arXiv:2603.00925v1 [cs.CL] 1 Mar 2026

DrawEduMath 的后续影响：视觉语言模型在面对学习困难的学生时表现不佳，并误诊错误

LiLucy[△] **AlbertZhang**[+] **Nathan Anderson**[÷] **Ryan Knight**[+] **Kyle Lo**[□△△]华盛顿大学 [+]Insource Services [÷]伍斯特理工学院[□]艾伦人工智能研究所
lucy3li@cs.washington.edu kylel@allenai.org

摘要

有效的数学教育需要识别学生的错误并作出回应。要让人工智能支持教学应用，模型必须在不同熟练程度的学生群体中表现良好。我们的研究全面呈现了 11 种视觉语言模型（VLM）在 DrawEduMath 上的表现；DrawEduMath 是一个问答基准，包含真实学生针对数学题的手写、手绘作答。我们发现，模型的弱点集中在数学教育的一个核心环节：学生错误。所有受评估的视觉语言模型在描述需要更多教学帮助的学生作答时都表现不佳；在所有问答中，它们在评估学生错误相关的问题上也最为吃力。因此，尽管视觉语言模型可能经过优化，成为数学题求解专家，我们的结果表明，要充分支持教育应用场景，还需要采用其他开发激励机制。

1 引言

视觉语言模型（VLMs）在教育中的应用，在学术研究（Küchemann 等人，2025年；Lee 等人，2025b）和商业人工智能产品中都日益受到关注。后者的例子包括集成了 Gemini 的 Google Classroom，[1] 以及可汗学院的人工智能导师 Khanmigo，[2] 由 OpenAI 模型提供支持。然而，将这些模型整合到辅导和课堂环境中，往往缺乏透明、开放且贴近现实的评估。鉴于这一空白，我们此前发布了 DrawEduMath（图1），这是一个由 2,030 张教师标注的真实学生手绘 K-12 数学题作答图像构成的数据集（Baral 等人，2025年）。相比于其他多模态数学理解或解题基准测试（Alshammari 等人，2026），DrawEduMath 包含从在线学习平台采集的嘈杂、自然真实的数据（图1）。自发布以来的一年里，我们持续更新了该基准测试的排行榜[3]，纳入了更新的模型。

本文呈现了 DrawEduMath 发布一年后 11 个视觉语言模型在该数据集上的表现概况（图 2）。我们总结出两项主要发现：

F1值：与不含错误的学生作业相比，视觉语言模型更难描述包含数学错误的学生作业内容。

F2：在与评估学生答案正确性相关的问题类型上，视觉语言模型仍然表现最差。

这些发现表明，视觉语言模型无法很好地服务于需要额外教学支持的学生（**F1值**），也无法恰当地识别出需要支持的情形（**F2**）。

为进一步探究这些模式，我们在§4-§8中进行了五项分析，针对可能与模型性能相关的因素。实验表明，即使控制题目因素（**§4**），并降低图像噪声（§5），F1 值方面的性能差距依然存在。此外，**F1 值**表现不佳的一种可能解释是，视觉语言模型预期输入图像中的数学内容是正确的。事实上，我们发现，对于有错误的学生作业，一些模型给出的错误预测答案与无错误学生作业的标准答案相似（§6）。

我们还发现，当提供学生作业的标准自然语言描述（**§F2**）时，模型在评估学生作答正确性（§7）方面确实有所改善。然而，在获得额外文本支持的情况下，它们在这些问题上的表现仍落后于其在其他问题类型上的开箱即用表现。最后，尽管模型在二元正确性问题（例如“学生是否___正确完成？”）上的表现似乎优于开放式问题

[1]https://blog.google/outreach-initiatives/ 教育/classroom-ai-features/[2]https://www.khanmigo.ai/

[3]https://drawedumath.org/

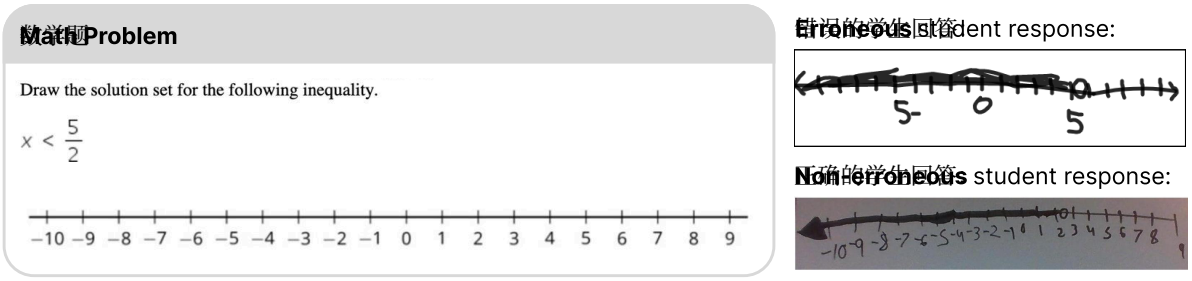


图1：左侧是一道数学题，要求学生在数轴上画出 $x < 5/2$。右侧展示了两个正确性不同的学生作答示例。DrawEduMath将每道数学题与一份学生作答配对，并提示视觉语言模型回答有关该学生作答的问题。

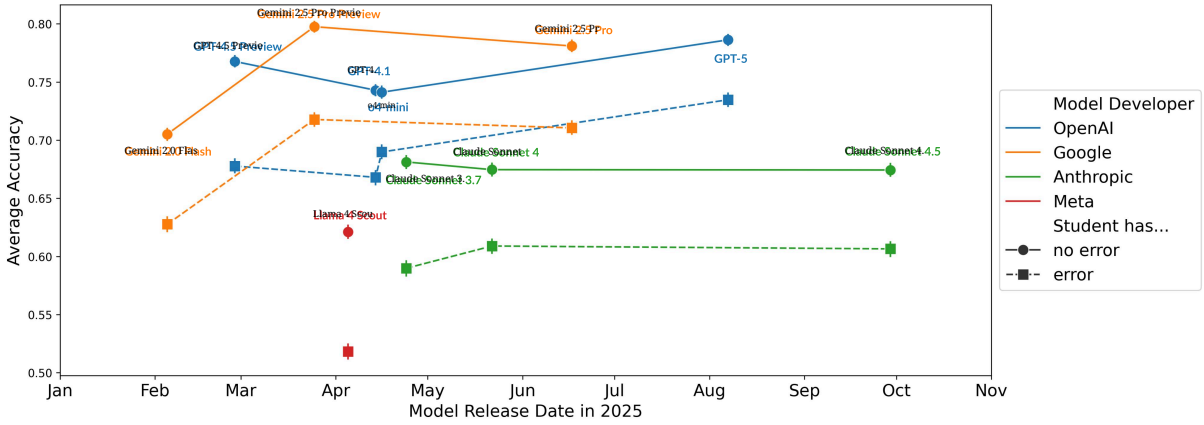


图2：在回答有关错误学生作答的 DrawEduMath 基准测试问题时，视觉语言模型的表现始终较差。正确学生作答的表现（○）标注了具体视觉语言模型的名称；同一模型在错误学生作答上的表现则显示在其正下方（□）。误差线表示95%置信区间。

（例如“学生在作答中犯了什么错误？”），但一些视觉语言模型的表现有时仅略高于随机水平（§8）。

总而言之，我们对视觉语言模型在真实学生数学作答上的表现进行了深入的错误分析，从而更清晰地展现了它们在支持 K-12数学教育方面的弱点。我们公开数据和脚本，以便复现研究结果。[4]

2 背景与相关工作

多模态数学基准测试。 数学内容中丰富的图表，为评估模型的多模态能力提供了绝佳素材。因此，许多视觉语言基准测试的创建工作都聚焦于数学（Alshammari 等人，2026年；Zhang 等人，2024；Lu 等人，2024；Yan 等人，2025年）。教育环境带来了额外挑战：评估视觉语言模型时，可能需要考察其进行更高层次教学推断，以及处理手写和手绘内容的能力（Parsaeifard 等人，2025年；Latif 等人，2025年；Nath 等人，2025年；Nguyen 等人，2025年）。例如，MathCog 要求模型通过一组二元“是/否”问题和一个数字化手写数据集来诊断学生的认知技能（Jin 等人，2025年）。在这一既有研究领域中，DrawEduMath 仍是一项重要的评估资源，因为它涵盖多种图像和问题类型，采用带有噪声的真实学生作业，并纳入资深教师标注（Baral 等人，2025年）。

AI 与学生错误。长期以来，学生错误与误解一直是教育研究的关注重点（Smith III 等人，1994年；Radatz，1979年；Borasi，1994年；Metcalfe，2017年）。随着人们日益关注将 AI 用作辅导教师和助教，研究重点转向模型识别学生错误的能力（Srivatsa 等人，2025年；Kochmar 等人，2025年；Daheim 等人，2024）；推理常见误解（Rittle-Johnson 等人，2025年；Ross 和 Andreas，2025年）；纠正错误（Mita 等人，2024年）；以及提供反馈（Kaliisa 等人，2026年；Botelho 等人，2023年；Stahl 等人，2024年）。此外，

[4] https://github.com/lucy3/aftermath_drawedumath

关于模型应对用户错误时鲁棒性的研究也有一些，但其中大部分关注提示词中的语言错误（Gan 等人，2024；Chatterjee 等人，2024）。关于数学错误如何影响模型表现的研究则很少：其中一个例子是Daheim 等人（2024），他们指出，与无错误的学生数学作答相比，语言模型在验证存在错误的学生数学作答是否正确时表现更差。我们的研究在更多问答类型和多模态数据上也得出了类似结论，不过，对于模型正确性和错误评估的结果则呈现出较为复杂的图景（§8）。

教育中的人工智能的风险。人工智能与教育领域可被视为一种旨在促进“社会公益”的人工智能应用（Cowls 等，2021），其目标是抵消人工智能对人类技能形成和认知思维的负面影响（Bastani 等，2025年；Klimova 和 Pikhart，2025年；Shen 和 Tamkin，2026；Lee 等，2025a）。然而，尽管其最终目标可能被乐观地描述，教育中的人工智能并非没有风险（Blodgett 和 Madaio，2021；Holstein 和 Doroudi，2021）。我们的研究与探讨人工智能可能对不同学生群体产生差异化影响的文献相一致（Schaller 等，2024；Capraro 等，2024；Hadar Shoval，2025年）；我们采取的独特方法是，根据学生展现出的数学熟练度对其输入进行分组。

3 使用 DrawEduMath 评估视觉语言模型

3.1 数据

DrawEduMath 是一个英语数据集，包含 2,030 张学生对 K-12 数学题的手写作答图像（Baral 等人，2025年）。图像由在线学习平台 ASSISTments 提供，其中的数学题来自开放教育资源（Heffernan 与 Heffernan，2014年；Feng 等人，2025年）。每张图像左侧为一道数学题，右侧为学生作答，并附有三类数据：

1. **自由形式描述**（2.0k+），由教师描述每张学生作答图像。
2. **合成问答对** (44.4k+)，由 Claude-3.5 Sonnet 和 GPT-4o 将教师图像说明的各个方面重新格式化为问答，例如 在图像左侧，学生写下了单词 *syrup* → 学生在图像左侧写下了什么单词？*Syrup*。
3. **教师编写的问答对**（11.6k+）。教师为每道数学题编写了一组共用问题，随后针对每道题的每份学生作答给出答案。教师还额外回答了两个通用问题：学生在作答中犯了哪些错误？ 以及学生使用什么策略来解决该问题？ ，范围涵盖所有题目和学生回答。

DrawEduMath 包含由七种问题类型构成的分类体系。在我们的分析中，我们将该分类体系简化为三种类型：图像创作与媒介（12.3%）、正确性与错误（8.5%）以及内容描述（79.2%）。前两类与基准测试的原始分类一致，而第三类问题类型是对其余所有问题类型的汇总，涵盖从学生的问题解决策略到绘制/书写元素的含义、位置和出现频率。我们在正文中汇总这些问题类型，因为我们的主要发现可推广至更细粒度的类型（附录 A）。原始 DrawEduMath 论文还包含更多数据集统计信息和问答示例（Baral 等人，2025年）。

3.2 评估设置

评分指标。 Baral 等人（2025年）使用 Mixtral 8x22B，根据 1（答案差异很大）至 4（基本相同）的量表评判视觉语言模型生成的答案与标准答案的相似度，并在计算模型准确率时将这些评分二值化。随着大语言模型不断更新，Mixtral 8x22B 等较旧模型在模型 API 服务中逐渐被弃用。因此，我们的所有评估均使用更新后的大语言模型评审器。我们采用三位评审器的多数投票结果：Claude Sonnet 4.5、Gemini 2.5 Pro 和 GPT-4o。[5] 我们更新后的评审器与 $\rho = 0$相同的一组人工判断之间的相关性（Spearman .808）与Baral 等人（2025年）的原始评审器（ρ = 0.801）相近。我们计算并报告二值化的模型准确率；1–2 分计为错误，3–4 分计为正确。

模型。 我们在 DrawEduMath 上评估了 11 个于 2025年发布的视觉语言模型。模型涵盖四家开发者：Open AI（GPT-4.1、GPT-4.5 Preview、o4-mini、GPT-5）、Anthropic（Claude Sonnet 3.7、Claude Sonnet 4、Claude Sonnet 4.5）、Google（Gemini 2.0 Flash、Gemini 2.5 Pro、Gemini 2.5 Pro Preview）、

[5]尽管 GPT-4o 是较旧的模型，但它与人工标注的个体相关性高于另外两个较新的模型。因此，我们将其纳入评审模型集合。

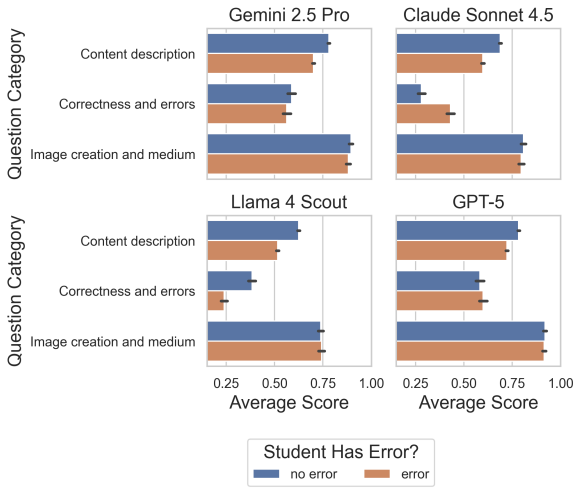


图3：内容描述问答始终是导致学生回答中存在错误与不存在错误时视觉语言模型性能差距的关键因素。附录A列出了更多视觉语言模型，进一步印证了这一发现。

以及 Meta AI（Llama 4 Scout）。尽管我们的主要发现 **F1值** & **F2**适用于所有这些模型，但第 §5-§8 节中的分析聚焦于四个具有代表性的模型：Gemini 2.5 Pro、Claude Sonnet 4.5、GPT-5 和 Llama 4 Scout。

学生错误标注。为了判断学生的数学作答是否包含错误，我们使用教师对以下问题的回答：学生在作答中犯了哪些错误？我们请 GPT-5-mini 解读每个开放式回答，并将其分类为是，即教师描述了某种错误；或否，适用于教师的回答为以下说法或其变体的情况：没有错误或学生没有犯错（附录B.1）。我们通过人工核查的随机抽样200个样例（F1 = 0.984）验证了这一语言模型标注器。

3.3 主要发现

平均而言，当学生作答中包含错误时，模型的表现往往更差（图2）。我们发现，这一模式主要由内容描述问答（**F1值**，图3）所驱动。此外，Baral 等人(2025年)在较早期视觉语言模型中报告的一项弱点，在新一代模型中依然存在：与学生作答正确性和错误相关的问题仍然最难回答（**F2**）。接下来的几节中，我们将深入探讨五个可能与这些发现有关的因素：题目效应（§4）、图像噪声（§5）、题目回答默认模式（§6）、视觉理解瓶颈（§7）以及问题开放性（§8）。

模型	β_1
Gemini 2.0 Flash	0.0944***
Gemini 2.5 Pro Preview	0.0888***
Gemini 2.5 Pro	0.0889***
GPT-4.1	0.0865***
GPT-4.5 Preview	0.0874***
o4-mini	0.0612***
GPT-5	0.0585***
Claude Sonnet 3.7	0.0922***
Claude Sonnet 4	0.0816***
Claude Sonnet 4.5	0.0842***
Llama 4 Scout	0.1013***

表1：学生作答正确性（β_1）对视觉语言模型在内容描述问答上的准确率的估计影响，其中所有 $p < 1.0^{-12}$。

4 即使控制题目因素，模型在错误学生作答上的表现仍较差

一种可能是，**F1值**所观察到的模型性能差距实际上并非受到学生错误有无的影响，而是因为某些数学题对视觉语言模型而言更难理解。DrawEduMath 包含 188 道独特的数学题，涉及从几何到分数等概念。这些题目涵盖多个年级，平均每道题有 12.64 份学生回答（Baral 等人，2025年）。在这里，我们表明，即使控制题目因素，学生错误对视觉语言模型内容描述问答表现的影响在统计上仍然显著。

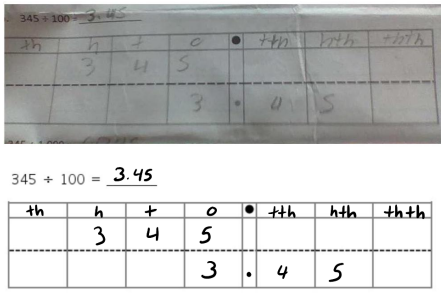
我们使用包含题目固定效应的普通最小二乘回归进行估计：

$$y_{ij} = \beta_0 + \beta_1(s_{ij}) + u_j + \epsilon_{ij}$$

在上述方程中，y_{ij} 表示模型对学生回答 i 和题目 j，u_j 在内容描述问答中的平均得分；每道题对应一个固定效应，ϵ_{ij} 则表示残差。如果学生作答正确，$s_{ij} = 1$；否则s_{ij} =为 0。表1 列出了 β_1 各视觉语言模型的取值。这些数值表明，即使控制了题目因素，无错误学生作答仍与视觉语言模型较高的表现显著相关。

5 个模型的表现差距并未受到图像噪声的显著影响

对**F1值**的另一种可能解释是，出现数学错误的学生提交的图像可能只是噪声更多。ASSISTments 上的学生可以通过数字绘图提交答案，也可以上传纸笔作业的照片，这些照片可能



345 ÷ 100 = 3.45

th	h	t	o	•	tth	hth	thth
	3	4	5				
			3	.	4	5	

图4：示例展示了学生作答图像（上）如何经过我们的数字重绘流程转换并修整（下）。这名学生使用位值表来展示 345 除以 100 后各位数字的数值如何变化。

包含污迹和模糊。在本节中，我们将探讨：即使以标准化方式将学生的回答重新绘制在数字画布上，**F1值**所揭示的模型性能差距是否依然存在?

5.1 实验设置

重绘图像是一项耗时的工作，需要仔细解读每道数学题以及学生原始作答的意图。因此，在本实验中，我们从每道题中分层抽取一个错误的学生回答和一个正确回答，总计得到 336 张图像。尽管样本较小，但在视觉语言模型对内容描述问答的表现方面，错误与非错误学生作答图像之间仍存在统计学显著差距（表 2）。

为确保一致性，第一作者使用数字笔和绘图应用 Procreate 重绘了所有抽样的学生作答。负责重绘的作者保留了学生原本的内容布局，并参考教师对图像的说明，以消除歧义，避免错误解读题目和学生回答。如有需要，作者还使用 Figma 重制了方格纸网格等元素，并输入了文字内容。图4展示了重绘如何改变学生作答的示例。

5.2 结果

对于重新绘制的图像，视觉语言模型的表现变化方向符合预期：图像噪声越少，得分通常越高（图5）。尽管要求学生只提交原生数字内容，可能会让 AI 更容易解读他们的作业，但并非所有教室都具备实施此类标准化要求所需的资源和政策。此外，纸笔作业仍然至关重要，研究

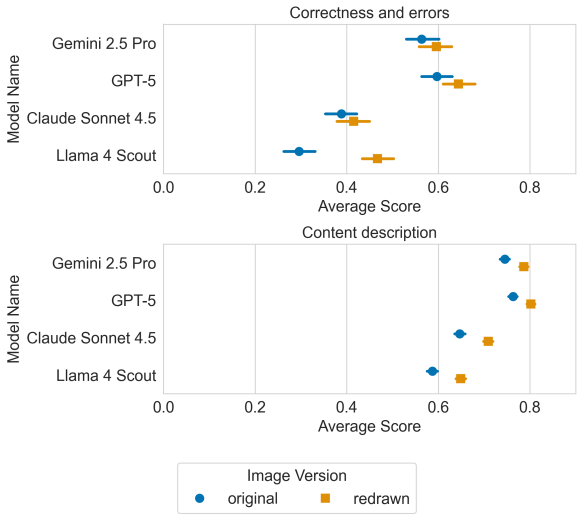


图5：模型在内容描述问答中的表现通常在图像被重绘后有所提升。误差线为95%置信区间。

模型	原始图像	重绘图像
Gemini 2.5 Pro	-0.089***	-0.096***
Claude Sonnet 4.5	-0.100***	-0.058**
GPT-5	-0.055**	-0.043**
Llama 4 Scout	-0.092***	-0.073***

表2：重绘后，错误的学生图像与无错误的学生图像在内容描述 QA 平均得分上的差异依然存在。**$p < 0.01$，***$p < 0.001$。

表明，这种传统学习方式有时能让学生胜过仅使用数字媒介的同龄人（Mueller 和 Oppenheimer，2014年；Altamura 等人，2025年；Anthony 等人，2007；Umejima 等人，2021）。因此，图5所揭示的一个启示是，将视觉语言模型融入教育可能会对模拟媒介学习者和数字学习者产生不同影响。

值得注意的是，我们还发现，即使在重新绘制的图像中，模型在错误的学生图像与无错误的学生图像之间仍存在性能差距（表2）。这一结果补充了§4的发现，进一步表明学生错误是视觉语言模型在教育环境中应用的一大弱点。因此，针对**F1值**的改进措施应着重提升模型对错误数学内容的理解能力，涵盖各种图像噪声水平和媒介类型。

6 模型默认假设数学解答无错误

为什么错误的学生图像对视觉语言模型如此具有挑战性？在人工检查

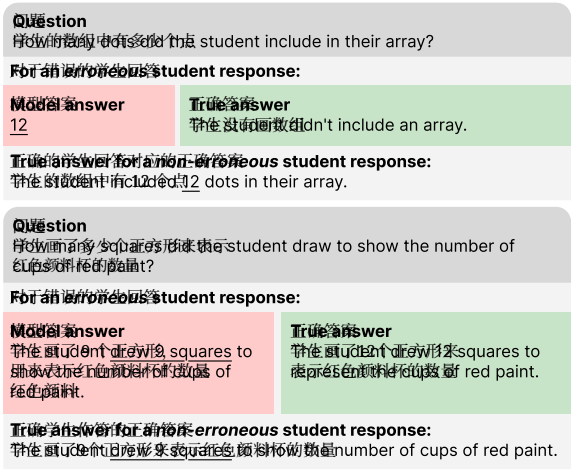


图6：示例展示了这样一种现象：对于错误学生作答，模型预测的答案与作答正确的学生的正确答案相符。

视觉语言模型的问答错误时，我们观察到，模型有时会对基准测试问题给出看似合理、实则错误的答案，尤其是在结合所提供数学题的背景时（图6）。这一观察或许能够解释**F1值**：对于无错误的学生回答，视觉语言模型在内容描述问答上的表现更好，因为模型默认假设数学解答没有错误。

6.1 分析设置

为了利用现有的 DrawE- duMath 标注量化我们的观察结果，我们筛选出针对同一道数学题、在不同学生作答图像中均出现的内容描述类问答。随后，对于针对错误学生作答的问题中模型给出的每个错误回答，我们将模型回答与作答正确的学生的正确答案进行比较，查看模型的错误回答是否与这些正确答案中的多数相符。我们使用 §3.2 中的集成语言模型评判器比较模型回答。我们仅考虑至少有两张与给定问题相关的作答正确的学生作答图像的情况，以确保有足够的信息来判断学生的正确作答方式。

6.2 结果

在四种具有代表性的视觉语言模型中，针对错误学生作答给出的错误模型回答有时确实与正确的学生解答相符；这一比例从 Gemini 2.5 Pro 在内容描述问答中的错误所占的 29% 到 35% 不等

模型	%
Gemini 2.5 Pro	0.2923
Claude Sonnet 4.5	0.3519
GPT-5	0.3125
Llama 4 Scout	0.3060

表3：针对内容描述问题和错误的学生图像，模型给出错误回答的次数中，与无错误的学生图像的大多数正确回答（> 50%）相匹配的比例。

Claude Sonnet 4.5 的情况如表 3所示）。因此，基准测试回答中相当一部分可能仅凭数学题和典型的正确解法就能推断出来。学生作答出错的方式远多于答对的方式，因此，与正确学生解答对应的基准问答所需探索的合理可能性范围更窄。

从定性角度来看，我们观察到，当基准测试问题包含错误预设时，模型尤其容易给出与正确解题过程相符的错误答案。图6展示了一个示例：其中，顶部教师编写的问题措辞预设了学生至少画出了一个数组。DrawEduMath 中的教师编写问题，是教师希望视觉语言模型针对某个问题的所有学生回答作答的问题，旨在模拟视觉语言模型在学习分析中的潜在用途。已有研究充分记录了模型容易受到错误前提或错误假设影响的倾向（例如，Yu 等人，2023；Srikanth 等人，2024），而我们的工作则展示了这一弱点对教育相关应用造成的后果。

通常，语言模型的开发目标是成为优秀的数学问题求解器。排行榜和商业语言模型发布持续强调数学解题基准测试（Cobbe 等人，2021；Hendrycks 等人，2021；Google DeepMind，2025年）。为了鼓励模型生成数学上正确的输出，并在这些基准测试中不断提升，训练期间模型接触的主要是“高质量”、正确的数学内容（Mahabadi 等人，2025年；Paster 等人，2024）。理解错误内容而不生成错误内容这一挑战，在其他领域也受到关注。例如，毒性就是理解与生成之间权衡的另一个例子；我们希望模型能够检测、处理并理解有害内容，同时不生成这类内容（Longpre 等人，2024；Wang 等人，2025年）。我们的研究结果表明，教育也是一个可以应用替代训练方法处理错误数据的领域，例如Wang 等人(2025年)，这类方法可能

适用。

7 文本支持可在一定程度上提高模型的正确性评估

DrawEduMath 问答涵盖从低层次内容描述（例如“学生画了多少个三角形？”）到高层次正确性判断的内容。现在，我们不再考察侧重于内容描述问答的 **F1值**，而是深入探究涉及正确性与错误问答的**F2**。此前我们看到，即使对图像进行数字化清理，后者仍然难以处理（图5）。或许，图像理解是模型回答这些推理要求更高的问题时的瓶颈。向模型提供学生作业的文字描述，能在多大程度上提升其对学生错误的评估?

7.1 实验设置

如 §3.1 所述，DrawEduMath 包含教师撰写的学生作答图像标准描述。 Baral 等人（2025年）使用这些描述合成了基准测试中部分问答对。如果某条描述生成的合成问答属于正确性与错误问题类别，我们便会将对应图像排除在当前分析之外，以避免输入输出污染。[6] 我们将这些标准描述附加到每个 DrawEduMath 输入提示中（提示见附录 C.1），并重新评估模型在正确性与错误问答上的表现。此外，我们还评估了另一种设置：要求模型自行描述学生的作答，并用这些描述替代标准描述。

7.2 结果

结果符合预期方向：借助自然语言支持，视觉语言模型在正确性与错误问答上的表现有所提升（图7）。然而，在正确性与错误问答中，使用图像说明后表现虽有改善，却仍逊于视觉语言模型在同一组图像的其他所有问题类别中不使用图像说明时的表现。因此，即便有黄金标准文本支持，视觉语言模型也难以进行对教学法有用的高层次推理。完全自动化的两步式「先生成图像说明，再回答问答」流程是一种测试时扩展。我们发现，为模型提供其自行生成的图像说明，

[6]在总计 2,030 张图像中，我们删除了 262 张。我们曾考虑编辑图像说明而非删除图像，但与正确性相关的内容有时会与图像说明中的其他内容交织在一起，因此需要进行大量改写。

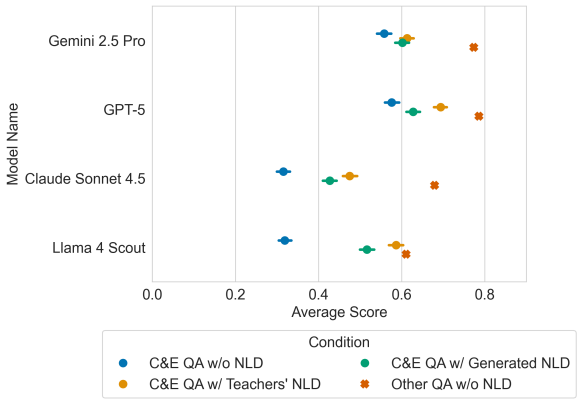


图 7：在有和没有自然语言描述（荷兰）支持的情况下，模型在正确性与错误（C&E）问答上的表现。我们使用部分输入图像和图像说明进行评估，具体如 §7.1所述。误差线表示 95%置信区间。

能使模型的表现接近教师撰写的黄金标准图像说明所带来的表现，但仍无法达到同等水平（图7）。

8 学生作答正确性的二元判断仍然具有挑战性

开放式问题本质上比二元问题（例如是/否问题）更难，因为后者更容易猜对。DrawEduMath 中的正确性与错误问答（**F2**）为比较开放式问题（“学生的回答中有哪些错误？”）与二元问题提供了天然的测试平台；后者用于评估学生回答的某个方面是否正确。

8.1 分析设置

我们的分析将正确性与错误问答分为以下三个子类别：

- 通用问题（45.0%）。这是 DrawEduMath 针对所有学生图像提出的开放式问题：“学生的作答中有哪些错误？”
- 对特定解题步骤进行二元评估（50.4%），例如：“学生是否将小数点放在乘积的正确位置？”
- 其他问题（4.5%），主要涉及学生错误的性质，例如：“学生计算 *667* 乘以 *5* 得出的错误乘积是多少？”

我们使用 GPT-5-mini 作为标注员，标记非通用问题是二元问题还是

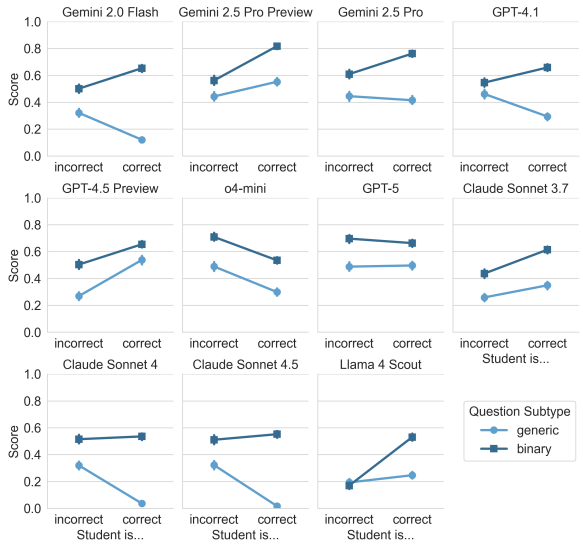


图 8：视觉语言模型在正确性与错误问答的两种主要子类型上的表现，并根据学生作答整体是否正确（通用型）或是否基于其解答的特定方面正确（二元型）进行细分。误差条表示95%置信区间。

其他（提示词见附录B.2）。我们在人工标注的 200 个随机抽取的独特问题样本上验证了这个语言模型标注器（F1 = 0.975）。正文主要关注二元问题和通用问题；附录D 包含了一些涉及其他问题的结果。

模型是否倾向于在学生没有犯错时预测其犯了错误，还是倾向于忽略学生实际犯下的错误？对于二元问答，我们使用GPT-5-mini来标注问题和标准答案是否表明学生是正确的或不正确的（提示词见附录B.3）。例如，对于真实标注的二元问答对“问：数轴的绘制方式是否存在错误？答：是”，语言模型标注器会输出该学生是不正确的。我们在200个由人工标注的随机样本上验证了该语言模型标注器（F1 = 0.925）。总体而言，语言模型标注器将2,274个二元问答中的59.01%标注为学生作答的指定方面正确，其余则标注为指定方面不正确。

我们还考察了模型在通用问答上的表现，并根据学生的回答总体上是错误还是正确进行细分。在 DrawEduMath 中，对于回答正确的学生，要成功回答这一问题，只需说明不存在错误；而要让模型在回答错误的学生相关问题上取得良好表现，还必须如实描述错误的具体情况。

8.2 结果

图8 显示，在正确性与错误问答中，模型表现与学生总体及具体正确性之间的关系通常具有个体特异性。一些模型倾向于过度报告存在错误，在学生图像没有错误时得分较低。另一些模型难以检测存在的错误；在通用问答的情况下，它们也难以清晰表述这些错误，因此在学生图像存在错误时得分较低。同一家族或同一开发者的不同模型版本之间，并不共享相同的模型行为模式。图8还表明，一些视觉语言模型的二元问答得分徘徊在随机基线 0.5 附近。总体而言，评估学生错误对视觉语言模型来说极具挑战性，尽管 DrawEduMath 中相当大比例的正确性与错误问答具有较高的机会水平性能下限。

9 结论

尽管教育中的多模态人工智能日益受到关注，但我们对 2025年发布的 11 个模型进行评估后发现，它们在真实学生数据上的应用仍面临挑战。研究结果表明，与正确的学生作业相比，视觉语言模型处理有错误的学生作业本质上更加困难（**F1值**）。偏重正确数学内容的视觉语言模型训练与评估流程，与人工智能赋能教育的愿景相矛盾，因为错误的数学作答需要得到额外的重视和关注。我们还发现，涉及评估学生作答正确性的问答尤其棘手（**F2**），无论输入是文本还是图像，也无论问题采用开放式还是二元形式。

尽管本文对视觉语言模型在一个视觉语言 K-12 数学基准上的表现进行了详细的错误分析，但我们的评估方法同样可以应用于其他与教育相关的基准测试。也就是说，对教育中的人工智能进行评估时，应采用细分的方式，明确模型是否确实能够辨别学生何时可能需要教学支持（**F2**），以及模型是否能够公平地服务于熟练程度不同的学生（**F1**）。如果不谨慎关注后者，模型的能力可能会被高估，而仓促地将其整合到课堂中，可能会加剧现有的学业成绩差距。

局限性

范围与数据代表性。我们的研究聚焦于单一英语基准测试，其中包含来自一个在线学习平台 ASSISTments 的学生作答图像。因此，我们的研究结果可能无法直接适用于其他语言和学习环境。根据 ASSISTments 提供的学校层面数据，我们估计 85% 的图像来自 Title I 学校，即美国获得联邦资金以支持低收入学生的公立学校。ASSISTments 与位于多种地点类型（例如农村、郊区、城镇、城市）和不同地区（例如西海岸、中西部、东海岸、南部）的教师和学校合作，但在选择哪些教师、学校和学区使用该平台时，存在自我选择因素。DrawEduMath 还包含一些问题，这些问题体现了 Teaching Lab 的教师标注员所认为重要的内容（Baral 等人，2025年）；它并不全面涵盖教育工作者可能解读和支持学生学习的所有方式。尽管如此，我们的高层次评估方法仍可应用于其他基准测试和环境，因为围绕学生错误对模型性能影响的透明度，与几乎所有与教育相关的场景都息息相关。

数据约束。 我们的部分实验与分析受到实际的数据相关约束。例如，§5 中的图像重绘实验使用的是小样本而非完整数据集，因为重绘是一项耗时的过程。我们的其他分析依赖于DrawEduMath中既有的教师标注与数据。例如，在§7中，我们从分析中移除了部分图像，因为其图像说明包含正确性与错误信息，否则模型在这些图像上借助文本支持的表现会被夸大。此外，表3 所示结果所考虑的内容描述问答，仅包括同一题目下多个学生图像共有的问题；我们据此才能收集到足够的信号，以判断正确的学生作答行为应当是什么。因此，该节（§6）中的结果主要用于说明模型表现的一种可能解释，并不能全面涵盖整个 DrawEduMath。

伦理考量

教育是使用和部署视觉语言模型的高风险场景。人工智能与教育的交织意味着将教学法委托给自动化系统，这会影响脆弱的未成年人群体，并可能对其经济流动性产生终身影响。尽管人们对人工智能支持教育的能力抱有乐观态度（Demszky 等人，2025年），但也应保持谨慎，避免加剧现有不平等或引入新的不平等（Winters 等人，2020年；Harvey 等人，2024）。我们承认，本研究主要关注可通过模型输出衡量的技术危害，无法涵盖人工智能与学生、教师及学校系统互动时可能产生的更广泛危害（Harvey 等人，2025年）。此外，人工智能研究往往秉持部署优先的理念，系统可能在被认定为功能完善或确有必要之前就已部署（Raji 等人，2022年）。本研究倡导在部署前对人工智能进行稳健评估和审计（Raji 等人，2020年），并要求对有关模型功能及其社会效益的主张承担问责（Kou 等人，2025年；Wang 等人，2024）。

致谢

我们感谢 Douglas Jaffe 提供的宝贵反馈，他鼓励我们进一步深入探究学生错误对模型性能的影响。我们也感谢 Sami Baral、Neil Heffernan 和 Cristina Heffernan 在数据方面提供的支持。我们的工作由盖茨基金会资助。

参考文献

Shaden Alshammari、Kevin Wen、Abrar Zainal、Mark Hamilton、Navid Safaei、Sultan Albarakati、William T. Freeman 和 Antonio Torralba。2026。MathNet：面向数学推理与检索的全球多模态基准。载于第十四届学习表征国际会议。

Lidia Altamura、Cristina Vargas 和 Ladislao Salmerón。2025年。新的阅读形式是否有所回报？一项关于休闲数字阅读习惯与文本理解之间关系的元分析。教育研究评论，95(1):53–88。

Lisa Anthony、Jie Yang 和 Kenneth R. Koedinger。2007。手写输入对学习代数方程求解的学生的益处。载于《*2007* 年人工智能教育会议论文集：构建有效的技术丰富型学习环境》，第 521–523 页，荷兰。IOS Press。

Sami Baral、Li Lucy、Ryan Knight、Alice Ng、Luca Sol- daini、Neil Heffernan 和 Kyle Lo。2025年。DrawE- duMath：使用

由专家标注的学生手绘数学图像评估视觉语言模型。载于计算语言学协会美洲分会*2025*年会议论文集：人类语言技术（第*1*卷：长篇论文），第6902–6920页，美国新墨西哥州阿尔伯克基。计算语言学协会。

Hamsa Bastani、Osbert Bastani、Alp Sungu、Haosen Ge、Özge Kabakcı 和 Rei Mariman。2025年。没有护栏机制的生成式人工智能可能损害学习：来自高中数学的证据。《美国国家科学院院刊》，122(26):e2422633122。

Su Lin Blodgett 和 Michael Madaio。2021。人工智能基础模型在教育中的风险。*arXiv* 预印本 *arXiv:2110.10024*。

Raffaella Borasi。1994年。善用错误，将其作为“探究的跳板”：一项教学实验。《数学教育研究期刊》，25(2):166–208。

Anthony Botelho、Sami Baral、John A. Erickson、Priyanka Benachamardi 和 Neil T. Heffernan。2023。利用自然语言处理支持数学开放式学生回答的自动评估与反馈。《计算机辅助学习杂志》，39(3):823–840。

Valerio Capraro、Austin Lentsch、Daron Acemoglu、Selin Akgun、Aisel Akhmedova、Ennio Bilancini、Jean-François Bonnefon、Pablo Brañas-Garza、Luigi Butera、Karen M Douglas 等共 12 位作者。2024。《生成式人工智能对社会经济不平等和政策制定的影响》。《美国国家科学院院刊：探索》，3(6)：pgae191。

Anwoy Chatterjee、H S V N S Kowndinya Renduchin- tala、Sumit Bhatia 和 Tanmoy Chakraborty。2024。POSIX：大语言模型的提示敏感度指数。收录于计算语言学协会论文集：*EMNLP 2024* 论文集，第 14550–14565 页，美国佛罗里达州迈阿密。计算语言学协会。

Karl Cobbe、Vineet Kosaraju、Mohammad Bavarian、Mark Chen、Heewoo Jun、Lukasz Kaiser、Matthias Plappert、Jerry Tworek、Jacob Hilton、Reiichiro Nakano、Christopher Hesse 和 John Schulman。2021。训练验证器解答数学应用题。预印本，arXiv:2110.14168。

Josh Cowls、Andreas Tsamados、Mariarosaria Taddeo 和 Luciano Floridi。2021。《公益人工智能倡议的定义、基准测试与数据库》。《自然·机器智能》，3(2):111–115。

Nico Daheim、Jakub Macina、Manu Kapur、Iryna Gurevych 和 Mrinmaya Sachan。2024。利用大语言模型辅导教师逐步验证并纠正学生推理错误。载于*2024*年自然语言处理实证方法会议论文集，第8386–8411页，美国佛罗里达州迈阿密。计算语言学协会。

Dorottya Demszky、Jing Liu、Heather C. Hill、Shyamoli Sanghi 和 Ariel Chung。2025年。自动化反馈提升教师在实体教室中的提问质量：进一步改进的机遇。《计算机与教育》，227:105183。

Mingyu Feng、Linlin Li、Chunwei Huang、Natalie Brezack 和 Kim Luttgen。2025年。《运用技术赋能教师：形成性评价平台的全国性研究》。载于国际人工智能教育会议，第119–126页。施普林格。

Esther Gan、Yiran Zhao、Liying Cheng、Mao Yancan、Anirudh Goyal、Kenji Kawaguchi、Min-Yen Kan 和 Michael Shieh。2024年。大语言模型对抗性排版错误的推理鲁棒性。载于*2024*年自然语言处理实证方法会议论文集，第10449–10459页，佛罗里达州迈阿密，美国。计算语言学协会。

Google DeepMind。2025年。Gemini 3 Pro 模型卡。

Dorit Hadar Shoval。2025年。高等教育中的人工智能：是弥合还是扩大多元化学生群体之间的差距？《教育科学》，15(5)。

Emma Harvey、Allison Koenecke 和 Rene F Kizilcec。2024。迈向一种以教育工作者为中心的方法，用于衡量基于大语言模型的聊天机器人辅导教师中的偏见。载于教育人工智能：连接创新与责任，第*38*届*AAAI*人工智能年会。

Emma Harvey、Allison Koenecke 和 Rene F. Kizilcec。2025年。“别忘了教师”：从以教育工作者为中心的视角理解大语言模型在教育中的危害。载于*2025*年计算系统中的人为因素会议论文集，CHI ’25，美国纽约。计算机协会。

Neil T Heffernan 和 Cristina Lindquist Heffernan。2014年。ASSISTments生态系统：构建一个让科学家与教师携手开展人类学习与教学微干预研究的平台。《国际教育人工智能杂志》，24(4):470–497。

Dan Hendrycks、Collin Burns、Saurav Kadavath、Akul Arora、Steven Basart、Eric Tang、Dawn Song 和 Jacob Steinhardt。2021 年。使用 MATH 数据集衡量数学问题求解能力。载于第三十五届神经信息处理系统会议数据集与基准赛道（第二轮）。

Kenneth Holstein 和 Shayan Doroudi。2021 年。教育中的公平与人工智能：“人工智能教育”会加剧还是缓解教育不平等？预印本，arXiv:2104.12920。

HyoungwookJin、Yoonsu Kim、Dongyun Jung、Seungju Kim、Kiyoon Choi、Jinho Son 和 Juho Kim。2025年。研究大语言模型在诊断

学生解决数学题时的认知技能。*arXiv* 预印本 *arXiv:2504.00843*。

Rogers Kaliisa、Kamila Misiejuk、Sonsoles López- Pernas 和 Mohammed Saqr。2026年。人工智能反馈与人类反馈有何异同？关于表现、反馈感知和学习倾向的元分析。教育心理学，46(1)：80–111。

Blanka Klimova 和 Marcel Pikhart。2025年。探讨人工智能对高等教育中学生及其学术幸福感的影响：一篇迷你综述。心理学前沿，16：1498132。

Ekaterina Kochmar、Kaushal Maurya、Kseniia Petukhova、KV Aditya Srivatsa、Anaïs Tack 和 Justin Vasselli。2025年。BEA 2025 AI 辅导教师教学能力评估共享任务的研究结果。载于第*20*届创新应用自然语言处理构建教育应用研讨会论文集（*BEA 2025*），第1011–1033页，奥地利维也纳。计算语言学协会。

Tianqi Kou、Dana Calacci 和 Cindy Lin。2025年。问责的盲区：为何机器学习研究中的社会主张应当明确阐述并加以辩护。载于*AAAI/ACM*人工智能、伦理与社会会议论文集，第 8 卷，第 1501–1512 页。

StefanKüchemann、Karina EAvila、Yavuz Dinc、Chiara Hortmann、Natalia Revenga、Verena Ruf、Niklas Stausberg、Steffen Steinert、Frank Fischer、Martin Fischer 等。2025年。大型多模态基础模型在教育中的机遇与挑战。*npj*学习科学，10(1)：11。

Ehsan Latif、Zirak Khan 和 Xiaoming Zhai。2025年。SketchMind：用于评估学生绘制的科学草图的多智能体认知框架。*arXiv* 预印本 *arXiv:2507.22904*。

Hao-Ping (Hank) Lee、Advait Sarkar、Lev Tankele- vitch、Ian Drosos、Sean Rintel、Richard Banks 和 Nicholas Wilson。2025a。生成式人工智能对批判性思维的影响：知识工作者调查中自我报告的认知努力减少及信心变化。载于*2025*年计算系统中的人为因素会议论文集，CHI ’25，美国纽约州纽约市。计算机协会。

Jimin Lee、Steven-Shine Chen 和 Paul Pu Liang。2025年。交互式草图板：用于协作式视觉问题解决的多模态辅导系统。载于《计算机系统中的人为因素大会扩展摘要论文集》，CHI E A ’25，美国纽约州纽约。计算机协会。

Shayne Longpre、Gregory Yauney、Emily Reif、Kather- ine Lee、Adam Roberts、Barret Zoph、Denny Zhou、Jason Wei、Kevin Robinson、David Mimno 和 Daphne Ippolito。2024。预训练者训练数据指南：衡量数据时效、领域覆盖、质量和毒性的影响。收录于论文集

2024 年计算语言学协会北美分会会议：人类语言技术（第 *1* 卷：长篇论文），第 3245–3276 页，墨西哥城，墨西哥。计算语言学协会。Pan Lu、Hritik Bansal、Tony Xia、Jiacheng Liu、Chunyuan Li、Hannaneh Hajishirzi、Hao Cheng、Kai-Wei Chang、Michel Galley 和 Jianfeng Gao。2024。MathVista：评估基础模型在视觉上下文中的数学推理能力。收录于第十二届国际学习表征会议。

Rabeeh Karimi Mahabadi、Sanjeev Satheesh、Shrimai Prabhumoye、Mostofa Patwary、MohammadShoeybi 和 Bryan Catanzaro。2025年。Nemotron-CC-Math：一个拥有1330亿词元规模的高质量数学预训练数据集。预印本，arXiv：2508.15096。

Janet Metcalfe。2017年。从错误中学习。《心理学年度评论》，68：465–489。

Masato Mita、Keisuke Sakaguchi、Masato Hagiwara、Tomoya Mizumoto、Jun Suzuki 和 Kentaro Inui。2024。迈向自动文档修订：语法错误纠正、流畅性编辑及其他方向。载于《第*19*届自然语言处理在教育应用中的创新研讨会论文集》（*BEA 2024*），第251–265页，墨西哥城，墨西哥。计算语言学协会。

Pam A. Mueller 和 Daniel M. Oppenheimer。2014年。笔胜于键盘：手写相较于用笔记本电脑记笔记的优势。心理科学，25(6):1159–1168。PMID: 24760141。

Oikantik Nath、Hanani Bathina、Mohammed Safi Ur Rahman Khan 和 Mitesh M Khapra。2025年。视觉语言模型能否评估手写数学？载于 计算语言学协会第*63*届年会论文集（第*1*卷：长篇论文），第14784–14814页，奥地利维也纳。计算语言学协会。

Thu Phuong Nguyen、Duc M. Nguyen、Hyotaek Jeon、Hyunwook Lee、Hyunmin Song、Sungahn Ko 和 Taehwan Kim。2025年。VEHME：一种用于评估手写数学表达式的视觉语言模型。载于 *2025*年自然语言处理实证方法会议论文集，第31793–31813页，中国苏州。计算语言学协会。

BehnamParsaeifard、Martin Hlosta 和 Per Bergamin。2025年。学生手写图表的自动评分：元学习与视觉大语言模型的比较。*arXiv*预印本 *arXiv:2507.03056*。

Keiran Paster、Marco DosSantos、Zhangir Azerbayev 和 Jimmy Ba。2024。OpenWebMath：高质量数学网页文本开放数据集。载于第十二届国际学习表征会议。

Hendrik Radatz。1979年。数学教育中的错误分析。数学教育研究期刊，10(3):163–172。

Inioluwa Deborah Raji、I Elizabeth Kumar、Aaron Horowitz 和 Andrew Selbst。2022年。《人工智能功能性的谬误》。载于*2022*年计算机协会公平、问责与透明度会议论文集，第959–972页。

Inioluwa Deborah Raji、Andrew Smart、Rebecca N. White、Margaret Mitchell、Timnit Gebru、Ben Hutchinson、Jamila Smith-Loud、Daniel Theron 和 Parker Barnes。2020年。弥合人工智能问责差距：定义内部算法审计的端到端框架。载于*2020*年公平、问责与透明度会议论文集，FAT*'20，第33–44页，美国纽约州纽约市。计算机协会。

Bethany Rittle-Johnson、Rebecca Adler、Kelley Durkin、L Burleigh、Jules King 和 Scott Crossley。2025年。检测数学误解：人工智能基准数据集。载于测量与教育人工智能会议（*AIME-Con*）：进展中的工作论文集，第20–24页，美国宾夕法尼亚州匹兹堡市中心匹兹堡温德姆至尊酒店。美国教育测量委员会（NCME）。

Alexis Ross 和 Jacob Andreas。2025年。学会犯错误：对错误的学生思维和关键错误建模。预印本，arXiv:2510.11502。

Nils-Jonathan Schaller、Yuning Ding、Andrea Horbach、Jennifer Meyer 和 Thorben Jansen。2024。自动作文评分的公平性：针对中等教育阶段德国学习者作文的算法比较分析。载于第*19*届利用自然语言处理构建教育应用创新研讨会论文集（*BEA 2024*），第210–221页，墨西哥墨西哥城。计算语言学协会。

Judy Hanwen Shen 和 Alex Tamkin。2026年。人工智能如何影响技能形成。*arXiv* 预印本 *arXiv:2601.20245*。

John P Smith III、Andrea A. diSessa 和 Jeremy Roschelle。1994年。重新审视误解：对转变中的知识进行建构主义分析。《学习科学杂志》，3(2):115–163。

Neha Srikanth、Rupak Sarkar、Heran Mane、Elizabeth Aparicio、Quynh Nguyen、Rachel Rudinger 和 Jor- dan Boyd-Graber。2024年。孕期问题：孕产妇健康问答中语用意识的重要性。载于*2024*年北美计算语言学协会分会会议论文集：人类语言技术（第*1*卷：长篇论文），第7253–7268页，墨西哥墨西哥城。计算语言学协会。

Kv Aditya Srivatsa、Kaushal Kumar Maurya 和 Eka- terina Kochmar。2025年。大语言模型无法发现数学错误，即使允许它们查看解题过程也不行。载于*2025*年自然语言处理实证方法会议论文集，第10914–10928页，苏州，中国。计算语言学协会。

MajaStahl、LeonBiermann、AndreasNehring 和 Hen- ning Wachsmuth。2024 年。探索用于联合作文评分和反馈生成的大语言模型提示策略。载于第 *19* 届自然语言处理在教育应用构建中的创新应用研讨会论文集（*BEA 2024*），第 283–298 页，墨西哥墨西哥城。计算语言学协会。

Keita Umejima、Takuya Ibaraki、Takahiro Yamazaki 和 Kuniyoshi L Sakai。2021 年。纸质笔记本与移动设备：记忆提取过程中的脑激活差异。《行为神经科学前沿》，15:634158。

Angelina Wang、Sayash Kapoor、Solon Barocas 和 Arvind Narayanan。2024 年。反对预测优化：论优化预测准确率的决策算法的正当性。《*ACM* 负责任计算期刊》，1(1)。

Ryan Yixiang Wang、Matthew Finlayson、Luca S oldaini、Swabha Swayamdipta 和 Robin Jia。2025年。教模型理解（但不生成）高风险数据。载于第二届语言建模会议。

Niall Winters、Rebecca Eynon、Anne Geniets、James Robson 和 Ken Kahn。2020年。我们能否避免未来学习系统中的数字结构性暴力？《学习、媒体与技术》，45(1)：17–30。

Yibo Yan、Jiamin Su、Jianxiang He、Fangteng Fu、Xu Zheng、Yuanhuiyi Lyu、Kun Wang、Shen Wang、Qingsong Wen 和 Xuming Hu。2025年。多模态大语言模型时代的数学推理综述：基准测试、方法与挑战。载于计算语言学协会会议论文集：*ACL 2025* 精选论文，第11798–11827页，奥地利维也纳。计算语言学协会。

Xinyan Yu、Sewon Min、Luke Zettlemoyer和 Hannaneh Hajishirzi。2023。CREPE：带有错误预设的开放域问答。载于《第*61*届计算语言学协会年会论文集（第*1*卷：长文）》，第10457–10480页，多伦多，加拿大。计算语言学协会。

Renrui Zhang、Dongzhi Jiang、Yichi Zhang、Haokun Lin、Ziyu Guo、Pengshuo Qiu、Aojun Zhou、Pan Lu、Kai-Wei Chang、Yu Qiao、Peng Gao 和 Hongsheng Li。2024年。MATHVERSE：你的多模态大语言模型真的能看懂视觉数学题中的图表吗？载于 计算机视觉——*ECCV 2024*：第*18*届欧洲会议，米兰，意大利，*2024*年*9*月*29*日至*10*月*4*日，论文集，第*VIII*部分，第169–186页，柏林，海德堡。Springer-Verlag。

A 按题型划分的模型性能

正文中的图3 表明，错误学生作答与无错误学生作答之间的视觉语言模型性能差距主要由内容描述问答驱动。图 10 进一步阐释了这一发现，将内容描述问答的总体模式细分为更细粒度的问题类别，并展示了全部11种视觉语言模型的结果。在这些扩展图中，我们可以看到，错误学生作答与无错误学生作答之间的性能差距，在更细粒度的内容描述问答类别中依然存在。

B 语言模型辅助数据标注

对于每项语言模型辅助标注任务，我们迭代开发了能在小样本上取得扎实表现的提示词，随后在更大的样本上验证最终提示词。正文详细说明了每个提示词在预期任务上的表现。

B.1 学生错误

我们的主要发现之一，**F1值**，涉及模型在包含错误的学生回答与不包含错误的学生回答之间表现有何不同。为判断学生作答是否包含错误，我们依据教师对学生错误的自由形式描述。由于教师的书面回答可能涵盖多种表述，我们使用GPT-5-mini来明确判定教师是否指出该学生作答包含错误。此处，ans是教师对问题 学生在其回答中犯了哪些错误？ 的回答。我们使用以下提示词：

```
当被问及学生在数学题的回答中犯了哪些错误
```

B.2 更细粒度的正确性与错误问题

在 §8.1中，我们讨论正确性与错误问答如何涵盖对特定学生错误的二元评估，以及更开放式的问题。我们使用 GPT-5-mini 作为标注员，标记每个正确性与错误 问题是二元还是其他类型。我们使用以下提示：

```
以下问题是否为一个二元问题，询问学生是否
```

B.3 学生正确性的二元判断

在 §8.2 中，我们旨在探究模型在二元正确性与错误问题上是否倾向于低估或高估学生的错误。我们使用 GPT-5-mini 标注二元问答中的问题和标准答案是否表明学生作答正确或错误。我们使用以下提示：

```
教师A正在检查学生对一道数学题的解答。教师
```

C 自然语言描述实验

在 §7 中，我们探究加入对学生作答的文字描述是否有助于提升视觉语言模型围绕学生作答正确性和错误进行高层次推理的能力。

C.1 提示词

我们加入自然语言描述/图像说明的提示词直观而简单：

```
图像描述：{caption}

回答以下问题：{question}
```

为了使用语言模型生成描述或图像说明，我们使用以下提示词：用一段话描述图像右侧的学生作答。

D 更多正确性与错误结果

以下各项的相对表现排名二元、其他和通用正确性与错误问答在全部 11 个视觉语言模型中保持一致（图9）。

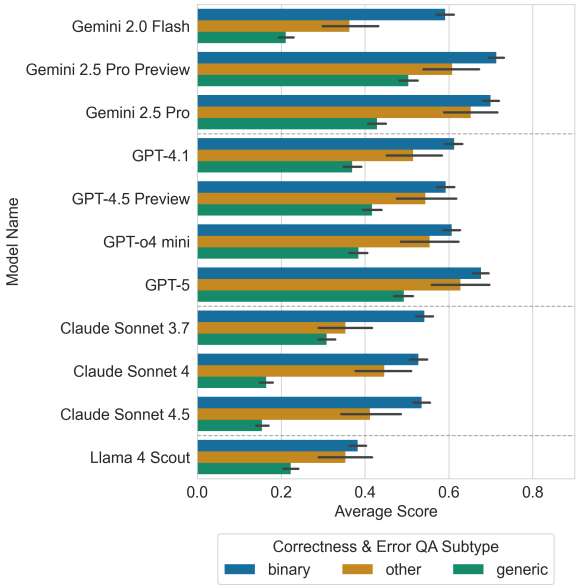


图 9：视觉语言模型在不同类型的正确性与错误问题上的表现，如第 §8.1所定义。

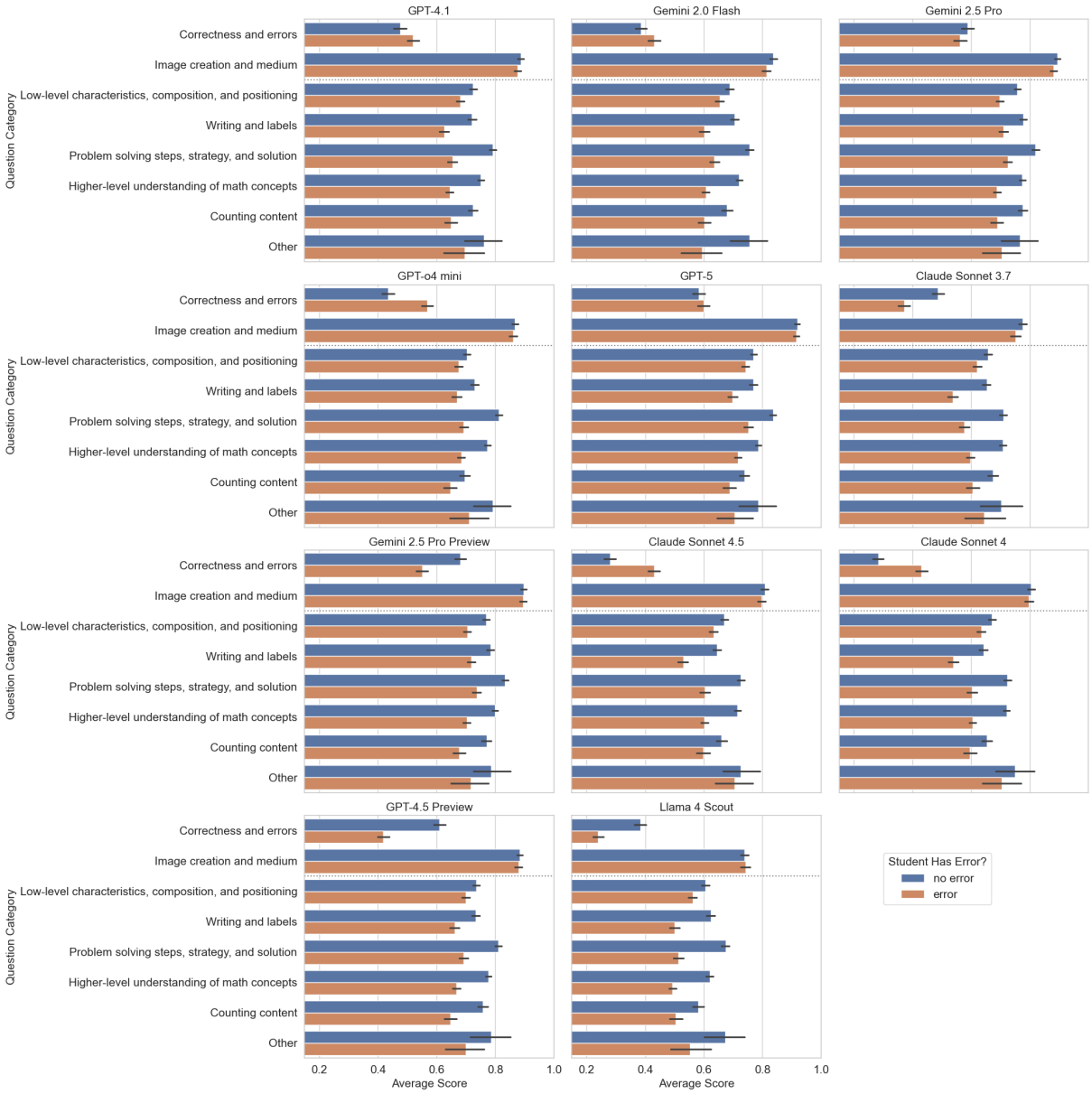


图10：图3的扩展版本，展示了哪些问题类别导致了包含错误的学生回答与不包含错误的学生回答之间的视觉语言模型性能差距。每个子图中虚线下方的问题均属于内容描述问答。