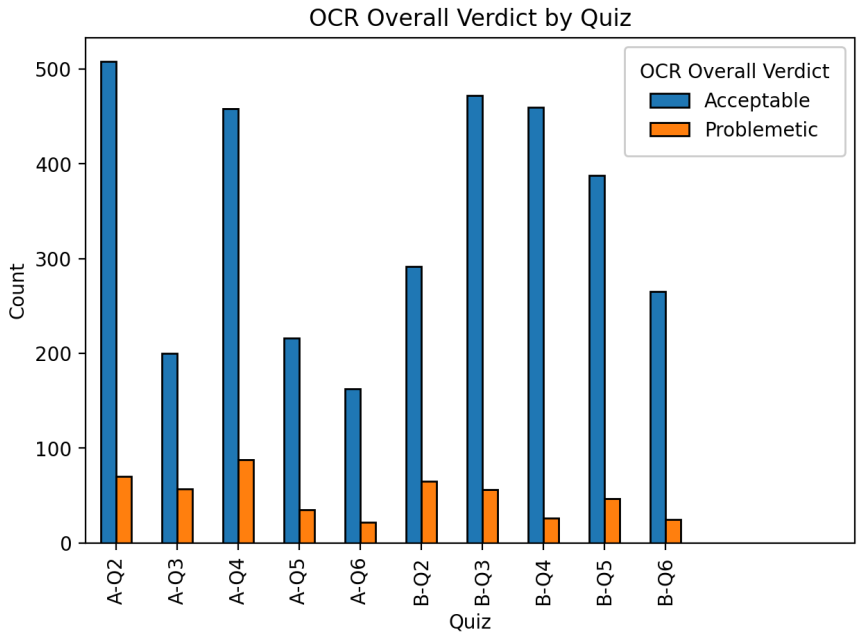


图5：光学字符识别判定分布：A-Math 2A、B-Math 2B

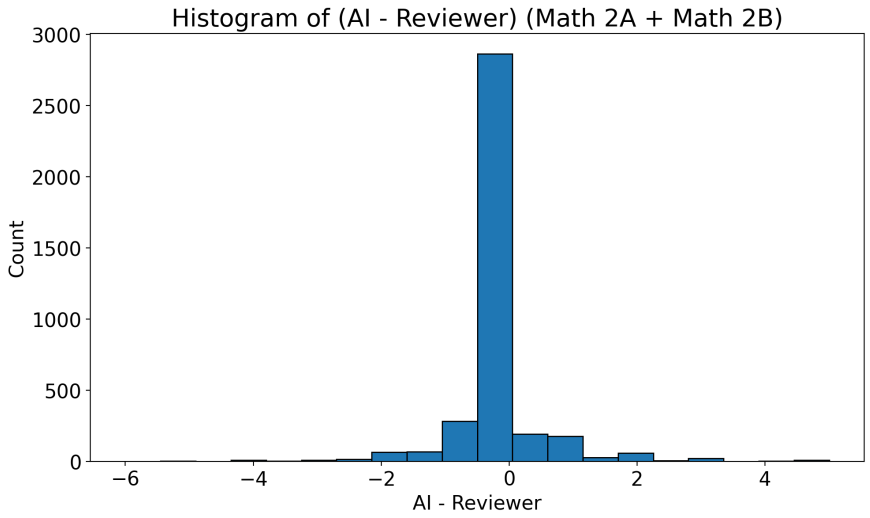


图6：人工智能与评审员的分数差异

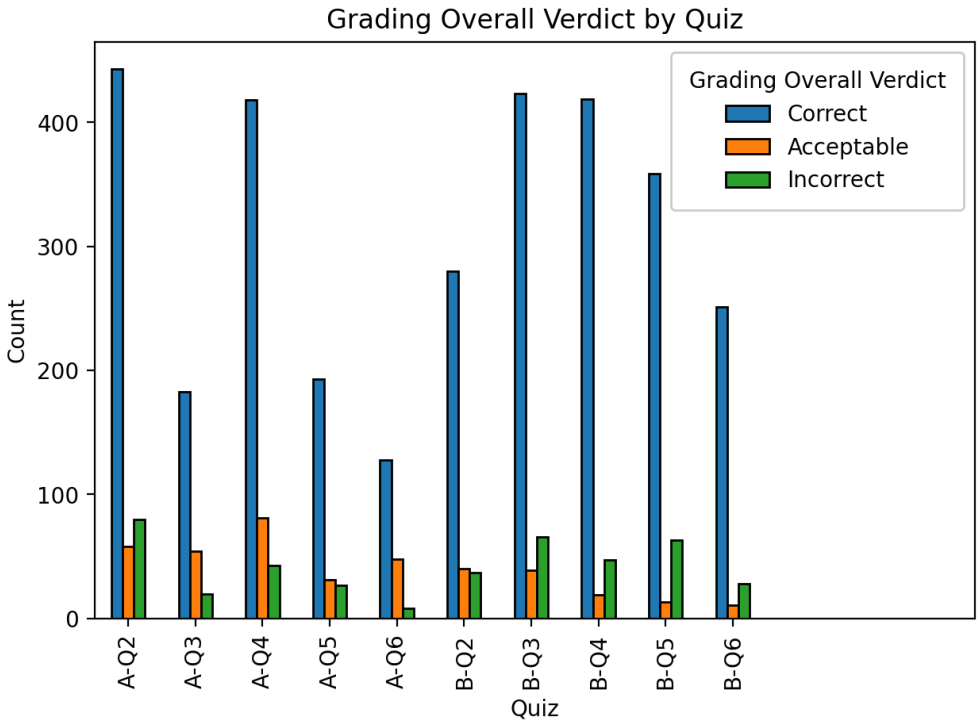


图 7：评分标准忠实度与反馈审计：A-Math 2A、B-Math 2B

7 局限性

我们的研究存在若干局限，也因此明确了后续研究的方向。

干预策略。 大规模部署的一项关键瓶颈，是决定何时信任自动评分，何时转而进行额外检查或人工审阅。尽管光学字符识别错误和评分标准与提示词的交互仍是导致失败的主要原因，但我们尚无系统化方法来决定何时触发此类检查或转交人工审阅。在实验中，我们仅标记少数几类问题（例如，最终答案正确但未获满分，或重复运行间结果方差较高），而这些触发条件只是简单的人工设计启发式规则，并非统一决策规则。关于学习型延迟决策、不确定性感知的风险规避型部署以及公平弃权的相关机器学习研究，为触发人工审阅提供了有原则的思路（例如，[9, 10, 11]），但如何将这些框架用于缺乏真实标注的、基于光学字符识别的部分得分评分，仍有待探索。

高风险部署。 我们的实验聚焦于低风险测验（占课程成绩的 ≈10%）；若将人工智能辅助评分扩展至期中考试和期末考试，错误的代价将会更高，因此需要采取严格得多的人机协同控制措施。

选择最准确的评分/反馈。 当有多个 OCR 引擎或评分者可供选择时（例如，基于GPT的OCR与Mathpix，或不同的大语言模型评分器），我们目前依靠人工评审员解决分歧，并判断哪种输出更准确。这种做法难以扩展，也会拖慢我们评估或引入新模型的速度。开发一种可靠的自动机制来检测分歧，并选择最可靠的分数+反馈，是未来的研究项目。

评分量规聚合（最大值规则）。 我们使用最大值规则来聚合评分量规层面的证据；在我们的实验中，这种方法与助教评分最为一致。不过，我们尚未将其与其他合理的选择（例如取平均或其他单调分数变换）进行全面比较，因此，关于最优聚合规则的结论仅适用于我们的具体情境。

研究范围。 我们的评估聚焦于一所院校的单变量微积分；还需要在不同课程和院校中开展更多部署，以验证其普适性，并了解其在不同评分文化和学生群体中的实际影响。

8 结论

我们对加州大学欧文分校单变量微积分课程中的人工智能辅助评分进行了大规模评估，涵盖近800名学生的数千份测验解答。我们的系统将光学字符识别、提示引导推理和基于评分细则的评分整合到自动化工作流程中，并通过与助教评分、结构化人工审阅及学生反馈进行比较来评估。

我们的结果表明，当前的人工智能系统已经可以作为可靠的评分助手：它们能够准确处理大多数常规题目，提供实用的反馈，并大幅减少人工评分工作量。仍有一些关键挑战有待解决，包括在图示或几何内容较多的作答以及约分过程中出现的光学字符识别失败、如何对多种解答变体进行一致的部分分数分配，以及在低置信度情况下触发人工审阅的合理策略。

展望未来，我们计划从收集的数据库中发布有完善文档说明的基准子集，以支持对光学字符识别和基于评分细则的评分进行可复现的评估。具体而言，我们将提供两个轨道：**A 轨**（干净版），包含校正后的转录文本以及经教师/助教批准的参考分数；**B 轨**（含噪版），保留原始 OCR 输出和审阅者标签，用于评估鲁棒性并为分流提供依据（即确定何时应转交人工审阅）。我们希望这些资源能够促进更加透明、可扩展的比较，并加快大规模招生的 STEM 课程中可靠自动化评估的发展。

致谢

我们谨向加州大学尔湾分校物理科学学院和数学系致以诚挚谢意，感谢他们持续支持这一人工智能评分项目。我们衷心感谢研究生评审员——Jongwon D. Kim、Yuanduan Lou、Shiyi Luo 和 Zeyi Xu；本科生评审员——Keith Bui、Xin Gao、Changze Li、Tony Li、Alex Lu、Eric Lyu、Xuaner Niu、Oleg Presnyakov、Scott Sibley、Vincent Wang、Qian Zhang 和 Mickey Zhao；以及高中生评审员——Eric Guo、Kevin Huang、Aaron Liu、Jasper Sun、An My Ngoc Tran 和 Lauren Zhang，感谢他们审阅人工智能评分结果，并提供建设性的反馈与建议。我们特别感谢 Oleg Presnyakov 提出最大值规则（从多个评分细则中选取最高分）。同时，特别感谢 MATH 2A/2B 任课教师——Rachel Cohen Lehman、Ziming Shi、Daniel V. Simonson 和 Yang Zhang—— 给予的大力支持，也感谢 MATH 2A/2B 助教在整个项目期间提供的宝贵协助。Long Chen 获美国国家科学基金会 DMS-2309777 项目部分资助，Jack Xin 获美国国家科学基金会 DMS-2309520 项目部分资助，Yifeng Yu 获美国国家科学基金会 DMS-2000191 项目部分资助。我们使用了基于大语言模型的语言工具辅助措辞和表达（包括改写和结构编辑）；所有技术内容和结论均由作者全权负责。

参考文献

[1] Nur Aisyah、Muhammad Daffa A. Kautsar、Asep Hidayat、Rifat Chowdhury 和 Fajri Koto。《从手写文字到反馈：评估视觉语言模型和大语言模型在印度尼西亚课堂中用于人工智能驱动评估的表现》，2025年。arXiv:2506.04822.[2] Ryan Mok、Faraaz Akhtar、Louis Clare、Christine Li、Jun Ida、Lewis Ross 和 Mario Campanelli。《使用人工智能大语言模型进行教育评分：物理学实践测试》，2024年。arXiv:2411.13685.[3] Adriana Caraeni、Alexander Scarlatos 和 Andrew Lan。《评估 GPT-4 对数学考试手写解答的评分能力》。发表于第*15*届国际学习分析与知识会议论文集，2024年。亦可见 arXiv:2411.05231。[4] Alberto Gandolfi。《教育中的 GPT-4：评估其在解答微积分问题和评分作业方面的适用性、可靠性及连贯性损失》。《教育人工智能国际期刊》，35:367–397，2025年。[5] Unggi Lee、Youngin Kim、Sangyun Lee、Jaehyeon Park、Jin Mun、Eunseo Lee、Hyeoncheol Kim、Cheolil Lim 和 Yun Joo Yoo。《我们能否在教育中使用 GPT-4 作为数学评估器？探索基于大语言模型的自动评估系统对开放式数学问题的效能与局限》。《教育人工智能国际期刊》，35:1560–1596，2025年。致编辑信。[6] Tianyi Liu、Julia Chatain、Laura Kobel-Keller、Gerd Kortemeyer、Thomas Willwacher 和 Mrinmaya Sachan。《人工智能辅助的自动简答题评分：大学数学手写考试》，2024年。arXiv:2408.11728。[7] Jean-Claude Nkoyo 等。《大语言模型自动评分的进展：跨学科综述》。发表于第*20*届教育应用构建自然语言处理创新应用研讨会（*BEA*）论文集，第477–498页，2025年。[8] Xuezhi Wang、Jason Wei、Dale Schuurmans、Quoc V. Le、Ed H. Chi、Sharan Narang、Aakanksha Chowdhery 和 Denny Zhou。《自洽性提升语言模型中的思维链推理》。发表于国际学习表征会议论文集，2022年。[9] David Madras、Toniann Pitassi 和 Richard Zemel。《负责任地进行预测：通过学习延迟决策提升公平性和准确性》。发表于神经信息处理系统进展，2018年。[10] Shayan Kiyani、George Pappas、Aaron Roth 和 Hamed Hassani。《保形预测的决策论基础：面向风险厌恶型主体的最优不确定性量化》。发表于国际机器学习会议论文集，2025年。[11] S´ılvia Casacuberta 和 Varun Kanade。《选择性全预测与公平弃权》。发表于国际机器学习会议论文集，2025年。

A 附录

A.1 不同验证协议下的数据集覆盖范围

由于实际技术和行政方面的限制，并非所有提交的测验都能纳入每项验证协议。我们区分 (i)适用于所有后续分析的普遍 排除项，以及 (ii) 特定协议额外施加的限制（例如，人工智能与助教评分比较或独立人工审查）。排除是由处理或后勤可行性决定的，而非评分结果；其目的在于避免对评分分歧产生选择性影响。

普遍排除项（适用于所有协议）。首先，我们的处理流程未能成功处理部分测验（最常见的原因是学生在指定方框外书写导致分割失败、扫描质量问题，或偶发的光学字符识别/输入提取失败）。此外，在模型迭代以及与课程工作人员协调的过程中，Math 2A 和 Math 2B 的第一次测验均未纳入评估集。最后，Math 2B 的一个助教班级出现了系统性的扫描伪影（例如扫描页面上出现多余线条），严重降低了分割和输入提取的质量；因此，我们将该班级从所有分析中排除。

人工智能与助教评分比较的其他限制（记录关联）。人工智能与助教的比较还需要通过指定的去标识化测试代码，将每份已处理的提交内容与 Gradescope 上的记录关联起来。由于部分学生漏填或错填了代码，这种关联有时会失败；此外，在某些情况下，由于我们无法控制的原因，Gradescope 导出的助教评分数据不完整或缺失，因此无法可靠地进行逐题评分比较。

独立人工审查的其他限制（审查安排）。独立评审按助教班级组织：每位评审员负责审阅来自一个助教班级的测验，但并非所有助教班级都能纳入审查。例如，当一位助教负责多个讨论课班级时，由于审查人员容量有限，我们有时只审阅其中一部分班级。少数已分配的审查任务未能完成，因为部分评审员没有完成审查。

A.2 答题纸设计

下面是两种答题纸设计。版本 1 允许答案留空，尽管有明确的光学字符识别指令，仍可能引发幻觉。版本 2 则通过在指定框内放置**“解答”**和**“最终答案”**来避免这一问题。

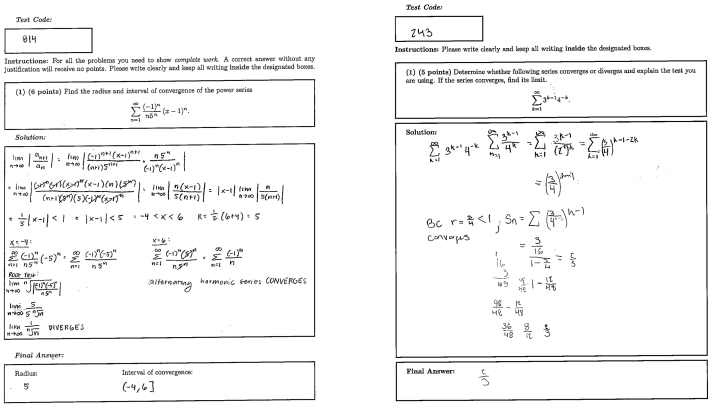


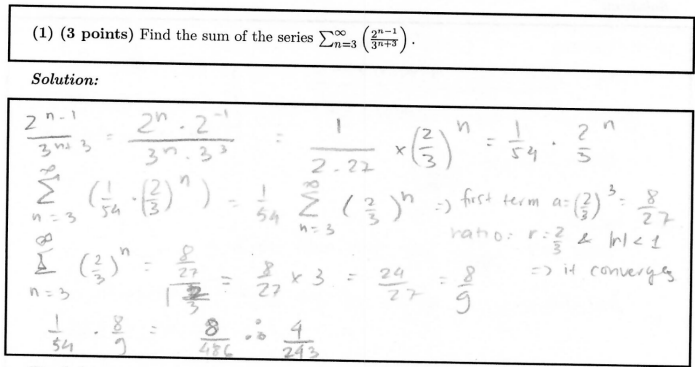
图 8：左：答题纸版本 1；右：答题纸版本 2。

A.3 GPT-4.1-mini 与 Mathpix

接下来，我们将展示几个有代表性的示例，说明：(1) ChatGPT 4.1-mini 表现出色；(2) Mathpix 与 ChatGPT 4.1 mini 的输出对比凸显了关键差异；(3) 如何有效缓解自动纠错和幻觉等问题。

示例 1. 阅读下面这份来自 Math 2B 测验的手写解答。题目要求求出该级数的和：

$$\sum_{n=3}^{\infty} \frac{2^{n-1}}{3^{n+1}}.$$



GPT-4.1-mini 的光学字符识别输出经由我们的 API 流水线以 LaTeX 格式返回，并在此处渲染为编译后的 PDF。转录文本几乎完美，下游评估也给出了正确分数。

$$\frac{2^{n-1}}{3^{n+3}} = \frac{2^n \cdot 2^{-1}}{3^n \cdot 3^3} = \frac{1}{2 \cdot 27} \times \left(\frac{2}{3}\right)^n = \frac{1}{54} \cdot \left(\frac{2}{3}\right)^n$$

$$\sum_{n=3}^{\infty} \left(\frac{1}{54} \cdot \left(\frac{2}{3}\right)^n\right) = \frac{1}{54} \sum_{n=3}^{\infty} \left(\frac{2}{3}\right)^n \quad \Rightarrow \textit{first term } a = \left(\frac{2}{3}\right)^3 = \frac{8}{27}$$

公比：$r = \frac{2}{3}$ & $|r| < 1$

$$\sum_{n=3}^{\infty} \left(\frac{2}{3}\right)^n = \frac{\frac{8}{27}}{1 - \frac{2}{3}} = \frac{8}{27} \times 3 = \frac{24}{27} = \frac{8}{9} \quad \Rightarrow \textit{it converges}$$

$$\frac{1}{54} \cdot \frac{8}{9} = \frac{8}{486} = \frac{4}{243}$$

此示例展示了 4.1-mini 基于推理的阅读能力。根据上下文，模型正确推断出了学生想要表达的式子：(1) 补上缺失的括号，修复了轻微的符号遗漏，并将该项理解为 $\left(\frac{2}{3}\right)^n$；(2) 对分母中的有歧义的记号作出正确判断（减号很淡，也可能缺失），将其理解为 $\frac{\frac{8}{27}}{1-\frac{2}{3}}$，而不是带分数 $1\frac{23}{27}$。因此，我们将该输出归类为几乎完美，因为它正确还原了原意，而非逐字转录。在这种情况下，严格按字面解读可能会导致评分时出现不合理的失分。

来自 Mathpix 的光学字符识别输出，通过我们的 API 流水线以 LaTeX 格式返回，并在此处渲染为编译后的 PDF： 转录文本中包含几个拼写错误，导致在下游评估环节无谓地失了一分。

$\frac{2^{n-1}}{3^{n+3}} = \frac{2^n \cdot 2^{-1}}{3^n \cdot 3^3} = \frac{1}{2 \cdot 27} \times \left(\frac{2}{3}\right)^n = \frac{1}{54} \cdot \frac{2}{3} n$

$\sum_{n=3}^{\infty} \left(\frac{1}{54} \cdot \left(\frac{2}{3}\right)^n\right) = \frac{1}{54} \sum_{n=3}^{\infty} \left(\frac{2}{3}\right)^n \Rightarrow \ firsterm\ a = \left(\frac{2}{3}\right)^3 = \frac{8}{27}$

$\frac{k}{\infty} \left(\frac{2}{3}\right)^n = \frac{8}{27} = \frac{8}{27} \times 3 = \frac{24}{27} = \frac{8}{9} \quad \Rightarrow \ it\ converges$

$n = 3 \quad \frac{8}{34} \cdot \frac{8}{9} = \frac{8}{486} \cdot \frac{4}{243}$

$\frac{1}{54}$

出于好奇，我们也使用在线版本 Mathpix Snip 处理了上述样例。以下是其中一个输出（以 PDF 格式渲染）。

$$\frac{2^{n-1}}{3^{n+3}} = \frac{2^n \cdot 2^{-1}}{3^n \cdot 3^3} = \frac{1}{2 \cdot 27} \times \left(\frac{2}{3}\right)^n = \frac{1}{54} \cdot \frac{2}{3}$$

$$\sum_{n=3}^{\infty} \left(\frac{1}{54} \cdot \left(\frac{2}{3}\right)^n\right) = \frac{1}{54} \sum_{n=3}^{\infty} \left(\frac{2}{3}\right)^n \Rightarrow \ first\ term\ a = \left(\frac{2}{3}\right)^3 = \frac{8}{27}$$

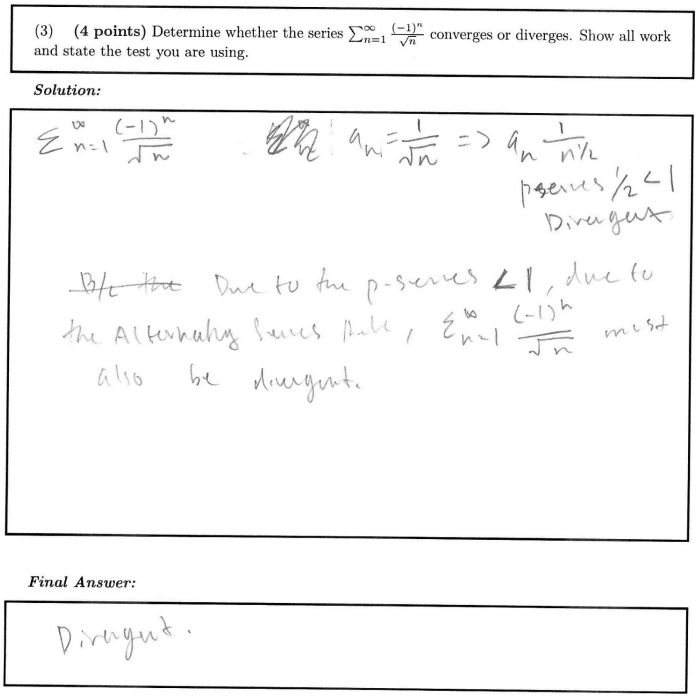
$$\frac{k}{n} \left(\frac{2}{3}\right)^n = \frac{\frac{8}{27}}{1\frac{2}{3}} = \frac{8}{27} \times 3 = \frac{24}{27} = \frac{8}{9} \quad \Rightarrow \ it\ converges$$

$$\frac{1}{54} \cdot \frac{8}{9} = \frac{8}{486} \therefore \frac{4}{243}$$

示例 2。下面是另一个示例，展示了 4.1-mini 在识别潦草的英文手写单词方面的优势。题目是：**（4 分）** 判断级数

$$\sum_{n=1}^{\infty} \frac{(-1)^n}{\sqrt{n}}$$

收敛还是发散。请展示所有解题过程，并说明所使用的判别法。


(3) **(4 points)** Determine whether the series $\sum_{n=1}^{\infty} \frac{(-1)^n}{\sqrt{n}}$ converges or diverges. Show all work and state the test you are using.

Solution:

$\sum_{n=1}^{\infty} \frac{(-1)^n}{\sqrt{n}}$ $a_n = \frac{1}{\sqrt{n}} \Rightarrow a_n \frac{1}{n^{1/2}}$

p-series $1/2 < 1$

Divergent

~~B/c the~~ Due to the p-series < 1, due to the Alternating Series Rule, $\sum_{n=1}^{\infty} \frac{(-1)^n}{\sqrt{n}}$ must also be divergent.

Final Answer:

Divergent.

请注意，即使对人类读者而言，英文手写单词也并不容易识别。

GPT-4.1-mini 的光学字符识别输出通过我们的 API 流水线以 LaTeX 格式返回，并在此处呈现为编译后的 PDF：转录文本几乎完美：解答部分完全正确，只有最终答案中有一个拼写错误（可能是由于缺乏上下文所致）。

解答部分：

$$\sum_{n=1}^{\infty} \frac{(-1)^n}{\sqrt{n}} \quad a_n = \frac{1}{\sqrt{n}} \implies a_n \frac{1}{n^{1/2}}$$

p级数 $\frac{1}{2} < 1$ 发散

由于 p级数 ≤ 1 根据交错级数判别法，$\sum_{n=1}^{\infty} \frac{(-1)^n}{\sqrt{n}}$ 也必定发散。——

最终答案：*Dirugnt*。

尽管最终答案中有一个拼写错误，下游评估模块仍正确地将“Dirugnt”识别为“divergent”的拼写错误。

Mathpix 的 OCR 输出经由我们的 API 流水线以 LaTeX 格式返回，并在此处渲染为编译后的 PDF：相比之下，Mathpix 的转录文本在数学符号和英文单词中都包含许多拼写错误：

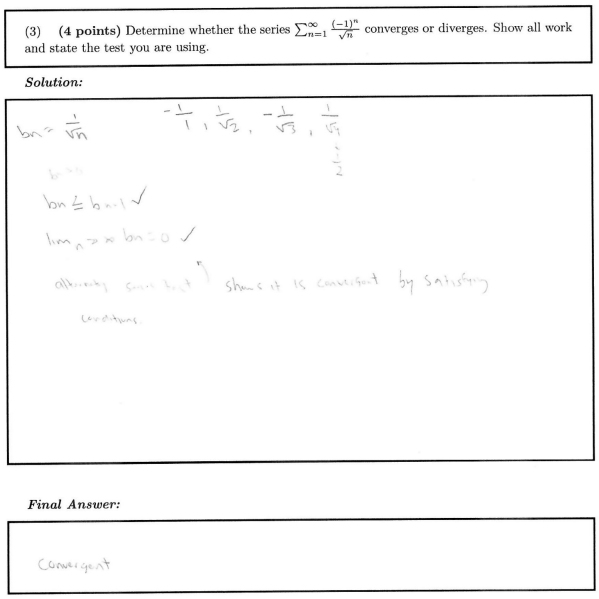
解答部分：$\sum_{n=1}^{\infty} \frac{(-1)^n}{\sqrt{n}}$ ~~an~~ $a_n = \frac{1}{\sqrt{n}} \Rightarrow a_n \frac{1}{n^{1/2}}$ *penns* $1/2 < 1$ 发散

由于 *Dare* 是 *p-serves* < 1，根据交错级数定理，$\sum_{n=1}^{\infty} \frac{(-1)^n}{\sqrt{n}}$ 也必须~~是~~发散的。

最终答案部分：发散。

我们想指出，尽管存在大量单词层面的拼写错误，下游模型仍能根据上下文还原学生的意图推理。尽管如此，也有一些例子表明，拼写错误造成的混淆确实导致了不必要的失分。

示例3。该示例表明，GPT-4.1-mini 能够正确读取非常浅淡的手写文字，并生成正确的下游解读；而 Mathpix 无法做到这一点，导致下游评估中出现了不应有的失分。


(3) **(4 points)** Determine whether the series $\sum_{n=1}^{\infty} \frac{(-1)^n}{\sqrt{n}}$ converges or diverges. Show all work and state the test you are using.

Solution:

$b_n = \frac{1}{\sqrt{n}}$

$b_n \leq b_{n+1}$ ✓

$\lim_{n \to \infty} b_n = 0$ ✓

alternating series test, since it is convergent by satisfying conditions.

Final Answer:

Convergent

GPT-4.1-mini 的光学字符识别输出通过我们的 API 流水线以 LaTeX 形式返回，并在此渲染为编译后的 PDF。： 转录文本完全准确。

解答部分：

$$b_n = \frac{1}{\sqrt{n}} \quad -\frac{1}{1}, \frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{3}}, \frac{1}{\sqrt{4}}$$

$$b_n \leq b_{n-1} \checkmark$$

$$\lim_{n\to\infty} b_n = 0 \checkmark$$

交错级数判别法 $\Rightarrow$ 表明它满足条件，因此收敛。

最终答案部分：收敛

来自 Mathpix 的 OCR 输出通过我们的 API 流水线以 LaTeX 格式返回，并在此处呈现为编译后的 PDF： 转录文本中的数学符号和英文单词都有多处拼写错误，而且遗漏了测试名称中的一个关键部分，导致后续评估失分 1 分（满分 4 分）。

解答部分：

$$b_n = \tfrac{1}{\sqrt{n}} \quad -\tfrac{1}{1}, \tfrac{1}{\sqrt{2}}, -\tfrac{1}{\sqrt{3}}, \tfrac{1}{\sqrt{4}}$$
$$b_n \leq b_{n-1} \surd$$
$$\lim_{n\to\infty} \to b_n = 0/$$
$$curcitiums$$

最终答案部分：收敛

A.4 4.1-mini OCR 失败示例

题目是 **(3 分数)** 求级数的和

$$\sum_{n=3}^{\infty} \frac{2^{n-1}}{3^{n+1}}.$$

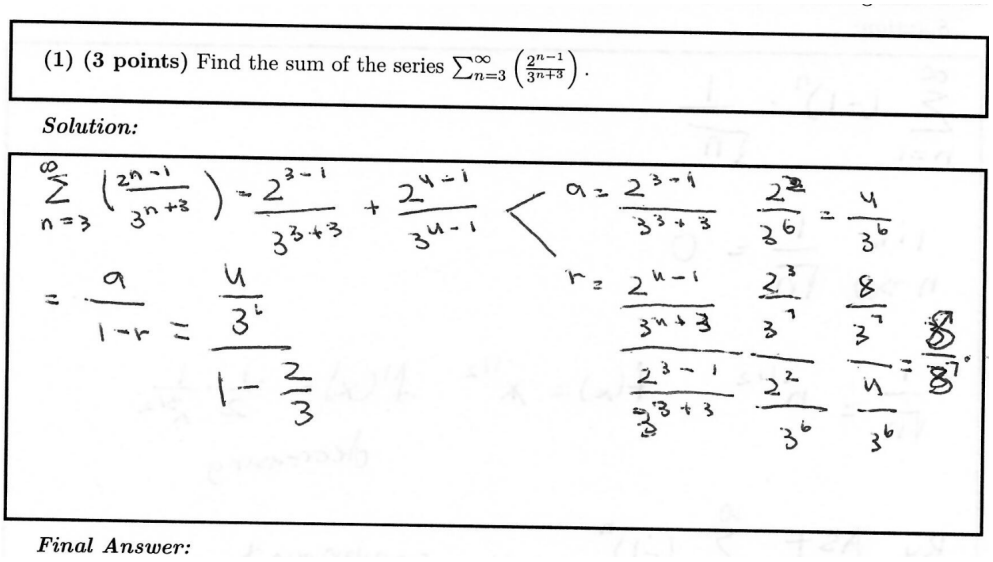


图9：典型的除法相关主要失败模式

来自 GPT-4.1-mini 的 OCR 输出，通过我们的 API 流水线以 LaTeX 形式返回，并在此渲染为编译后的 PDF：下方输出存在两个问题。(1) 内容缺失：分式表达式的一部分被省略，很可能是因为存在多个嵌套分式层。(2) 非预期的自动纠错：模型以视觉上并无依据的方式改动了一个指数，将学生所写的 $2^{4-1}/3^{4-1}$ 改为 $2^{4-1}/3^{4+3}$；后者在此语境下在数学上是自洽的，但并未忠实反映学生的书面作答。

解答部分：

$$\sum_{n=3}^{\infty}\left(\frac{2^{n-1}}{3^{n+3}}\right)=\frac{2^{3-1}}{3^{3+3}}+\frac{2^{4-1}}{3^{4+3}} \quad \leftarrow \quad a=\frac{2^{3-1}}{3^{3+3}} \quad \frac{2^2}{3^6}=\frac{4}{3^6}$$

$$r=\frac{2^{n-1}}{3^{n+3}} \quad \frac{2^3}{3^7} \quad \frac{8}{3^7} \quad \frac{8}{3^7}$$

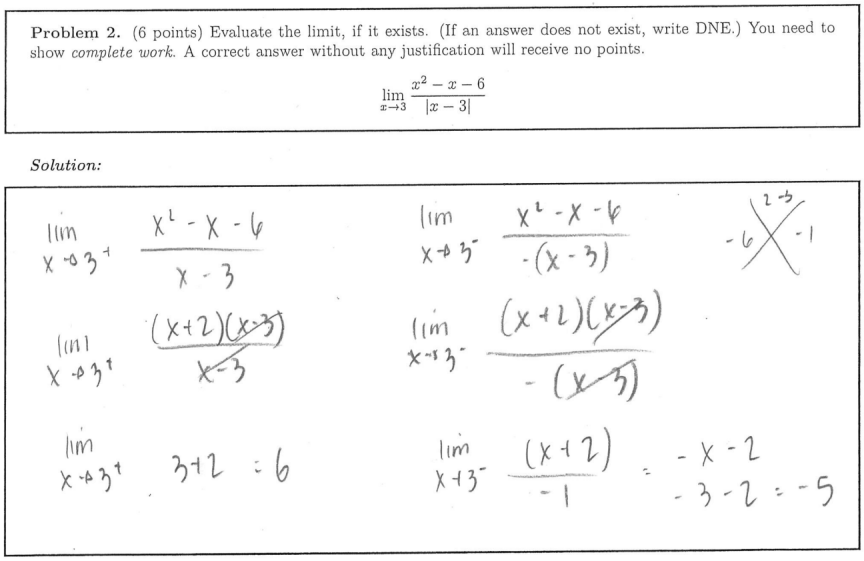
$$=\frac{a}{1-r}=\frac{\frac{4}{3^6}}{1-\frac{2}{3}}$$

A.5 自动校正问题

在本示例中，我们说明了在 OCR 阶段明确指示模型“不要纠正任何数学或逻辑错误或拼写错误”（OCR 提示词，见第3节）以减少不必要的自动纠错的重要性。

示例 4. 请看以下来自 Math 2A 测验的手写解答：求极限（若存在）。（若答案不存在，请写不存在。）你需要展示完整的解题过程。没有论证的正确答案将不得分。

$$\lim_{x\to 3^+}\frac{x^2-x-6}{|x-3|}.$$



Problem 2. (6 points) Evaluate the limit, if it exists. (If an answer does not exist, write DNE.) You need to show *complete work*. A correct answer without any justification will receive no points.

$$\lim_{x\to 3}\frac{x^2-x-6}{|x-3|}$$

Solution:

$\lim_{x\to 3^+}\frac{x^2-x-6}{x-3}$ $\lim_{x\to 3^-}\frac{x^2-x-6}{-(x-3)}$ $-6 \quad 2,-3 \quad -1$

$\lim_{x\to 3^+}\frac{(x+2)(x-3)}{x-3}$ $\lim_{x\to 3^-}\frac{(x+2)(x-3)}{-(x-3)}$

$\lim_{x\to 3^+} 3+2=6$ $\lim_{x\to 3^-}\frac{(x+2)}{-1}=-x-2$ $-3-2=-5$

请注意，学生犯了一个明显的算术错误，写成了 $3+2=6$。我们使用 GPT-4.1-mini 光学字符识别和转录提示词运行了三次，**但未**包含“请勿更正”这条指令：

Output 1:

$$\lim_{x\to 3^+}\frac{x^2-x-6}{x-3}$$
$$=\lim_{x\to 3^+}\frac{(x+2)(x-3)}{x-3}$$
$$=\lim_{x\to 3^+}(x+2)=3+2=6$$

$$\lim_{x\to 3^+}\frac{x^2-x-6}{|x-3|}$$
$$=\lim_{x\to 3^+}\frac{x^2-x-6}{-(x-3)}$$
$$=\lim_{x\to 3^+}\frac{(x+2)(x-3)}{-(x-3)}$$
$$\lim_{x\to 3^+}\frac{x+2}{-1}=\lim_{x\to 3^+}-(x+2)$$
$$=-(3+2)=-5$$

Output 2:

$$\lim_{x\to 3^+}\frac{x^2-x-6}{x-3}$$
$$=\lim_{x\to 3^+}\frac{(x+2)(x-3)}{x-3}$$
$$\lim_{x\to 3^+}(x+2)=3+2=6$$

$$\lim_{x\to 3^-}\frac{x^2-x-6}{|x-3|}$$
$$=\lim_{x\to 3^-}\frac{(x+2)(x-3)}{-(x-3)}$$
$$\lim_{x\to 3^-}\frac{x+2}{-1}=-(x+2)$$
$$=-3-2=-5$$

Output 3:

$$\lim_{x\to 3^+}\frac{x^2-x-6}{x-3}$$
$$=\lim_{x\to 3^+}\frac{(x+2)(x-3)}{x-3}$$
$$=\lim_{x\to 3^+}(x+2)$$
$$=\mathbf{3+2=5}$$

$$\lim_{x\to 3^-}\frac{x^2-x-6}{-(x-3)}$$
$$=\lim_{x\to 3^-}\frac{(x+2)(x-3)}{-(x-3)}$$
$$=\lim_{x\to 3^-}\frac{x+2}{-1}$$
$$=-x-2$$
$$=-3-2=-5$$

正如我们所见，第三次运行时，人工智能悄悄地纠正了学生的错误：

$$3+2=6 \quad \Rightarrow \quad 3+2=5.$$

这种类型的隐式自动纠错在评分数百份试题时可能会成为一个严重问题，因为它改变了原始的学生作答内容。

在我们将“不要纠正任何数学错误、逻辑错误或拼写错误”的指令添加到 OCR 提示词后，即使经过多次运行，也不再出现这种自动纠错行为。

我们注意到，无法完全消除不必要的自动纠错错误（见 A.4节中的示例）。在我们的实验中，此类情况很少见，出现比例低于样本总数的 2%。

A.6 展示系统提示作用的示例

在本节中，我们通过两个示例来说明系统消息（4.1节）中提示词1、2和3的作用。

第一个示例涉及提示词1和2，它们强调评估中的逻辑一致性。

第一个是　　关于提示词 1 和 2，它们在系统消息中强调逻辑一致性　　.

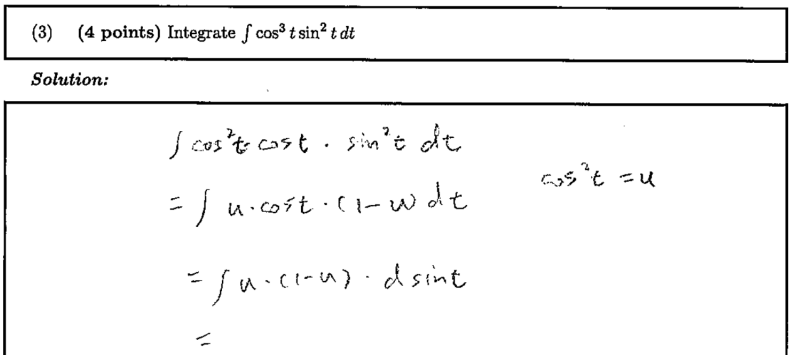
Ex例 5. 以下是来自 Math 2B 的一个样例。该问题是求积分　　e

$$\int \cos^3 t\,\sin^2 t\,dt.$$

标准方法是使用换元

$$u=\sin t.$$

以下是一份不完整的学生解答：



在没有提示词 1 和提示词 2 的情况下，GPT-4.1-mini 的一项评估示例如下：

> 学生首先将被积函数改写为 $\cos^2 t \cos t \sin^2 t$，这是正确的。然而，换元 $u = \cos^2 t$ 并不是求解该积分的常用或有效换元。随后，学生尝试用 u 表示 $\sin^2 t$，但这些步骤在数学上并不成立：$\cos^2 t = u$ 并不能推出 $\sin^2 t = 1 - u$，因为 $\sin^2 t = 1 - \cos^2 t$，而且换元过程也处理得不正确。

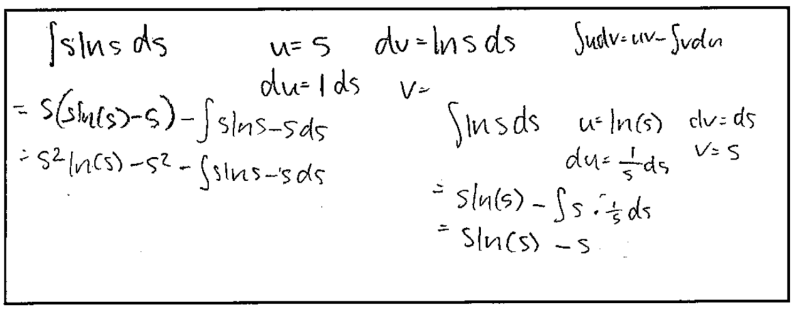
在系统提示词中为提示词 1 和 2 添加更明确的指示，可以减少此类情况。

更普遍地说，这一观察表明，通过 API 部署基于大语言模型的评分器时，需要进行明确的一致性检查。一个有前景的方向是引入轻量级验证步骤，例如采用两阶段流程，重新评估生成的解释是否具有内部逻辑一致性；这将留待未来研究。

学生一开始尝试解题时出错，将错误步骤留在页面上，随后继续求解并得到正确答案，这种情况很常见。下面的例子说明了提示 3 的重要性：它能确保在评估过程中不会因这类探索性的错误尝试而扣分。

示例 6。 以下是 Math 2B 中的一个示例。题目要求计算积分

$$\int s \ln s \, ds.$$



如我们所见，学生先进行了错误的换元，随后又改正了错误，但没有擦除之前的步骤。在没有提示 3 的情况下，ChatGPT 4.1-mini 给出了以下评价，并因最初的错误尝试扣了 1 分（摘录相关内容如下）：

> 学生最初尝试使用分部积分法，但错误地选择了 $u = s$ **and** $dv = \ln(s)\, ds$，这既不合理也没有帮助，因为 $\ln(s)\, ds$ 的积分并非简单可求，且未得到正确计算。

然而，学生后来正确地将问题重新表述为 $u = \ln(s)$ 和 $dv = s\,ds$，这是正确的设定，因此因选择和设定正确而获得 *1* 分（*.....*此处省略部分内容）。然而，由于学生解答的初始部分在应用分部积分法时存在严重错误和混乱，总分被降至 ***2*** 分。学生展现出对正确方法的理解，并得出了正确的最终答案，但在最初的尝试中犯了重大错误。

在系统消息中加入提示 *3* 后，这个问题得到了解决。以下是 *GPT 4.1-mini* 生成的修订版输出：

学生最初尝试使用分部积分，但在解答的第一部分出现了几处符号和设定错误，例如错误地设定了 $u = s$ 和$dv = \ln\, s\, ds$，并混淆了 du和 dv。**不过，学生正确地放弃了错误的开头，随后给出了完全正确且完整的最终解答**，使用了恰当的选取方式 $u = \ln s$ 和 $dv = s\,ds$。最终答案与官方正确答案完全一致，包括积分常数 C。由于最终解答和答案正确且逻辑连贯，因此尽管之前有错误的尝试，仍给予满分（*3* 分）。

A.7 两种人工智能评分量规的比较：案例研究

本节将展示一组具有代表性的评分细则（灵活型与固定型），并以以下测验题为例，说明评分细则框架在实践中的应用：

题目：（3 分）求级数的和

$$\sum_{n=3}^{\infty} \left(\frac{2^{n-1}}{3^{n+3}} \right).$$

评分细则 1：灵活评分标准（数学逻辑；接受重新索引和两种方法）。

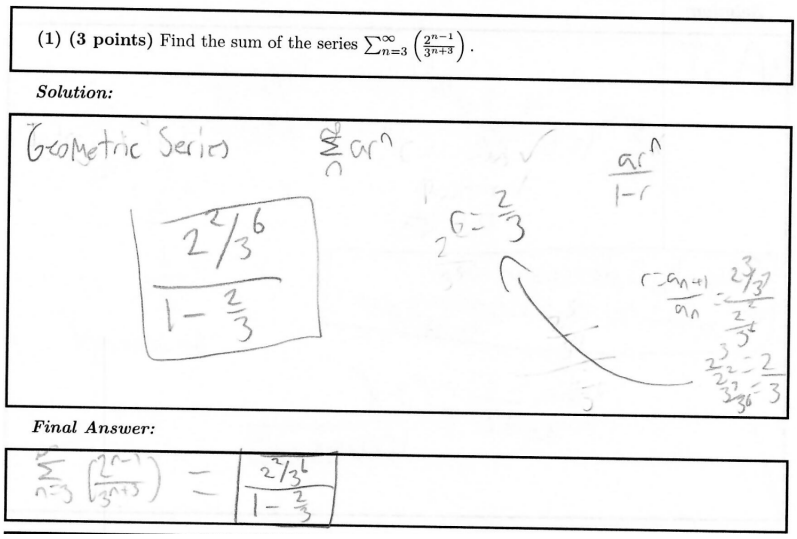
- **30 分** 推理连贯，且最终答案可化简为 $\frac{1}{54}\left(\frac{2}{3}\right)^{n}$和$\frac{1}{81}\left(\frac{2}{3}\right)^{n-1}$，无论使用直接的无穷级数求和公式，还是先求部分和再取极限法。
- **部分分** 可针对任何正确的组成部分给分：正确识别几何级数结构／公比；根据学生的索引方式正确确定首个纳入项；正确设定部分和并包含 $|r| < 1 \Rightarrow r^N \to 0$；等等。

重新索引／等价形式（政策）。任何正确的重新索引（例如 $n \mapsto n + k$、改变上下界、移出常数）均应给满分，前提是学生的首个纳入项与公比一致。将代数等价形式视为相同（例如 $\frac{1}{54}\left(\frac{2}{3}\right)^{n}$和$\frac{1}{81}\left(\frac{2}{3}\right)^{n-1}$）。不要因规范化方式而扣分；只有当首项或上下界不一致时才扣分。

在其他题目中，我们可以通过为关键评分标准项目指定有界分数区间（例如，对证明某个关键中间步骤有效性给 1–2 分），进一步提高评分的灵活性，而不是像下面的固定评分标准那样只采用固定扣分方式。

评分标准2：固定评分标准（分数映射；明确处理重新索引）。1. 等比级数识别与公比（1.0 分）。指出或运用该级数是等比级数，并正确求出 $r = \frac{2}{3}$（或将其改写为等价的等比形式）。**2. 首个纳入项／正确重新索引（1.0 分）**。根据其索引方式求出正确的首个纳入项（例如 $a_3 = \frac{4}{729}$），或采用正确的重新索引，并保持首项一致。**3. 求和步骤与最终值（1.0 分）**。正确运用 $S = \frac{a}{1-r}$，或给出正确的部分和公式，并通过 $|r| < 1$ 证明极限。最终答案必须化简为 $4\,243$。若解题思路完全正确，但有轻微的算术错误，此处给 $0.5/1.0$ 分。**评分指南**。3.0 — 三项均正确。2.0–2.5 — 轻微的代数／索引错误；或采用部分和解法但未明确写出极限，不过显然蕴含 $r^N \to 0$。1.0–1.5 — 有部分进展（例如求出 r 或首项），但求和或极限错误／不完整。0–0.5 — 几乎没有相关解题过程，或方法错误（例如裂项求和）。

示例 7. 我们通过此例说明，由于设计原则不同，两种评分细则对同一份学生解答的评价也有所不同。在本例中，OCR 输出可以接受，学生的解题过程在数学上是正确的，但呈现方式并非标准形式，且略显随意。**评分细则 1（灵活、基于推理）给出满分（3.0 分，与助教给出的分数相同），而评分标准2（固定、按分值映射）则给出部分分（2.0 分）**。



评分细则 ***1*** 的反馈：

学生正确地将该级数识别为几何级数，并正确写出了无穷等比级数的求和公式。他们正确地将首个纳入项识别为$\frac{2^2}{3^6}$，并将公比识别为$\frac{2}{3}$。给出的最终答案 $\frac{2^2/3^6}{1-\frac{2}{3}}$ 在代数上等价于正确的和$4\,243$。尽管有些中间步骤没有充分解释，或略显

不清楚（例如，含有 $\frac{r-a^{n+1}}{a^n}$ 的那一行并非标准记法，而且似乎没有必要），但关键要点——识别几何级数结构、首项、公比并应用无穷和公式——都已包含且正确。最终答案的数值与参考解答一致。因此，给予满分。

评分标准*2*的反馈：

学生正确识别了公比 $r = 2/3$，满足第一条标准（*1.0*分）。然而，围绕首项和重新索引的记号与解释不够清晰且前后不一致。学生将首项写为 $2^2/3^6 = 4/729$，这是正确的，但周围的表达式以及涉及 r 和 a^{n+1} 的分数令人困惑，未能清楚地确立一致的基准项或索引方式。因此，针对首个纳入项/正确重新索引的第二项分数仅部分满足，故此处给予*0.5*分。对于求和步骤，学生将和式写为$(2^2/3^6)/(1-2/3)$，这是从 $n = 3$ 开始、公比为 $2/3$ 、首项为 $4/729$的几何级数求和的正确公式。尽管学生没有明确将分数化简为 $4/243$，也没有论证 r^N 的极限为零，但整体设定基本正确，且最终框出的答案与正确公式相符。因此，求和步骤给予*0.5*分。总体而言，该解答表明学生对几何级数和公比有所理解，但围绕首项和重新索引的解释与记号不完整且略显混乱，求和的论证也较为简略。因此，总分为：$1.0 + 0.5 + 0.5 = 2.0$。

这个示例凸显了两种评分标准的互补特性。固定评分标准按各个组成部分分别给分，并对不清晰的索引和符号表示进行扣分；而灵活评分标准则评估整体数学一致性和代数等价性，因此给出满分，与助教给出的分数一致。尽管如此，我们想强调，灵活性并不会系统性地提高分数；在其他情况下，如果局部正确的部分在整体上存在矛盾，灵活评分标准给出的分数会更低。

A.8 代表性评分标准指南（节选）

以下节选展示了我们在实验中使用的评分标准说明，旨在减少因呈现方式不同或OCR错误导致的不合理扣分。本节选具有代表性，但并未涵盖所有情况。

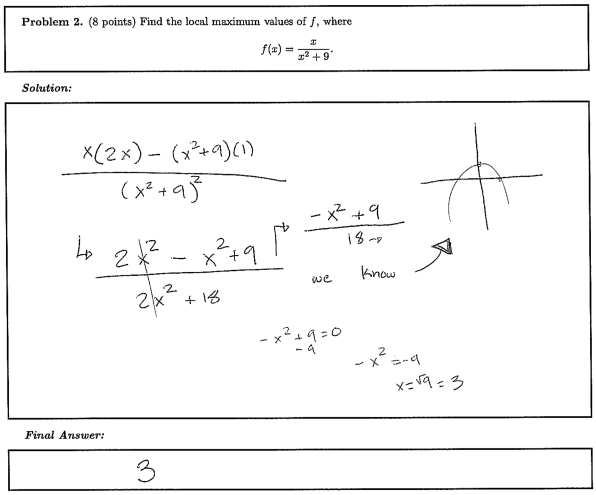
不因格式扣分。如果数学内容和所需论证均正确，不要因手写文字质量、轻微拼写错误、非正式措辞或不常见但清晰的符号表示等呈现问题而扣分。如果学生的最终答案看起来有轻微的拼写或格式问题，或存在可能由*OCR*导致的细微数值或字符差异，但解答步骤明确支持正确的数值，则不要因此扣分。

A.9 AI评分与助教评分：案例研究

在本小节中，我们展示一个有代表性的案例：人工智能给出的分数高于助教给出的分数。这个例子说明，即使双方的评分决定都有合理依据，仍可能出现评分差异。

例 8. 以下问题选自 Math 2A 的一次测验：求 f 的局部最大值，其中

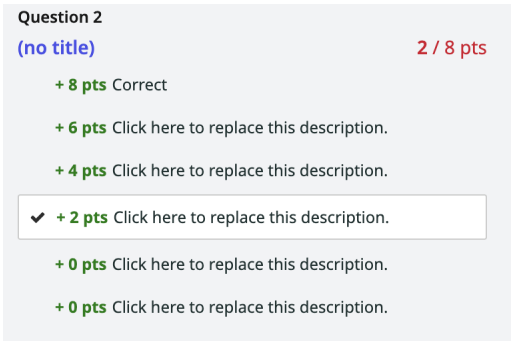
$$f(x) = \frac{x}{x^2 + 9}.$$



对于这份提交内容，助教给了 2 分，而 AI评分器给了 5 分。以下是 AI评分器的评语：

> 学生尝试使用商法则，但在化简导数表达式时出现代数错误，因此得到的导数形式不正确。这导致其在求导部分失去 *2* 分（标准 *1*）。学生正确求出了 $x = 3$ 是一个临界点，但漏掉了 $x = -3$，因此在求临界点部分得 *1* 分（标准 *2*）。学生没有提供分类或解释来确认 $x = 3$ 是一个局部最大值，因此仅因找出最大值而未提供论证，获得 *1* 分（标准 *3*）。学生没有计算或报告最大值 $f(3) = 1/6$，因此在报告最大值部分得 *0* 分（标准 *4*）。总的来说，最终答案不完整且缺乏充分论证，但找到了关键临界点 $x = 3$。

AI评分器采用基于标准的评分细则，分别评估（i）求导的正确性，（ii）临界点的识别，（iii）极值的分类，以及（iv）最大值的报告，并对每个部分分别给予部分分。相比之下，助教评分采用下方所示的 Gradescope 粗略评分标准，其中只有少数几个分数档：



这个例子表明，较大的分数差异可能源于评分标准结构不同，而非评分错误。粗略的助教评分标准将部分理解的多个方面合并到单一的低分档中，而更细致的人工智能评分标准则汇总解答多个维度的部分分。因此，尽管解答明显不完整，人工智能给出的总分仍然更高。重要的是，依据各自的评分标准，这两个分数都有其合理性。这个案例说明，AI 与人工评分差异可能反映的是评分分辨率和评分标准的表达能力，而非任何一方评分者存在系统性偏差或错误。

A.10 GPT-4.1-mini 与 o3-mini：比较结果

使用 GPT-4.1-mini 生成的固定 OCR 输出，我们在一组包含 57 份 Math 2B 测验（171 个题目实例）的数据上评估评分表现，这些测验对应于正文（第3节）所述的同一**171 份解答子集**。每份测验都包含相同的三道高难度方程。对于每个问题，GPT-4.1-mini 和 o3-mini 均在相同条件下运行三次。设 $\mu_{m,\ i}$ 和 $\sigma_{m,\ i}$ 分别表示模型$m \in \{$4.1-mini、o3-mini$\}$ 对问题 i给出的平均分和运行间标准差，并定义每题的平均分差

$$\Delta_i = \mu_{\text{o3},i} - \mu_{\text{4.1},i}.$$

A.10.1 模型内评分稳定性

表2总结了模型内的评分稳定性。与 o3-mini 相比，GPT-4.1-mini 的运行间变异性显著更低，表现为平均标准差更小，且零方差的问题占比更高。这表明，在低温度设置下，GPT-4.1-mini 在重复评估中给出的分数更加稳定。

模型	# 问题	均值 σ	Pr($\sigma = 0$)
GPT-4.1-mini	171	0.083	0.772
o3-mini	171	0.319	0.725

表2：使用固定OCR输入时，每个问题进行三次独立运行所得的模型内评分稳定性。 σ 表示每个问题的运行间标准差在所有问题上的平均值。

A.10.2 跨模型分数差异

表3报告了跨模型分数差异的总体统计数据。平均而言，o3-mini 与 GPT-4.1-mini 的平均分差异较小，约45%的问题得分完全一致。这些结果表明，尽管两种模型的运行间稳定性有所不同，但在总体层面上，它们给出的分数大体相似。

统计量	数值
问题数	171
均值 Δ_i	0.087
均值 $\lvert\Delta_i\rvert$	0.315
Pr($\Delta_i = 0$)	0.464

表3：o3-mini 与 GPT-4.1-mini 的跨模型分数比较。对于每道题， i, Δ_i表示各模型三次运行的平均分之差。

A.10.3 人工评审员比较

为进一步评估评分变异性，我们使用两位独立人工评审员的成对判断来比较模型表现。两位评审员均认为，GPT-4.1-mini 和 o3-mini 的表现最常被判定为相近；在评分接近度和反馈准确性方面，平局约占 60–72%。若观察到差异，评审员更常倾向于 o3-mini。此外，两位评审员都更倾向于对多次运行的 o3-mini 分数取平均。在接下来的两张表中，百分比四舍五入至小数点后两位，因此总和可能并非恰好为 100%。

比较	选项	百分比
评分准确度比较 哪个模型更接近审稿人的评分?	4.1 更接近 o3 更接近 T（平局）	7.74% 29.17% 63.09%
评估准确率对比 哪个模型提供的反馈更准确?	4.1 更准确 o3 更准确 T（平局）	3.59% 23.35% 73.06%
取平均能提高准确性吗? （o3 运行三次；选择最接近平均分）	Yes No	62.22% 37.78%

表4：审稿人1：GPT-4.1-mini（“4.1”）与 o3-mini（“o3”）的成对比较。

比较	选项	百分比
评分准确度比较 哪个模型的评分更接近审稿人的评分?	4.1 更接近 o3 更接近 T（平局）	14.97% 17.96% 67.06%
评估准确率对比 哪个模型提供的反馈更准确?	4.1 更准确 o3 更准确 T（平局）	9.58% 25.15% 65.28%
取平均能否提高准确性? （将 o3 运行三次；选择最接近平均分的那次运行）	Yes No 不清楚	51.11% 26.67% 22.22%

表 5：审稿人 2：GPT-4.1-mini（“4.1”）与 o3-mini（“o3”）的成对比较。

与多数投票自洽性的关系 [8]。 我们的多次运行稳定化会从三次运行中选出分数最接近均值的那次，而不是像多数投票自洽性那样选择出现频率最高的分数。为考察两者在实践中的差异，我们分析了 45 个 o3-mini 三次运行结果并不完全相同的案例。在其中大多数案例中，两次运行的分数相同，因此多数投票与选择最接近均值的结果等效。只有五个案例出现了三个不同的分数；这些案例在不同评审员之间的结果不一，没有哪种规则表现出一致的优势。这表明，在课程规模下运行三次时，选择最接近均值的结果在大多数情况下与多数投票的表现相似，仅在少数确实存在歧义的案例中有所不同。在课程规模部署中，我们将运行次数限制为三次，以兼顾稳定性和成本；对大量样本进行多数投票并不实际。在少数三次运行结果全不相同的案例中，这种差异反映了评分本身存在的歧义，因此会采取保守处理（例如标记出来以供后续审核）。下一步，我们将量化三方意见分歧在大规模应用中出现的频率；如果仍然很少，我们将采用保守的

策略，将此类案例交由人工审阅；如果这类情况更为频繁，则需要制定额外的稳定化策略，同时考虑分数和反馈，而非仅仅汇总分数。

A.10.4 评阅者间差异

最后，为了将 AI 与人工评分差异置于适当背景下，我们使用与 AI评分器 相同的基于标准的评分细则，考察人工评分者之间的差异。两位独立评审员对相同的 57 份测验（171 个题目实例）进行了评分。图 10 展示了每道问题的分数差异分布。虽然相当一部分问题（约 45%）的评分完全一致，但跨越多个分数等级的显著差异也很常见。这说明，即使采用共同且明确的评分标准，人工评分仍然具有固有的主观性，也为解读 AI 与人工评分差异提供了一个自然的基准。

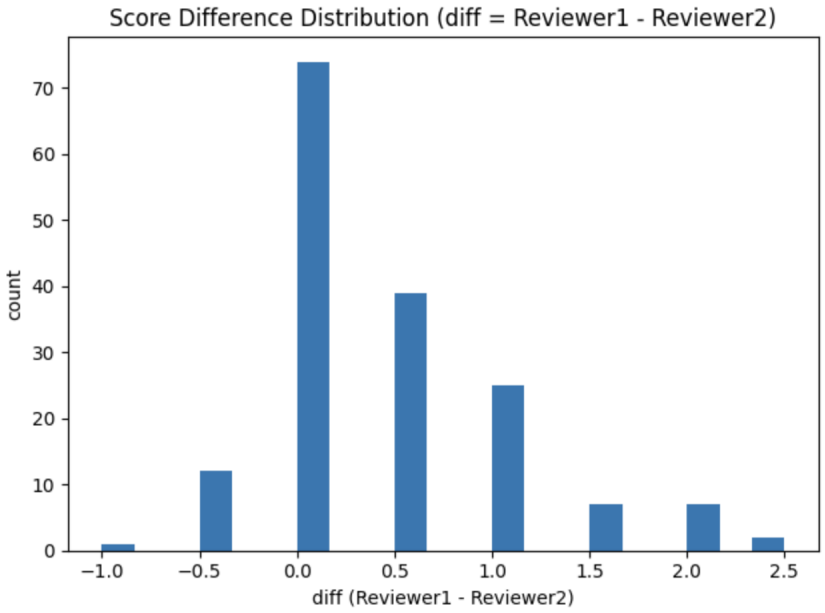


图 10：两位人工评审员在每道问题上的分数差异分布（审稿人1减去评审员 2）。

A.11 面向学生的示例消息

下面是一则发送给 Math 2B 学生的匿名示例消息，其中包括给出的分数及相应反馈。所有消息均采用相同格式，只有每道问题的具体内容有所不同。

> 亲爱的（学生姓名）：
>
> 感谢您参与人工智能评分测试。您的试题 AI 评分结果如下：
>
> 问题 1：
> 分数：5

评估：学生的解答正确。他们正确运用了分部积分，正确求出了原函数，并准确计算了积分限。最终表达式 $(5/9)e^6 + 1/9$ 与预期答案 $1/9(5e^6 + 1)$ 在代数上等价。计算和代数运算均无错误，因此给予满分。

问题 2：

分数：2

评估：你正确回忆起了恒等式 $sec^2x = 1 + tan^2x$，并尝试据此改写被积函数，这是一个不错的开始（标准 1）。不过，问题出在你将 sec^6x 表示为 $(sec^2x)^3$，然后将每个 sec^2x 替换为 $1+tan^2x$。正确的方法是将 sec^6x 因式分解为 $sec^4x \cdot sec^2x$，这样就可以使用换元 $u = tanx$（其中 $du = sec^2xdx$），并将 sec^4x 改写为 $(1+tan^2x)^2$，而不是 $(1+tan^2x)^3$。改写被积函数时的这一步错误导致多项式不正确（标准 3），进而使原函数也不正确（标准 4 和 5）。虽然你正确地对所得各项进行了积分，但代数运算设置有误。因此，你获得 5 分中的 2 分。

总分：7

请注意：AI评分仅供参考。助教在 Gradescope 中给出的分数为官方分数。我们的AI评分包含两个基于大语言模型的步骤——（1）光学字符识别以读取手写作业，以及（2）对光学字符识别文本的评估。由于大语言模型具有概率性，偶尔可能出现错误。