

意图重要性：通过 细粒度教学意图注释增强人工智能辅导

Kseniia Petukhova, Ekaterina Kochmar

Mohamed bin Zayed University of Artificial Intelligence

{ kseniia.petukhova, ekaterina.kochmar } @mbzuai.ac.ae

Abstract

大型语言模型 (LLM) 在教育应用中，尤其是在智能辅导系统中具有巨大的潜力。然而，有效的辅导需要与教学策略相一致——这点当前的 LLM 在没有特定任务适应的情况下是所欠缺的。在这项工作中，我们探讨了教师意图的精细标注是否可以提高 LLM 生成的辅导回应的质量。我们关注 MathDial，一个数学教学对话数据集，并应用一个自动化标注框架使用详细的十一类教学意图分类法重新标注数据集的一部分。然后我们使用这些新的标注来微调一个 LLM，并将其性能与基于原始四分类法训练的模型进行比较。自动和定性评估均表明，精细模型生成了更符合教学目标且更有效的回应。我们的研究结果强调了意图的具体性在教育设置中的受控文本生成的重要性，我们发布了我们的标注数据和代码以促进进一步的研究：<https://github.com/Kpetykhova/autoTree/tree/main/mathdial>

1 介绍

人类导师是教育发展的基石，在增强学习者和促进社会进步方面发挥着重要作用。一对一辅导长期以来被认为是非常有效的 (Bloom, 1984)；然而，由于合格导师的有限可用性，其广泛实施受到限制。近年来，在教育背景中，LLMs 显示出了极大的潜力 (Wang et al., 2024; Gan et al., 2023)，这导致了基于 LLM 的智能辅导系统 (ITS) 的出现 (Pal Chowdhury et al., 2024; Liu et al., 2024) 以及通过高级提示策略使用 LLMs 作为导师 (Denny et al., 2024; Mollick and Mollick, 2024)。这些 AI 导师服务于一系列教育目标 (Wollny et al., 2021)，其中最突出的是纠正学生错误和困惑的补救——这一领域继续推动 AI 辅导系统的发展 (Macina et al., 2023; Wang et al., 2023)。

虽然大型语言模型在生成类似人类的对话和处理各种推理任务（例如常识推理和基础数学推理 (Achiam et al., 2023; Kojima et al., 2022; Laskar et al., 2023; Yang et al., 2024)）方面表

现良好，但它们不能在未经重大调整的情况下直接应用于教育系统。有效的辅导不仅仅需要流利的对话——更重要的是引导学习者自主发现答案。好的导师不是简单地提供解决方案，而是使用策略，例如给予提示、在苏格拉底式对话中提问 (Carey and Mullan, 2004)，并鼓励积极地解决问题。因此，基于大型语言模型的导师理想情况下应与人类的辅导策略 (Nye et al., 2014) 和已被证明可提高学生成果的主动学习实践 (Freeman et al., 2014) 保持一致。

为了拥有这样的模型，我们需要对话辅导数据集。MathDial (Macina et al., 2023) 就是这样一个数据集，包含围绕数学推理任务的导师-学生对话。每个教师的发言都被标记为 Macina et al. (2023) 中的四种教学动作类型之一：焦点（指导任务进度）、测试（鼓励概念探索）、叙述（在学生遇到困难时提供帮助），或通用（非教学对话回合）。这些注释是在数据收集期间由教师提供的，以更好地支撑学生的学习。虽然这种四类别的分类法提供了有用的高层结构，但它缺乏高级应用如控制响应生成、教学分析和 AI 导师中的行为建模所需的细粒度细节。同时，更细粒度的注释可能能够提升可解释性、改善教学一致性，并在指导学生学习体验方面提供更大的灵活性。

尽管 MathDial 的原始分类法仅包括四个广泛的类别，作者也提供了一组扩展的十一种细粒度意图，这可能为 AI 生成的辅导响应提供更大的控制和多样性。在此基础上，在这项工作中，我们应用了一个完全自动化的会话话语标注框架 (Petukhova and Kochmar, 2025) 到使用更细化的十一意图分类法对 MathDial 数据集的一部分进行重新注释。该标注框架使用 LLMs 自动从分类法构建决策树并用其来标记语句，提供了一种可扩展的替代手动标注的方法。这种方法在使用言语功能分类标注对话方面，比众包工人的表现更为出色 (Eggins and Slade, 2004)。

我们的目标是评估如此详细的注释是否能够通过在原始数据和重新注释的数据上微调模型来提高基于大型语言模型的辅导质量。具体来

说，我们在原始粗粒度注释和新的细粒度注释上微调 Mistral-7B-Instruct，并使用自动指标和人工评估对生成的辅导响应进行比较。我们的结果表明，细粒度模型产生的响应在教育对齐和效果上更优。为了促进进一步的研究和开发，我们发布了一个公共存储库，包含代码和重新注释的数据集。¹

2 背景 & 相关工作

2.1 MathDial 数据集

我们建立在 Macina et al. (2023) 的基础工作之上，其数据集为推进教育对话系统提供了无价的基础。MathDial 是一个大型、高质量的对话辅导数据集，专注于多步骤数学推理问题。与之前在教育质量低、规模小或缺乏基础等方面存在不足的数据集不同，MathDial 提供了基于现实学生疑惑和教学策略的丰富注释。作者引入了一种新的半合成数据收集框架，将专家教师与模拟学生及其错误的大型语言模型 (LLMs) 配对，能够可控且大规模地创建与真实辅导情景紧密相似的教育对话。该方法有效地解决了与众包或课堂录音相关的隐私问题和质量问题。

作者的方法论由一个受“绿野仙踪”启发的框架 (Kelley, 1984) 组成，其中专业教师与充当学生的 LLM 进行一对一辅导对话。这些学生模型经过精心设计，附有学生档案，并通过对 LLM 生成的多样推理路径进行温度采样产生的常见概念错误。所使用的数学文字题 (MWP) 来源于 GSM8K (Cobbe et al., 2021)。教师被指导使用四种教育举措的分类结构来支持学生理解：焦点、探测、陈述和通用的，并包含更细致的意图 (见 Table 1)。

关键是，在撰写回应之前，教师必须标注所采用的教学策略，鼓励更有目的性的策略使用。这些对话也基于元数据，包括具体的困惑、完整的问题、逐步解决方案，以及困惑是否得到解决，从而为培训 AI 导师提供丰富的信号。

实证评估表明，在 MathDial 上微调的模型在正确性和公平的辅导方面显著优于零射和指令微调的大型 LLM，例如 ChatGPT。值得注意的是，微调的开源模型在学生问题解决成功率上取得了类似的成绩，同时减少了“告知”——过早给出解决方案的事件。人工评估证实，这些微调模型在连贯性、正确性和教学效果上都优于使用提示的大型模型。

¹可在 <https://github.com/Kpetyxova/autoTree/tree/main/mathdial> 获得。

2.2 标注框架

尽管手动话语注释代价高昂且耗时，但基于大型语言模型 (LLM) 的注释进展提供了一种很有前途的替代方案，在速度、一致性和成本效益方面显示出改进 (Gilardi et al., 2023; Hao et al., 2024)。Petukhova and Kochmar (2025) 最近提出了一个开源管道，用于利用 LLM 进行全自动话语注释。具体而言，该管道自动构建层次树注释方案以及对对话中语句的注释，使其成为一种有前景且可扩展的方法，可以通过更详细的教师意图注释来丰富 MathDial 数据集。

Petukhova and Kochmar (2025) 探索了树构建和注释的多种配置，包括二叉和非二叉结构、基于频率的分组和最佳分割策略，并报告说，使用 GPT-4o 的基于频率指导的最佳分割选择在对话注释任务中基于言语功能分类时优于众包工作者 (Eggins and Slade, 2004)，同时将总注释时间从超过 30 小时减少到不到 1.5 小时。因此，在我们的工作中，我们采用这个公开可用的实现配置。²

2.3 受控生成

控制文本生成 (CTG) 旨在指导语言模型生成符合特定属性或约束 (例如情感、风格或意图) 的输出。CTG 中一种常见的方法是通过包含明确意图标签的提示对模型进行微调，从而生成与所需行为一致的文本 (Liang et al., 2024)。

指令微调已经成为一种有效的策略。通过在数据集上训练模型，其中的提示被增加了自然语言指令或意图标签，模型学会相应地调整它们的输出。例如，InstructCTG 框架展示了在自然语言描述和约束演示的条件下，模型如何在不变解码过程的情况下生成满足各种要求的文本 (Zhou et al., 2023)。

这种方法在教育环境中特别有利，因为将生成的内容与教学策略对齐是至关重要的。通过使用指定教学目的的提示来微调模型，AI 导师可以为学习者提供更有效和量身定制的支持 (Jia et al., 2025)。

3 重新注释 MathDial 数据集

为了构建 MathDial 中提出的扩展分类树，我们使用了 Petukhova and Kochmar (2025) 中的最佳框架配置——频率导向的最佳分割选择和回溯。该方法通过对候选分割进行评分并选择排名最高的分割来迭代选择，如果无法形成可行的划分，则采用回溯。此外，该方法偏向于向更频繁的类别构建树，使它们更快达到，并生成更好地反映现实世界类别分布的树。该树

²<https://github.com/Kpetyxova/autoTree>

Category	Intent	Example
Focus	Seek Strategy	So what should you do next?
	Guiding Student Focus	Can you calculate ...?
	Recall Relevant Information	Can you reread the question and tell me what is ...?
Probing	Asking for Explanation	Why do you think you need to add these numbers?
	Seeking Self Correction	Are you sure you need to add here?
	Perturbing the Question	How would things change if they had ... items instead?
	Seeking World Knowledge	How do you calculate the perimeter of a square?
Telling	Revealing Strategy	You need to add ... to ... to get your answer.
	Revealing Answer	No, he had ... items.
Generic	Greeting/Farewell	Hi ..., how are you doing with the word problem? Good Job! Is there anything else I can help with?
	General Inquiry	Can you go walk me through your solution?

Table 1: 教师从 MathDial 到 (Macina et al., 2023) 的话语及其意图的示例。

是基于十一种意图名称及其对应的例子生成的 (见表 1)。以文本形式在 Figure 1 中展示的最终树具有深度为二的结构, 从根节点分出五个分支。

有趣的是, 大多数意图按照在 Macina et al. (2023) 中定义的高级类别进行分组, 除了探测意图, 这些意图被分成两个独立的组: (1) 请求解释和寻求自我纠正, 以及 (2) 扰动问题和寻求世界知识。虽然这种划分不是预先定义的, 但它是可以解释的: 第一组以促使学生反思和评估他们的推理为中心, 而第二组则鼓励他们探索超出当前问题的更广泛或外部概念。

3.1 注释

在 2,861 个对话中, 我们随机选择了 500 个用于训练, 100 个用于验证, 100 个用于测试。

在 MathDial 中, 原始导师意图注释的一个例子显示在 Figure 2 中。在原始注释中, 每个教师话语都应用一个标签, 这在进行高层次分析时是有效的, 但可能会限制需要更细粒度控制的下游应用的灵活性。例如, 一个话语 [I see.]₁ [但我们在处理的是整个馅饼, 而不是切片。]₂ [如果你有一个生日蛋糕, 并且派对上有很多客人, 你不能只是不停地切蛋糕片。]₃ [你能想到另一种方法来确保每个人都有一块吗?]₄ 在数学对话中被标注为探测。然而, 这句话包含几个具有不同功能的话语单元: 片段 [1] 似乎是通用的, 片段 [2] 因其转移学生的注意力而与焦点 (特别是指导学生焦点) 相符, 片段 [3] 适合探究类别, 而片段 [4] 则对应于焦点 (寻求策略), 因为它促使学生思考替代解决方案。

相反, 在其他情况下, 指定的标签似乎与话语的最后部分相符。例如, 话语 [但是牧羊人派总是有 4 片, 所以使用总的片数可能没有帮助。]₁ [有没有其他可以用来除以派里片数的数量?]₂ 被标记为焦点。在这里, 虽然焦点应用于第二个句子 [2], 更合适的做法是将第一个句子 [1] 标记为叙述。这种不一致性——有时标签基于第一段, 有时基于最后一段——

强调了为了某些下游任务, 采用更细致和一致的注释方法可能带来的好处。

理想情况下, 注释应在基础话语单元 (EDUs) 层面上进行, 而不是整个话语。EDUs 是通常对应于句子的文本片段 (Jurafsky and Martin)。因此, 在这项工作中, 我们首先通过将教师的话语拆分为 EDUs 来预处理数据。

由于目前没有现成的方法可以自动将文本划分为 EDU, 我们采用以下策略: (1) 标点符号去除: 首先, 我们去除话语中的所有标点符号; (2) 标点符号恢复: 接下来, 使用为此任务训练的模型恢复标点符号;³ (3) 比较与分割: 最后, 我们将标点符号恢复后的版本与原始话语进行比较。如果恢复后的标点在原文中用句号、问号或感叹号替换了逗号, 我们就在该逗号处分割话语, 从而创造出独立的 EDU。默认情况下, 我们还将不同的句子分割为独立的 EDU。通过此过程从原始话语中产生的每个 EDU 继承了在 MathDial 中分配给完整话语的原始标签 (即四个高级类别之一)。

在数据被分割成 EDU 后, 训练集中的教师话语数量为 5,174。验证集和测试集也同样被分割成 EDU, 并且每个集被限制为 100 条教师话语。

利用生成的树结构, 应用了一种基于 GPT-4o 的标注流程来自 Petukhova and Kochmar (2025)。由于该树的结构与 MathDial 的作者提出的分层意图关系相符, 我们可以合理地预期基于此树的标注将反映这些关系。例如, 如果使用该树进行标注并赋予“扰动问题”的标签, 则原始标注相应地应该包含“探测”, 等等。基于这种对齐, 我们可以评估标注质量, 至少是在与原始高层级标注一致性方面。

Table 2 展示了在训练集上比较低层次意图注释与原始高层次教师行为类别时的加权精度 (P_w)、召回率 (R_w) 和 F1 ($F1_w$), 以及宏 F1 ($F1$) 分数。由于原始教师话语被分割为 EDU, 同时保留相同标签, 因此预期这些得分

³<https://huggingface.co/oliverguhr/fullstop-punctuation-multilang-large>

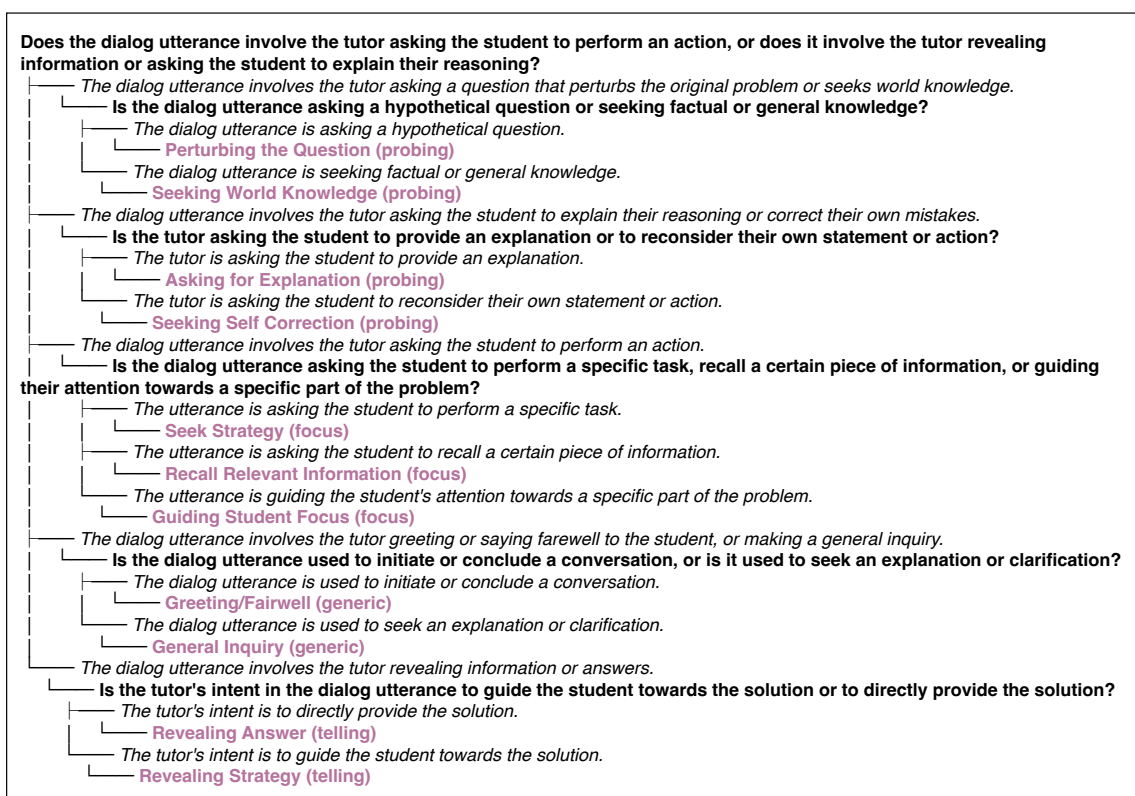


Figure 1: 使用 Petukhova and Kochmar (2025) 框架为 MathDial 数据集的扩展分类创建的树。对应树节点的问题用粗体显示，代表分支的可能答案用斜体显示，表示十一种意图的叶节点在 紫色粗体 中。

较低。如前所述，同一话语中的不同 EDU 通常会属于不同类型，这在原始注释中未被考虑在内 MathDial。

P_w	R_w	$F1_w$	$F1$
0.40	0.38	0.36	0.27

Table 2: 在训练集上对 11 标签标注的评估，使用来自 Petukhova and Kochmar (2025) 的标注框架，将新的对齐方式与来自 MathDial 的原始 4 标签标注进行比较。

在 5,174 个教师话语中，有 1,319 个保持不变，因为它们最初只包含一个 EDU。这些话语的标注结果在 Table 3 中展示。虽然这些指标高于 Table 2 中的指标，但它们仍然显示出相对较差的表现。

P_w	R_w	$F1_w$	$F1$
0.48	0.45	0.43	0.31

Table 3: 对在训练集中保持不变的语句（即，最初由一个单一的 EDU 组成）的 11 标签标注进行评估，使用来自 Petukhova and Kochmar (2025) 的标注框架，将新的对齐方式与来自 MathDial 的原始 4 标签标注进行比较。

然而，手动分析显示原始标注中存在显著的

不一致性。请考虑以下说明性例子：

- 一名学生起初将 14 确认为任务的正确最终答案。然而，在讨论过程中，该学生错误地将最终解答重述为 $10 + 10 + 4 = 24$ 。教师回应说，“那是 14 吗？”——指的是早些时候当 14 被正确确认为预期答案的时刻（完整对话见 Appendix A）。基于树的注释将教师的这句话分类为 寻求自我修正，对应于 探查。然而，在原始数据集中，它被标记为 讲述，我们认为这并不准确。
- 导师说，你需要在 $(8-2)$ 加上括号，并记住运算的顺序。学生回答，是的，我现在明白了。正确的方程应该是 $6 + (8 + 8) - 2 = 22$ 本新书。老师回复，不，我说的是 $8-2$ ，不是 $8+8$ 。虽然基于树的注释将这段话标记为揭示答案（告知），但原始注释将其标记为通用，可能反映了不同的解释或背景判断。
- 一位学生陈述， $6 + 8 + (8 - 2) = 22$ 。老师回答，请解释你是如何得到 22 的。基于树的注释将这个话语分类为 请求解释，它对应于 探测。然而，在原始数据集中，它

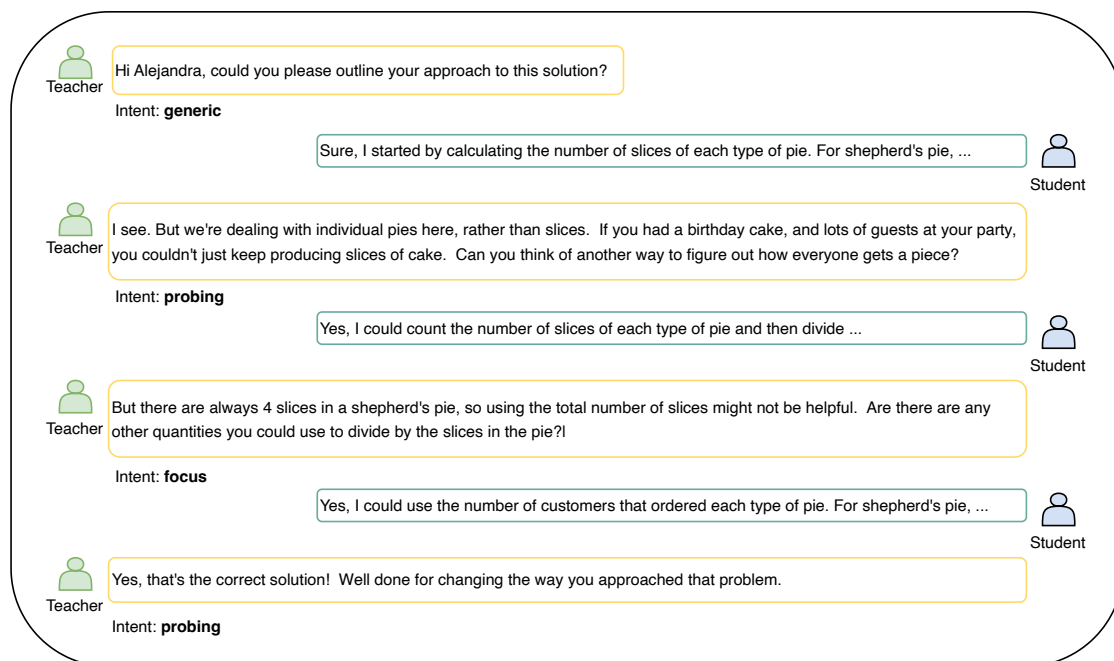


Figure 2: MathDial 中的教师话语及其标注意图的一个例子。

被标记为通用，这与话语的意图不太一致。

鉴于数据集的原始注释中存在如此多不明确或模棱两可的案例，我们不能断定基于树的注释是不准确的。相反，原始注释中的这些不一致性表明，评估指标中的差异至少部分可能是由于原始数据集中的歧义造成的。

十一种预测意图在所有数据集分割（训练集、验证集和测试集）中的分布如 Figure 3 所示。

4 可控生成

为了展示使用 Petukhova and Kochmar (2025) 框架收集的注解扩展分类法的好处，我们微调了一个 LLM，以预测下一步教师的发言。该模型使用数学任务描述、其标准解答、学生的解答、对话历史以及由注解框架预测的教师下一步发言意图进行训练。

此外，我们使用原始的四意图注释微调了同一模型的第二个版本。然后，我们将这两个微调后的模型相互比较，并与在零样本设置中相同的 LLM 进行比较。

我们使用 Mistral-7B-Instruct 作为微调的基础模型，特别是它来自 Hugging Face 的 4 位量化版本。最大序列长度设置为 1600。我们使用 QLoRA（量化低秩适配）对模型进行微调，这是一种参数高效的方法，通过使用量化的低秩适配器来减少内存和计算成本。我们使用的

秩为 $r = 32$ ，缩放因子为 $\alpha = 32$ 。微调在一个周期内进行，学习率为 $2e^{-5}$ ，批大小为 8，梯度累积为 4。我们采用 AdamW 优化器，线性调度并进行热身 (0.1)，权重衰减为 0.1，并每 50 步使用 SACREBLEU 进行评估。

我们将带注释的样本转换为提示和金标准输出的对，其中每个提示包括一个指令、数学任务、任务的金标准解答、学生的解答、对话历史和后续教师话语的意图（这些均可从带注释的数据中获得）。虽然在训练和评估期间您可以访问意图作为注释——因为我们可以访问金标准的下一个教师话语并能够对其意图进行分类——但对于现实世界的应用，这一意图需要由一个单独的模型进行预测，作为受控生成管道的一部分。提示模板如 Appendix B 所示。

我们使用基于参考的指标对生成输出进行自动评估，包括 CHRF++（字符 n -gram F 得分）(Popović, 2017)，SACREBLEU（ n -gram 精确度分数的加权几何平均值），以及 ROUGE-1，ROUGE-2 和 ROUGE-L（ n -gram 重叠的以召回为导向的测量）(Lin, 2004)。此外，我们进行一个小规模的人类评估。

结果 Table 4 展示了零次学习和微调两种设置下的生成结果，并比较了两种注释方案：MathDial 数据集中提供的原始四个教师意图以及扩展的十一种意图。正如预期的那样，经过微调的 LLM 表现优于零次学习的基线，并且在更细粒度的十一种意图注释下训练的模型

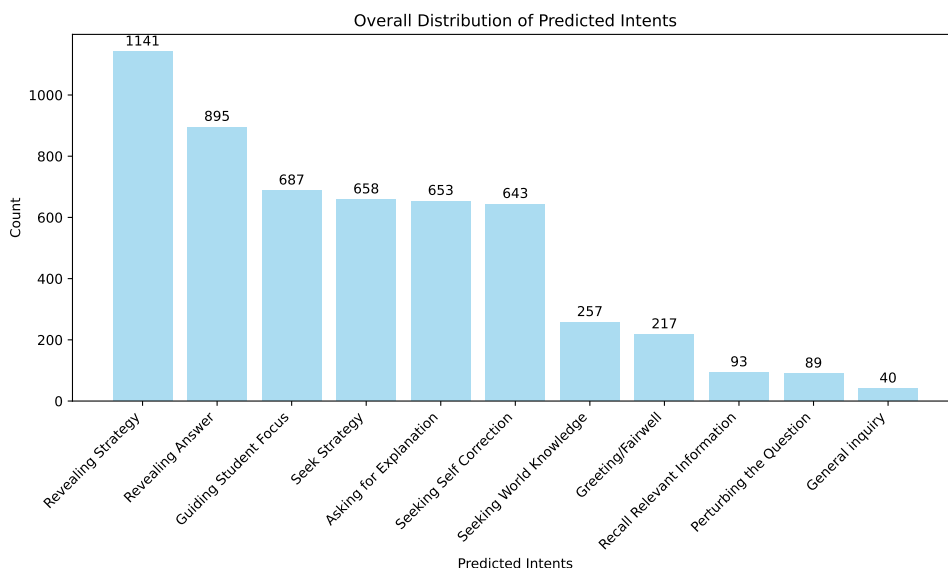


Figure 3: 对重新标注数据集中预测的十一种意图的总体分布。

Configuration	CHRF++	SACREBLEU	ROUGE-1	ROUGE-2	ROUGE-L
Zero-Shot, 4 intents	16.50	0.93	8.93	2.19	7.10
Zero-Shot, 11 intents	17.11	0.73	8.72	1.95	6.87
Fine-Tuning, 4 intents	16.82	2.67	17.13	5.61	15.95
Fine-Tuning, 11 intents	18.06	4.59	20.73	7.39	19.28

Table 4: 在零样本设定和通过微调 Mistral 对 MathDial 的测试集上进行受控生成的评估，比较使用来自 MathDial 的原始四个意图和使用 Petukhova and Kochmar (2025) 的框架注释的十一种意图。

在所有指标上均能持续获得更高的分数。

除了自动化指标外，我们还进行了人工评价，参与者为四位至少拥有自然语言处理硕士学位的标注者。我们从测试集中随机选择了七个对话，得到 30 对响应——一个来自在四个意图上微调的模型，一个来自在十一个意图上微调的模型。每位标注者都会看到这些对并被要求决定哪个响应更好，或者两个都同样好或差（见 Figure 4）。根据多数投票，56.7% 的情况下更偏好 FT-11 模型的响应。⁴ 使用 Fleiss’ Kappa 测量的标注者之间的协议为 $\kappa = 0.33$ ，表明协议程度为一般。

手动分析（见 Figure 5）表明，在以下观察中，基于十一种意图微调的模型（FT-11）生成的响应优于基于四种意图微调的模型（FT-4）生成的响应：

- FT-11 一贯表现出对会话策略的更深刻理解，例如使用更有效的提问技巧。例如，在处理错误的舞会情侣计算时（见第一个例子 Figure 5），FT-11 明确提示学生重新考虑原始条件（那么，如果我们知道舞会上有 123 名学生，有多少对情侣？）。相比之下，FT-4 只是简单地重述了错误的情境

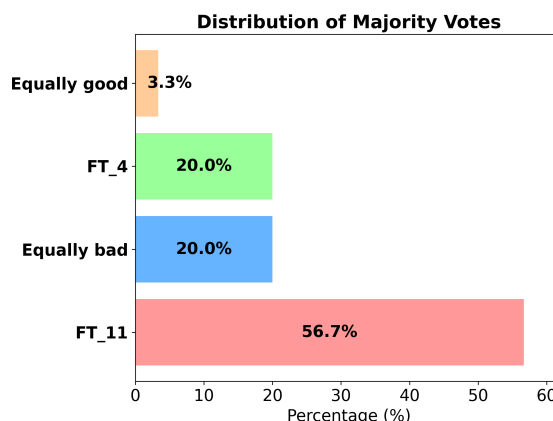


Figure 4: 人工评估的结果。我们邀请了四位注释者选择更好的回答，或者指出如果两个回答同样好或同样差。最终决定是通过多数投票确定的。

⁴在多数投票中没有出现平票现象——每个例子都得到了明确的决定。

(那么, 如果我们有 120 对情侣, 有多少学生参加舞会?), 这在引导学生意识到他们的错误方面的效果较差。

- FT-11 更直接地解决了学生的误解。在 Figure 5 的第二个例子中, FT-11 直接质疑学生的任意假设 (你是怎么得到 100 头牛的?), 与老师的答案非常接近 (克莱尔, 你为什么认为农夫有 100 头牛?)。FT-4 则不太集中, 要求学生解释计算过程, 而不是解决误解的根本原因。
- FT-11 的回答往往简明且相关, 促使学生对他们的推理进行批判性反思, 而不是重复先前的陈述。例如, 在 Figure 5 的第三种情境中, FT-11 简明地承认正确性 (正确。), 与实际的教师回应 (没错。) 很好地吻合, 而 FT-4 不必要地重复之前的问题, 表现出较低效的对话管理。

5 结论

在这项工作中, 我们考察了细粒度注释对 MathDial 数据集中受控响应生成的影响。通过将教师行为的原始分类从四个大类扩展到十一种更具体的意图, 并使用来自 Petukhova and Kochmar (2025) 的自动化树创建和注释框架, 我们证明了这种方法能增强微调后的 LLM 在生成有意义的响应时的表现。

结果证实, 在具有高粒度标签的数据集上进行微调可以更好地与期望的教师响应对齐, 优于零样本设置和在原始四类注释上进行的微调。这表明意图标签的具体性对于提高模型生成有针对性和有效的辅导响应的能力至关重要。

此外, 对注释过程的人工分析揭示了原始数据集中的不一致, 突显了高质量、结构良好的意图注释的重要性。Petukhova and Kochmar (2025) 中提出的基于树的注释框架提供了一种系统的方法来改进此类数据集, 使其更适合用于训练可控生成模型。

由于资源限制, 我们仅重新标注了 MathDial 数据集的一个子集, 并使用一个小型语言模型进行微调。

这些局限性提出了未来工作的几个方向: (1) 重新标注整个数据集以便训练质量更高的模型, 以及 (2) 探索规模更大的开放源代码模型以改进微调性能。

此外, 我们承认我们的结论是基于自动化指标和小规模人工评估。为了展示这些发现对实际世界的影响, 并评估生成的辅导干预的教学价值, 未来的工作应考虑通过实际教师和学生来验证这些结论。

6

伦理考虑 由于这项工作是探索性的, 并且本研究使用的模型输出尚未在真实学生中进行测试, 我们不预计此工作或使用重新标注的数据集会存在任何重大风险。同时, 我们承认该工作使用了大语言模型 (LLMs), 此类模型在实际教育场景中应用时可能会带来风险, 因为它们可能生成看似合理但实际上不准确或无意义的输出, 这反过来可能导致错误的决策和偏见的传播。虽然我们不认为我们的论文中提出的研究存在任何直接风险, 但如果基于此研究的未来工作将所提出的方法应用于实际场景, 应适用适当的安全保障措施。

我们感谢谷歌通过 2024 年度谷歌学术研究成果奖 (GARA) 对本研究的支持。

References

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Benjamin S Bloom. 1984. The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational researcher*, 13(6):4–16.
- Timothy A Carey and Richard J Mullan. 2004. What is Socratic questioning? *Psychotherapy: theory, research, practice, training*, 41(3):217.
- Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, and 1 others. 2021. Training verifiers to solve math word problems, 2021. URL <https://arxiv.org/abs/2110.14168>, 9.
- Paul Denny, Sumit Gulwani, Neil T Heffernan, Tanja Käser, Steven Moore, Anna N Rafferty, and Adish Singla. 2024. Generative AI for education (GAIED): Advances, opportunities, and challenges. *arXiv preprint arXiv:2402.01580*.
- Suzanne Eggins and Diana Slade. 2004. *Analysing casual conversation*. Equinox Publishing Ltd.
- Scott Freeman, Sarah L Eddy, Miles McDonough, Michelle K Smith, Nnadozie Okoroafor, Hannah Jordt, and Mary Pat Wenderoth. 2014. Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the national academy of sciences*, 111(23):8410–8415.
- Wensheng Gan, Zhenlian Qi, Jiayang Wu, and Jerry Chun-Wei Lin. 2023. Large language models in education: Vision and opportunities. In *2023 IEEE international conference on big data (BigData)*, pages 4776–4785. IEEE.

Task	Student's Incorrect Solution	Dialog
Chalktown High School had their prom last weekend. There were 123 students who attended. If 3 students attended on their own, how many couples came to the prom?	Let's call the number of couples that attended the prom "c". We know that there were 123 students total, and that 3 of them attended on their own. This means that the rest of the students attended as couples. So we can set up the equation: couples + 3 = 123 Subtracting 3 from both sides: couples = 120 So there were 120 couples that attended the prom. 120	<p>Teacher: Hey Heidi, do you know what a couple is?</p> <p>Student: Yes, a couple is two people who are together, usually a romantic couple.</p> <p>Teacher: So if we had 120 couples attending the prom, how many students would that be in total?</p> <p>Student: 120 couples would be 240 students in total, since each couple is made up of two people.</p> <p>Teacher: Great.</p> <p>Teacher: would you like to check your solution again, bearing in mind what youve just told me, please?</p> <p>Student: Sure. So if we have 120 couples, that would mean there are 240 students in total. We know that 3 students attended on their own, so if we subtract 3 from 240, we get 237. So there were 120 couples and 237 students in total at the prom.</p> <p>Teacher: We need to read the question again.</p> <p>Teacher: the question tells us how many people were at the prom.</p> <p>Next Teacher Utterance:</p> <p>Gold: lets start over from the beginning and ill help you if needed.</p> <p>FT-4 (focus): So, if we have 120 couples, how many students attended the prom?</p> <p>FT-11 (Seek Strategy): So, if we know that there were 123 students at the prom, how many couples were there?</p>
40% of a farmer's cattle are males. The rest are females. If a female cow produces 2 gallons of milk a day, how much milk will the farmer get a day if he has 50 male cows?	If 40% of the cattle are males, then $100\% - 40\% = 60\%$ are females. Let's say the farmer has a total of 100 cows, then 40% of 100 = 40 cows are males. So 60 cows are females. If each female cow produces 2 gallons of milk a day, then $60 \times 2 = 120$ gallons of milk will be produced a day. Since the farmer has 50 male cows, they don't produce milk, so the total milk produced a day is 120 gallons.	<p>Teacher: Hi, could you please walk me through your solution?</p> <p>Student: Sure. I started by figuring out that 40% of the cattle were males, so $100\% - 40\% = 60\%$ were females. Then I assumed that the farmer had a total of 100 cows, so 40% of 100 = 40 cows were males and 60 cows were females. I then multiplied the number of female cows (60) by 2 gallons of milk a day to get 120 gallons of milk produced a day. Since the farmer only has 50 male cows, they don't produce any milk, so the total milk produced a day is 120 gallons.</p> <p>Next Teacher Utterance:</p> <p>Gold: Claire, why did you assume that the farmer had 100 cows?</p> <p>FT-4 (generic): Hi, could you please explain how you got 120 gallons of milk from 60 cows?</p> <p>FT-11 (Seek Strategy): How did you get 100 cows?</p>
It takes 3 men an hour to complete a job. If they are hired to work on 5 such similar jobs by a company that pays each of them \$10 per hour, calculate the total amount of money the three will earn when they complete the jobs?	It takes 3 men an hour to complete a job, so it will take them $3 \times 5 = 15$ hours to complete the 5 jobs. If they are paid \$10 per hour, the total amount of money each of them will earn when they complete the jobs is $\$10 \times 15 = \150 . Therefore, the total amount of money the three will earn when they complete the jobs is $\$150 \times 3 = \450 .	<p>Teacher: Hi Brenda.</p> <p>Teacher: lets take this step by step.</p> <p>Teacher: How much would one man earn for one job?</p> <p>Student: Hi, one man would earn \$10 for one job.</p> <p>Teacher: How many jobs are there?</p> <p>Student: There are 5 jobs.</p> <p>Teacher: How much would one man doing the 5 jobs earn?</p> <p>Student: One man doing the 5 jobs would earn $\\$10 \times 5 = \\50.</p> <p>Teacher: Great, How many men are there?</p> <p>Student: There are 3 men.</p> <p>Teacher: So each man is paid 50 total for the 5 jobs and there are 3 men.</p> <p>Teacher: so how much is the total for the 3 men?</p> <p>Student: The total for the 3 men is $\\$50 \times 3 = \\150.</p> <p>Next Teacher Utterance:</p> <p>Gold: Thats right.</p> <p>FT-4 (generic): So how much is the total for the 3 men?</p> <p>FT-11 (Revealing Answer): Correct.</p>

Figure 5: 由在原始四意图注释集上微调的模型（FT-4）和使用 Petukhova and Kochmar (2025) 中的框架对具有十一种意图的数据进行微调的模型（FT-11）生成的响应示例，以及来自教师的实际响应（Gold）。

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120.

Jing Hao, Yuxiang Zhao, Song Chen, Yanpeng Sun, Qiang Chen, Gang Zhang, Kun Yao, Errui Ding, and Jingdong Wang. 2024. Fullanno: A data engine for enhancing image comprehension of mlms. *arXiv preprint arXiv:2409.13540*.

Linzhaio Jia, Changyong Qi, Yuang Wei, Han Sun, and Xiaozhe Yang. 2025. Fine-Tuning Large Language Models for Educational Support: Leveraging

Gagne's Nine Events of Instruction for Lesson Planning. *arXiv preprint arXiv:2503.09276*.

Daniel Jurafsky and James H Martin. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition.

John F Kelley. 1984. An iterative design methodology for user-friendly natural language office information applications. *ACM Transactions on Information Systems (TOIS)*, 2(1):26–41.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yu-

- taka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.
- Md Tahmid Rahman Laskar, M Saiful Bari, Mizanur Rahman, Md Amran Hossen Bhuiyan, Shafiq Joty, and Jimmy Xiangji Huang. 2023. A systematic study and comprehensive evaluation of ChatGPT on benchmark datasets. *arXiv preprint arXiv:2305.18486*.
- Xun Liang, Hanyu Wang, Yezhaohui Wang, Shichao Song, Jiawei Yang, Simin Niu, Jie Hu, Dan Liu, Shunyu Yao, Feiyu Xiong, and 1 others. 2024. Controllable text generation for large language models: A survey. *arXiv preprint arXiv:2408.12599*.
- Chin-Yew Lin. 2004. **ROUGE: A package for automatic evaluation of summaries**. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.
- Rongxin Liu, Carter Zenke, Charlie Liu, Andrew Holmes, Patrick Thornton, and David J Malan. 2024. Teaching CS50 with AI: leveraging generative artificial intelligence in computer science education. In *Proceedings of the 55th ACM technical symposium on computer science education V. 1*, pages 750–756.
- Jakub Macina, Nico Daheim, Sankalan Pal Chowdhury, Tanmay Sinha, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. 2023. Mathdial: A dialogue tutoring dataset with rich pedagogical properties grounded in math reasoning problems. *arXiv preprint arXiv:2305.14536*.
- Ethan Mollick and Lilach Mollick. 2024. Instructors as innovators: A future-focused approach to new AI learning opportunities, with prompts. *arXiv preprint arXiv:2407.05181*.
- Benjamin D Nye, Arthur C Graesser, and Xiangen Hu. 2014. AutoTutor and family: A review of 17 years of natural language tutoring. *International Journal of Artificial Intelligence in Education*, 24:427–469.
- Sankalan Pal Chowdhury, Vilém Zouhar, and Mrinmaya Sachan. 2024. Autotutor meets large language models: A language model tutor with rich pedagogy and guardrails. In *Proceedings of the Eleventh ACM Conference on Learning@ Scale*, pages 5–15.
- Kseniia Petukhova and Ekaterina Kochmar. 2025. A Fully Automated Pipeline for Conversational Discourse Annotation: Tree Scheme Generation and Labeling with Large Language Models. *arXiv preprint arXiv:2504.08961*.
- Maja Popović. 2017. **chrF++: words helping character n-grams**. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.
- Rose E Wang, Qingyang Zhang, Carly Robinson, Susanna Loeb, and Dorottya Demszky. 2023. Bridging the novice-expert gap via models of decision-making: A case study on remediating math mistakes. *arXiv preprint arXiv:2310.10648*.
- Shen Wang, Tianlong Xu, Hang Li, Chaoli Zhang, Joleen Liang, Jiliang Tang, Philip S Yu, and Qingsong Wen. 2024. Large language models for education: A survey and outlook. *arXiv preprint arXiv:2403.18105*.
- Sebastian Wollny, Jan Schneider, Daniele Di Mitri, Joshua Weidlich, Marc Rittberger, and Hendrik Drachler. 2021. Are we there yet?-a systematic literature review on chatbots in education. *Frontiers in artificial intelligence*, 4:654924.
- Jingfeng Yang, Hongye Jin, Ruixiang Tang, Xiaotian Han, Qizhang Feng, Haoming Jiang, Shaochen Zhong, Bing Yin, and Xia Hu. 2024. Harnessing the power of llms in practice: A survey on chatgpt and beyond. *ACM Transactions on Knowledge Discovery from Data*, 18(6):1–32.
- Wangchunshu Zhou, Yuchen Eleanor Jiang, Ethan Wilcox, Ryan Cotterell, and Mrinmaya Sachan. 2023. Controlled text generation with natural language instructions. In *International Conference on Machine Learning*, pages 42602–42613. PMLR.

A 树创建步骤的输出

Task: Harry decided to buy some balloons for his friend's birthday party. One balloon costs \$0.5, and a pack of 10 balloons is cheaper and costs only \$3. Finally, Harry wants to bring to the party exactly 14 balloons. How much did he need to pay for them?

Correct Solution: To buy 14 balloons, Harry should buy one pack of 10, and $14 - 10 = 4$ separate balloons.
The price of 4 balloons is $4 \times 0.5 = \$2$.
In total, the pack and separate balloons would cost Harry in total $3 + 2 = \$5$.

5

Student's incorrect answer: Harry needs to buy 2 packs of 10 balloons, which will give him 20 balloons in total. He will then need to buy 4 more individual balloons to get to 14.

2 packs of 10 balloons cost $2 \times \$3 = \6 .

4 individual balloons cost $4 \times \$0.5 = \2 .

Therefore, Harry needed to pay $\$6 + \$2 = \$8$ for 14 balloons.

8

Dialog:

Teacher: (generic)Hi Michael, could you please walk me through your solution?

Student: Sure. I calculated that Harry needed to buy 2 packs of 10 balloons, which cost $2 \times \$3 = \6 . Then, he needed to buy 4 individual balloons, which cost $4 \times \$0.5 = \2 . So, the total cost was $\$6 + \$2 = \$8$.

Teacher: (focus)So can you please tell me why he needs two packs of 10 (which is 20) and another 4 on top? What's $20 + 4$?

Student: Harry needs two packs of 10 balloons because each pack costs \$3 and buying two packs is cheaper than buying 20 individual balloons. Then, he needs to buy 4 more individual balloons to get to 14 balloons in total. $20 + 4 = 24$, so Harry has enough balloons for the party.

Teacher: (focus)Please could you tell me how many balloons Harry requires in the math problem.

Student: Harry needs 14 balloons for the party.

Teacher: (focus)So is that 24?IEOM!Student: No, 14. He needs two packs of 10 balloons and 4 individual balloons, which gives him a total of 24 balloons, but he only needs 14.

Teacher: (focus)So how many packs of 10, plus single balloons does he really need to equal 14?

Student: Harry needs two packs of 10 balloons and 4 individual balloons to equal 14.

Teacher: (probing)so what's $10+10+4$?

Student: $10+10+4 = 24$.

Teacher: (telling)Is that 14?

Student: No, 24 is not 14. Harry needs two packs of 10 balloons and 4 individual balloons, which gives him a total of 24 balloons, but he only needs 14.

Teacher: (telling)So Harry needs one pack of 10, which equals \$3. Plus he needs four separate balloons at \$0.5 each, can you tell me how much he needs to spend?

Student: Harry needs to spend \$3 for one pack of 10 balloons and \$2 for four individual balloons, so the total cost is $\$3 + \$2 = \$5$.

Teacher: (generic)Brilliant, can you see where you were duplicating the packs of 10

Student: Yes, I was calculating that Harry needed to buy two packs of 10 balloons, which cost $2 \times \$3 = \6 . Then, I was calculating that he needed to buy 4 individual balloons, which cost $4 \times \$0.5 = \2 . So, the total cost was $\$6 + \$2 = \$8$, which was incorrect because I was duplicating the packs of 10.

Figure 6: 来自 MathDial 数据集的对话示例。

B 用于微调的提示模板

```
[INST]
You are a math tutor. You will receive a task, a correct ("gold") solution, a student's
incorrect solution, and an ongoing dialog between the student and the tutor.
Your objective is to generate the next tutor response based on the given intent.

Task:
{task}

Gold Solution:
{gold_solution}

Student's Incorrect Solution:
{student_incorrect_solution}

Dialog:
{dialog}

Intent for the Next Tutor Utterance:
{intent}

[/INST]
### Tutor:
```

Figure 7: 用于在 MathDial 上微调的提示模板。