arXiv:2611.28359v1 [cs.CL] 20 Nov 2026

通过双分支监督的对比式在策略蒸馏

Zhibo Hou、Fan Zhao、Yiping Li、ZhiyuAn、WanDu加利福尼
亚大学默塞德分校{zhou6,fanzhao,yipingli,zan7,wdu3}
@ucmerced.edu

摘要

对于在线策略蒸馏（OPD），能力更强的模型自然是教师模型的首选。然而，任务表现更好并不一定意味着可教性更强。一项受控令牌干预研究表明，即使模型弱于学生模型，也可以作为有参考价值的对比参考。基于这一观察，我们提出对比式在策略蒸馏（cOPD），通过在学生生成的前缀处对比强教师、弱教师和学生模型的预测，提供双分支监督。我们的分析解释了在理想化模型下，这种三方比较如何帮助筛选出与更好或更差推理结果相关的令牌候选项。我们在涵盖不同训练阶段、参数规模和模型系列的多种教师—学生配置上评估 cOPD，任务包括数学推理、指令遵循以及带弃答机制的问答。在这三项任务中，cOPD 的表现均达到所有评估基线方法中的最佳或顶尖水平；在数学推理任务上，其整体 Pass@16 相较最强基线方法最高提升 7.51%。进一步分析表明，cOPD 能抑制不良响应模式，并显示其有助于促进探索。

1 引言

蒸馏已迅速成为大语言模型（LLM）后训练的核心技术。尤其是，在线策略蒸馏（OPD）（Agarwal 等，2024）从学生模型中采样轨迹，并在所得前缀处从教师模型的下一词元分布中获取监督，从而提供密集的令牌级指导。近期的大语言模型，包括 Qwen3（Yang 等，2025）、DeepSeek-V4（Xu 等，2026）和 GLM-5（Zeng 等，2026），已将 OPD 纳入其后训练流程，用于能力迁移与巩固，凸显了其与监督微调（SFT）和强化学习（RL）并列的日益重要的实际价值。

由于 OPD 使用教师模型的预测分布作为训练目标，因此，使用能力更强的模型作为教师模型既是自然的选择，也是一种广泛采用的设计方案（Jin等人，2026；Yang 等，2025；Gu 等，2024）。然而，近期研究表明，更高的教师模型准确率并不一定能够转化为
通过 OPD 带来更大的学生模型提升（Li 等，2026b；Fu 等，2026）。因此，模型解决任务的能力与其在蒸馏特定学生模型时的可教性并不相同。尽管性能较低的模型可能不是直接蒸馏时那么可靠的目标，但其分布仍可能包含有助于提升学生模型的信息。

这一关键观察引出了本文的研究问题：**表现不如学生模型的模型，是否仍能提供有益的监督？**我们研究较弱模型作为辅助教师的作用，将其预测分布与主要的、更强教师模型的预测分布进行对比。我们的假设是，比较二者的词元级偏好，有助于识别应当强化的有益词元选择和应当抑制的有害选择。词元干预研究为这一假设提供了初步支持。图 1 显示，当倾向于强教师模型时，由弱教师和强教师对比所引导的接管能够提升学生准确率；而当倾向于弱教师时，则会降低学生准确率。

本文提出了对比式在策略蒸馏（cOPD）。在学生模型生成的前缀处，cOPD 会选择这样的令牌候选项：它们在学生模型下的概率介于强教师和弱教师赋予的概率之间。通过对比教师模型的对数概率，cOPD 提供

双分支监督：正分支强化强教师偏好的选定倾向，负分支则抑制弱教师偏好的倾向。我们的分析阐释了在理想化模型下，比较强教师、学生模型和弱教师如何帮助选出与更好或更差推理结果相关的令牌候选项。我们在数学推理、指令遵循和带弃答机制的问答等任务上，针对多种教师—学生配置评估了 cOPD，结果表明，在所评估的基线方法中，cOPD 的表现最佳或位居前列。进一步的实证分析显示，cOPD 能有效利用弱教师的信息来抑制不理想的响应模式；定性证据还表明，它有助于探索更多有效响应。

我们的具体贡献如下：

1. 我们通过受控的词元接管实验表明，弱教师为识别有益和有害的词元偏好提供了互补信息。2. 我们提出了 cOPD，它利用强弱教师对比，有选择地强化有益的词元偏好并抑制有害的词元偏好。3. 我们在三个任务和多种教师—学生配置上提供了广泛的实证证据，表明 cOPD 在抑制不理想失败模式的同时，能够取得最佳或接近最佳的性能。

2 预备知识

令 $x \sim \mathcal{D}$ 表示一个输入提示，$y = (y_1、\ldots\ldots、y_L)$ 表示一个响应，其中 $y_{<t} = (y_1、\ldots\ldots、y_{t-1})$。我们将学生模型和教师模型分别记为 π_θ abd π_T，其下一词元分布为 $\pi(\cdot|x、y_{<t})$。

在线策略蒸馏。 OPD 使用教师监督，在学生模型自身生成的轨迹上训练学生策略。给定来自训练集 $\mathcal{D}$ 的提示 x，学生模型生成响应 $y \sim \pi_\theta(\cdot|x)$。在每个位置 t，教师模型都会提供下一词元分布 $\pi_T(\cdot|x、y_{<t})$，该分布以提示和学生模型生成的前缀为条件。学生模型通过最小化教师模型与学生模型分布之间的词元级散度进行更新：

$$\mathcal{L}_{OPD}(\theta) = \mathbb{E}_{x\sim\mathcal{D}, y\sim\pi_\theta(\cdot|x)}[\frac{1}{L}\sum_{t=1}^{L} D(\pi_T(\cdot|x, y_{<t}), \pi_\theta(\cdot|x, y_{<t}))]$$

其中 D 是散度度量，并且在每次更新期间，将采样得到的轨迹视为固定不变。

蒸馏目标。 词元级散度的常见选择包括正向 KL 散度、反向KL散度和詹森-香农散度（JSD）（Song和Zheng，2026）。令 $p(v) = \pi_T(v|x, y_{<t})$ 和 $q(v) = \pi_\theta(v|x, y_{<t})$，其中 $v \in \mathcal{V}$ 和 $\mathcal{V}$ 是完整词汇表。两个分布 $r(v)$ 和 $s(v)$ 之间的Kullback-Leibler（KL）散度（Kullback和Leibler，1951）定义为：

$$D_{KL}(r||s) = \sum_{v\in\mathcal{V}} r(v)\log\frac{r(v)}{s(v)}$$

正向 *KL* 散度，$D_{FKL}(p, q) = D_{KL}(p||q)$ 通过惩罚学生模型在教师模型支持的词元上概率不足的情况，促使其表现出覆盖模式行为。

反向*KL*散度，$D_{RKL}(p, q) = D_{KL}(q||p)$ 通过惩罚学生模型在教师模型认为不太可能出现的词元上分配概率质量，促使其表现出寻模行为。

詹森-香农散度，$D_{JSD}(p, q) = \frac{1}{2}D_{KL}(p||m) + \frac{1}{2}D_{KL}(q||m)$，其中 $m = \frac{1}{2}(p + q)$，通过将两个分布都与其混合分布进行比较，提供了一种对称且有界的替代方案。

大多数 OPD 实现都使用词汇表中的 top-k 子集来近似散度（Jin等人，2026年；Li 等，2026b）。

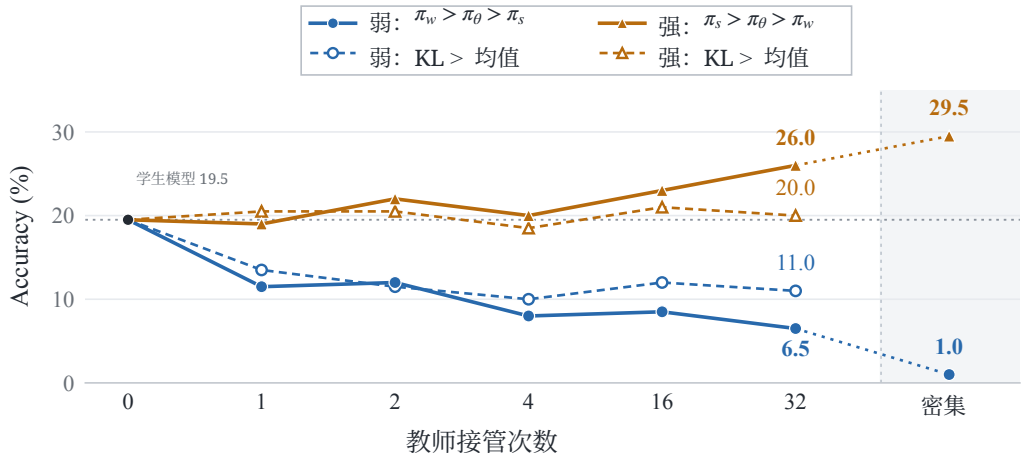

图 1：教师接管实验。随着接管预算增加，弱教师接管的负分支通常会降低学生准确率，而强教师接管的正分支通常会提高学生准确率。与基于 KL 的选择相比，这两种效果都更显著。

3 方法

3.1 CANA 弱教师能提供帮助吗?

OPD 在学生生成前缀处使用教师监督，以调整学生模型对下一个词元的预测。我们设计了一项受控的词元干预实验，旨在探究弱教师能否作为对比参考，用于识别影响学生模型最终答案正确性的关键位置和词元分布。该实验使用临时教师接管，作为 OPD 监督效果的一种简化行为代理。

我们使用 DeepSeek-R1-Distill-Qwen-1.5B（DeepSeek-AI，2025年）作为学生模型 π_θ，使用 Qwen-2.5-Math- 1.5B（Yang 等人，2024年）作为弱教师 π_w，并使用 JustRL-DeepSeek-1.5B（He 等人，2025年）作为强教师 π_s。我们从 DAPO-Math-17k 数据集（Yu 等人，2025年）中选取固定的 50 道题目，这些题目满足弱教师的 pass@$8 = 0$。每种干预设置均使用四个随机种子在该集合上进行评估。给定当前前缀 $h_t = (x,\ y_{<t})$，三个模型分别为候选词元 $v \in \text{Top}_8(\pi(\cdot|h_t))$ 分配概率。我们定义了两种接管条件：

$$\text{Weak teacher takeover:} \quad \pi_{\text{w}}(v \mid h_t) > \pi_\theta(v \mid h_t) > \pi_{\text{s}}(v \mid h_t),$$
$$\text{Strong teacher takeover:} \quad \pi_{\text{s}}(v \mid h_t) > \pi_\theta(v \mid h_t) > \pi_{\text{w}}(v \mid h_t),$$

在学生模型生成过程中，遇到第一个符合条件的位置时，相应的教师模型会接管生成并连续生成四个词元。随后，学生模型恢复生成，直到下一个符合条件的位置。此过程重复进行，直至达到接管预算，然后由学生模型完成响应。我们评估了预算为 $n \in \{1, 2, 4, 16, 32, \infty\}$ 次接管的情况。我们还评估了另一种单独的控制方式：当 KL 散度超过平均 KL 散度时触发接管。

如图 1 所示，随着预算增加，基于对比的选择在强教师接管下使学生准确率整体提高，而在弱教师接管下则使其降低。这些影响比基于 KL 的选择下的接管更为显著，这表明弱教师或许有助于提供额外信息，以识别能够提升或降低学生准确率的接管位置和分布。这促使我们将弱教师与强教师之间的这些对比纳入 OPD 监督，以强化有益的分布并抑制有害的分布。

3.2 对比式在线策略蒸馏

受这些发现的启发，我们提出对比式在策略蒸馏（cOPD），通过比较强教师、学生模型和弱教师在学生生成前缀处的三方概率，构建词元级监督。

3.2.1 对比式词元掩码

我们首先定义一个词元级掩码，用于筛选学生模型的概率介于弱教师和强教师概率之间的令牌候选项，这与动机实验中的做法相同。为排除近似平局的情况，我们要求两个概率差都大于间隔 δ。

对于学生生成的前缀 $h_t = (x, y_{<t})$，我们分别将 π_s、π_w 和 π_{ref} 设为强教师、弱教师和参考学生策略，其中 $v \in \text{Top}_{16}(\pi_{ref}(\cdot|h_t))$。我们定义了分别对应正分支和负分支的两个二值掩码：

$$\begin{aligned} m_v^+ &= \mathbf{1}[\pi_s(v|h_t) - \pi_{ref}(v|h_t) > \delta \wedge \pi_{ref}(v|h_t) - \pi_w(v|h_t) > \delta], \\ m_v^- &= \mathbf{1}[\pi_w(v|h_t) - \pi_{ref}(v|h_t) > \delta \wedge \pi_{ref}(v|h_t) - \pi_s(v|h_t) > \delta], \end{aligned} \tag{1}$$

其中，$\mathbf{1}[\cdot]$ 是指示函数。我们在每个训练周期内保持 π_{ref} 不变，并在训练周期结束时设置 $\pi_{ref} \leftarrow \pi_\theta$。正分支选择相对于参考学生受到强教师偏好、而不受弱教师偏好的词元。负分支则选择排序相反的词元。不满足上述任一条件的词元归入弃权情况。总体选择掩码为

$$m_v = m_v^+ + m_v^-. \tag{2}$$

3.2.2 对比式目标

我们首先通过对数概率对比来量化强教师和弱教师的相对偏好：

$$\Delta_v = \log \pi_s(v) - \log \pi_w(v) = \log \frac{\pi_s(v)}{\pi_w(v)}. \tag{3}$$

正的 Δ_v 表示强教师相较于弱教师更偏好词元 v，而负值则表示相反情况。其绝对值反映了二者相对偏好之间的差异。

设 $q_\theta(v) = \pi_\theta(v|h_t)$ 为当前学生模型的概率。我们将教师模型的对比信号与词元掩码结合，定义逐位置损失：

$$\ell_{cOPD}(\theta; h_t) = -\sum_{v \in \text{Top}_{16}(\pi_{ref})} m_v q_\theta(v) \tanh \frac{\Delta_v}{\tau}. \tag{4}$$

其中，温度 $\tau > 0$ 控制对比信号的尺度。正分支词元具有 $\Delta_v > 0$，并鼓励学生模型将概率质量分配给这些词元；负分支词元具有 $\Delta_v < 0$，并惩罚学生模型分配给这些词元的概率质量。弃权词元不会对损失产生任何贡献。

3.3 对比式监督分析

To 为了理解弱教师如何提供互补信息以选择训练目标 g
sign，我们给出理论分析。

先前工作通过下游成功概率分析令牌级教师引导，并区分有益、中性和有害的贡献（Armandpour 等人，2026年）。对于一个固定的学生生成前缀 h，我们将自回归语言策略的理想化下一个词元分布建模为以下混合分布：

$$p_i = \alpha H_i + \beta_i G + \gamma_i B, \quad i \in \{s, ref, w\}, \tag{5}$$

其中，p_s、p_{ref}、p_w 分别对应强教师、参考学生和弱教师，且 $\alpha + \beta_i + \gamma_i = 1$。归一化分布 G 和 B 在各模型之间共享，表示优质和劣质的推理候选项；在固定的学生模型续写下，它们的正确性概率分别高于和低于参考学生在 h处的预期正确性。任务中性组件 H_i 可能因模型而异，例如在上下文中两个表达语义等价时，某个模型可能偏好使用“but”而不是“however”。我们假设 G、B以及合并后的中性区域的支集彼此不相交，并且 $0 < \alpha < 1$ 和 $0 < \beta_w < \beta_{ref} < \beta_s < 1 - \alpha$。

命题 1（推理偏好的对比选择）。 在共享组件模型下，推理组件中的令牌候选项满足

$$\begin{aligned} v \in \text{supp}(G): &\quad p_s(v) > p_{ref}(v) > p_w(v), \qquad m_v^- = 0, \\ v \in \text{supp}(B): &\quad p_w(v) > p_{ref}(v) > p_s(v), \qquad m_v^+ = 0. \end{aligned}$$

表 1：数学推理结果（%）。Avg. 表示从 AIME24/AIME25/AIME26/AMC23 汇总的所有问题的总体 Pass@16。粗体和下划线分数分别表示最佳和次佳结果。

方法	AIME24 Pass@16 ↑	AIME25 Pass@16 ↑	AIME26 Pass@16 ↑	AMC23 Pass@16 ↑	Avg. Pass@16 ↑	Omni-Math Pass@1 ↑
JustRL / Qwen2.5-Math / DeepSeek-R1-Distill						
Base	70.00	46.67	40.00	87.95	69.36	16.84
强教师	76.67	60.00	40.00	93.98	75.72	28.18
弱教师	23.33	13.33	10.00	66.27	39.88	10.35
FKL	<u>76.67</u>	<u>53.33</u>	<u>43.33</u>	91.57	73.99	12.19
RKL	73.33	46.67	40.00	<u>93.98</u>	72.83	15.14
JSD	<u>76.67</u>	**56.67**	<u>43.33</u>	92.77	<u>75.14</u>	14.14
MAD-OPD	73.33	50.00	40.00	87.95	70.52	16.66
W2S-OPD	<u>76.67</u>	46.67	36.67	<u>93.98</u>	72.83	**25.49**
cOPD（我们的方法）	**80.00**	**56.67**	**46.67**	**95.18**	**77.46**	<u>23.82</u>
Qwen3-4B / 0.6B / 1.7B						
Base	33.33	30.00	36.67	74.70	53.18	11.63
强教师	46.67	46.67	53.33	83.13	65.32	20.60
弱教师	13.33	13.33	10.00	56.63	33.53	11.34
FKL	<u>46.67</u>	33.33	26.67	<u>78.31</u>	56.07	<u>16.13</u>
RKL	36.67	<u>36.67</u>	20.00	77.11	53.18	15.92
JSD	43.33	30.00	<u>30.00</u>	<u>78.31</u>	55.49	15.77
MAD-OPD	36.67	23.33	16.67	75.90	49.71	15.95
W2S-OPD	40.00	<u>36.67</u>	<u>30.00</u>	**81.93**	<u>57.80</u>	16.02
cOPD（本文方法）	**63.33**	**50.00**	**36.67**	**81.93**	**65.32**	**16.41**

在推理组件中，m_v^+ 和 m_v^- 分别作为良好和不良推理选择的选择性代理。*cOPD*损失利用这些代理，鼓励将概率质量分配给选定的良好推理候选项，并惩罚分配给选定的不良推理候选项的概率质量。

证明。 对于 $v \in \text{supp}(G)$，不相交支撑假设给出 $p_s(v) - p_{ref}(v) = (\beta_s - \beta_{ref})G(v)$ 以及 $p_{ref}(v) - p_w(v) = (\beta_{ref} - \beta_w)G(v)$。两个间隔均为正，因此得到 $p_s(v) > p_{ref}(v) > p_w(v)$，而负分支仍处于非激活状态。由于 $\gamma_i = 1 - \alpha - \beta_i$，对$B$ 应用相同的论证，但教师排序反转，从而得到 $p_w(v) > p_{ref}(v) > p_s(v)$ 以及 $m_v^+ = 0$。教师对数比在 G 上为正、在 B上为负，因此在损失中产生相应的强化和抑制方向。

概率间隔 δ 进一步限制了所选候选项，但不会改变这些方向性结论。该分析说明，即使教师模型为良好推理分配的概率质量较低，仍然可以提供有信息量的对比参考。cOPD 利用这一额外参考，从强教师模型和弱教师的相对预测中选择监督信号。

4 实验

我们通过使用弱教师作为对比参照，评估 cOPD 是否能够改进在线策略蒸馏。我们的实验涵盖数学推理、指令遵循，以及带有拒答机制的问答。教师—学生配置覆盖不同的后训练背景、参数规模和模型系列。

4.1 实验设置

我们采用三种模型配置，以下按强教师、弱教师和学生模型的顺序列出：

1. JustRL-DeepSeek-1.5B（He 等人，2025年）、Qwen2.5-Math-1.5B（Yang 等人，2024年）和 DeepSeek-R1-Distill-Qwen-1.5B（DeepSeek-AI，2025年）。这些模型均以 Qwen2.5-1.5B 为基座模型，但后训练方式各不相同。2. Qwen3-4B、Qwen3-0.6B 和 Qwen3-1.7B（Yang 等，2025年）。这些模型属于同一系列，但参数规模不同。3. Llama-3.1-8B-Instruct、Llama-3.2-1B-Instruct 和 Llama-3.2-3B-Instruct。这些模型涵盖了 Llama 系列的不同版本和参数规模。

The 各任务所用的模型配置，以及相应的训练和评估 n
设 置，将在以下小节中介绍。
我们将 cOPD 与六种基线方法进行比较：初始学生模型（基座模型）、采用正向 KL 散度（FKL）、反向KL散度（RKL）或 Jensen–Shannon 散度（JSD）的 OPD、多教师辩论 OPD（MAD- OPD）（Wang 等人，2026a），以及弱到强多教师 OPD（W2S-OPD）（Yu 等，2026）。所有模型均在非思考模式下运行，详细的训练超参数见附录 A。

4.2 数学推理

设置与评估。 我们使用 DAPO-Math-17k（Yu 等人，2025）和三种模型配置进行训练。由于在数学推理任务中，教师模型与学生模型之间未呈现明显的性能差距，因此我们排除了 Llama 配置。对于每种配置，我们保留这样的提示：弱教师在八个样本中没有生成任何正确响应，而强教师至少生成了一个正确响应。经过筛选，JustRL 和 Qwen3 配置分别得到 6,016 和 7,488 条训练提示。我们使用 Pass@16 在 AIME24（Zhang 和 Math-AI，2024年）、AIME25（Zhang 和 Math-AI，2025年）、AIME26（Zhang 和 Math-AI，2026 年）和 AMC23 上进行评估，并使用 Pass@1 在 Omni-Math （ Gao 等人，2025 ）上进行评估。

结果。 我们在表 1 中报告数学推理结果。在两种配置下，cOPD 在所有蒸馏方法中都取得了最高的 Pass@16，使用 JustRL 时达到 77.46%，使用 Qwen3 时达到 65.32%。这些结果比最强基线方法高出最多 7.51 个百分点。更重要的是，在 JustRL 配置下，cOPD 超越了强教师；在 Qwen3 配置下，cOPD 的表现与强教师持平。在 Omni-Math 上，使用 Qwen3 时，cOPD 取得了最佳学生性能；使用 JustRL 时，其排名仅次于 W2S-OPD。
我们进一步分析图3(a)中 JustRL 配置下作为一种失败模式的截断。弱教师频繁出现截断错误且准确率较低，而强教师的准确率更高，截断错误也更少。仅强教师 OPD 传递了这种有利行为，使学生模型的截断错误从 321 降至 54–65，同时提高了准确率。

然而，纳入弱教师并不能始终保留这些优势。MAD-OPD 出现了 318 次截断错误，准确率并未高于初始学生模型。这表明，在辩论过程中，响应经常被截断的弱教师可能只能提供有限的有效反馈，从而限制学生模型的提升。W2S-OPD 的准确率与强教师相当，但截断错误多于单教师 OPD，这表明纳入弱教师可能会削弱对截断失败的抑制。

cOPD 将截断错误减少至 56 次，同时取得最高准确率，甚至超过强教师。这些结果表明，cOPD 能够利用弱教师提供的互补信息来提升性能，同时抑制其失败模式。

4.3 指令遵循

设置与评估。 我们使用多约束 IF-RLVR（Pyatkin 等，2025）数据集，每个提示最多包含五项可验证约束。我们留出一组固定的 7,000 条提示用于测试。对于每种模型配置，我们保留弱教师模型无法给出任何正确响应、而强教师模型在八次采样中至少给出一次正确响应的训练提示，因此 Qwen3 和 Llama 分别得到 6,892 条和 6,903 条训练提示。

表2：指令遵循（IF）和带弃答机制的问答（QA w/ Abstention）结果。所有分数均为百分比，Abst. 表示正确弃答率。粗体表示每列中学生模型的最佳结果，下划线表示次佳结果。

	Qwen3-4B/ 0.6B/ 1.7B				*Llama-3.1-8B/ 3.2-1B/ 3.2-3B*			
	IF	带弃答的问答			IF	带弃答的问答		
方法	准确率↑	EM↑	F1↑	拒答↑	准确率↑	EM↑	F1↑	拒答↑
Base	20.36	60.14	65.64	55.68	17.81	54.20	60.82	46.24
强教师	27.13	68.35	74.39	62.32	18.74	67.07	71.75	69.50
弱教师	18.33	35.01	39.80	30.63	11.69	16.63	21.23	8.53
FKL	23.20	<u>71.43</u>	<u>76.49</u>	<u>69.86</u>	19.35	65.71	70.05	68.23
RKL	<u>23.46</u>	67.75	73.33	63.13	19.11	66.43	70.65	68.73
JSD	22.48	70.04	75.40	68.33	18.66	<u>67.36</u>	<u>71.75</u>	<u>70.58</u>
MAD-OPD	20.10	63.18	69.97	57.78	16.07	48.75	51.10	64.44
W2S-OPD	22.81	66.17	71.28	59.60	<u>19.77</u>	3.40†	4.01†	1.95†
cOPD（我们的方法）	**24.26**	**72.09**	**77.31**	**71.59**	**21.73**	**68.97**	**73.17**	**73.04**

† 对于 Llama QA，W2S-OPD 出现了导致验证器失败的灾难性格式错误。忽略该错误后，EM、F1值和正确弃答率分别为 34.69、41.27 和 13.29。

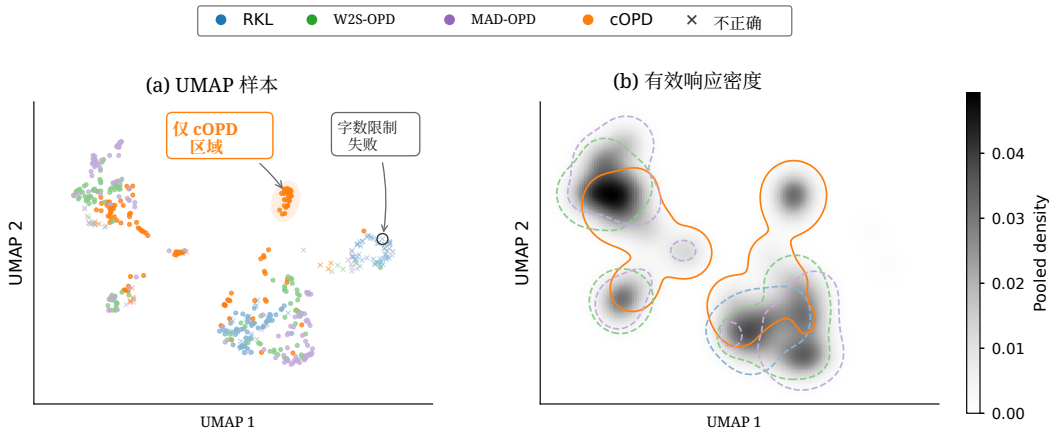

图2：共享UMAP空间中的指令遵循响应分布和约束满足情况，其中每个点代表一个响应，叉号表示违规。cOPD 在更多区域生成有效响应，同时基本避开了明显的字数限制失败区域。

结果。 评估结果见表2。两种模型配置下，cOPD 的表现均始终优于所有蒸馏基线，Qwen3 的准确率达到 24.26%，Llama 则达到 21.73%。这些结果分别比各配置下最强的基线方法——Qwen3 的 RKL 和 Llama 的 W2S-OPD——高出 0.80 和 1.96 个百分点。此外，在 Llama 上，cOPD 的表现比强教师高出 3.00 个百分点。

除准确率结果外，我们还在图2中，将使用相同提示词生成的响应映射到共享UMAP空间中进行可视化。我们的方法覆盖了包含有效基线响应的主要区域，同时还触及了基线样本中不存在的独特正确响应簇，展现出其在满足所有指令约束的同时探索额外区域的能力。与此同时，cOPD在很大程度上避开了显著的字数限制失败区域，证实了我们的假设：cOPD能够抑制有害响应模式。我们在附录C中提供了具有代表性的响应摘录。

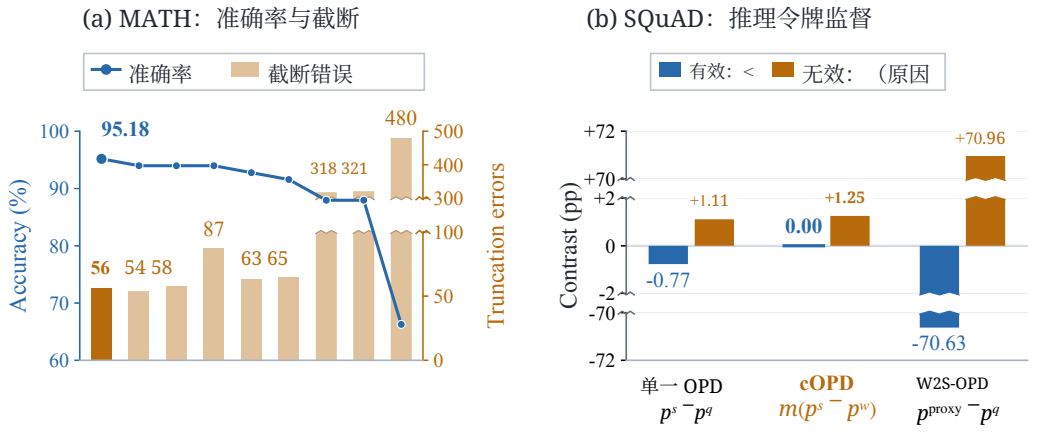

图3：**失败模式分析**。(a) 数学推理准确率（Pass@16）和截断错误计数。cOPD 在超越强教师模型准确率的同时，抑制了截断失败。(b) 对有效和无效推理格式的开头标记进行监督。W2S-OPD 过度放大了无效标记的相对权重，而 cOPD 中对应的差距可以忽略不计。

4.4 带弃答的问答

设置与评估。我们使用SQuAD2.0（Rajpurkar 等，2018），其中包含可回答问题和不可回答问题。给定上下文和问题后，模型会被指示根据上下文作答；当上下文不支持某个答案时，则应拒答。在训练集划分中，我们保留这样的提示：弱教师在八个样本中未生成任何正确响应，而强教师至少生成了一个正确响应。经过筛选后，Qwen3和Llama分别得到8,080和16,359个训练提示，其中不可回答问题分别占训练子集的26.51%和49.91%。我们在包含11,873个问题的完整官方开发集上进行评估。

结果。 我们在表 2 中报告了带弃答的问答结果。在两种配置下，cOPD 在精确匹配、F1值和正确弃答率方面均优于所有蒸馏基线，甚至超过了强教师。Qwen3 上，cOPD 的正确弃答率达到 71.59%，比最强基线高出 1.73 个百分点；Llama 上则达到 73.04%，比最强基线高出 2.46 个百分点。

我们进一步研究了 Llama 配置下 W2S-OPD 在 SQuAD 上的灾难性退化。我们观察到，训练后的模型错误地以 (reason 开始推理，而不是以 ＜开始，因此违反了所要求的 <reasoning> 格式。为考察这一行为，我们抽取了 128 个训练提示，并计算初始学生模型和两位教师模型在第一个响应位置的分布。图3(b)比较了 OPD 的单个教师模型、W2S-OPD 的代理教师以及学生模型之间，这两个词元的教师模型与学生模型概率差。对于 cOPD，我们报告了公式 2 和 3 中的掩码概率对比 $m_v($ $p_s - p_w)$，将其作为诊断性监督系数。

W2S-OPD 过度放大了无效标记（reason）的相对权重，使其平均概率从强教师下的 1.25% 升至代理教师下的 71.10%。这一失真的目标解释了观察到的格式错误。

相比之下，cOPD 使用边际门控掩码，有选择地应用对比式监督。对于无效标记，其估计的监督系数仍然很小，没有出现 W2S-OPD 中观察到的极端放大。这些结果表明，cOPD 能够利用弱教师提供的信息，同时对失败模式保持稳健。

4.5 结果讨论

我们的实证证据可以对提出研究问题作出肯定回答：**当一个表现逊于学生模型的模型与更强的教师模型搭配使用，作为对比参考时，它仍能提供有益的监督**。在数学推理、指令遵循和带弃答机制的问答任务中，cOPD 在评估的在线策略蒸馏方法中取得了最佳或并列最佳的表现。这些提升涵盖不同的模型系列、参数规模和后训练背景，证明弱教师对比在所评估的各种设置中具有价值。

这些优势还体现在抑制不理想的响应模式上。在数学推理任务中，cOPD 在所分析的配置中取得了最高准确率，同时大幅减少了截断错误（图3(a)）。在指令遵循任务中，它基本避免了与字数限制违规相关的显著失败区域（图2）。在 SQuAD2.0 上，它避免了 Llama 配置下 W2S-OPD 出现的灾难性格式错误（图3(b)）。总体而言，这些结果表明，cOPD 能够利用弱教师的信息作为反例，从而避免弱教师的失败模式。

此外，指令遵循可视化结果表明，cOPD 在探索方面可能具有优势：它能够达到一个基线方法未能触及的、包含令人满意响应的独特区域（图2）。附录B中的消融实验展示了监督分支和数据集选择各自的贡献。

5 相关工作

5.1 在线策略蒸馏

OPD 在学生模型自身生成的轨迹上使用令牌级教师监督进行训练，减少训练与推理之间的分布不匹配（Song和Zheng，2026年；Agarwal 等，2024年；Kim & Rush，2016年）。后续研究改进了蒸馏目标，以考虑教师模型不确定性（Jin等人，2026年；Li等人，2026a），并考察了教师模型与学生模型之间的兼容性如何影响知识迁移（Niu et al., 2026年；Wang 等人，2026c；Liu et al., 2026年）。标准 OPD 对单个教师模型的依赖促使人们探索来自多个模型的互补监督。

5.2 多教师 OPD

多教师在线策略蒸馏方法大体遵循两种路径：教师路由和教师聚合。基于路由的方法通过将每条由学生生成的轨迹分配给领域专属教师模型，在问题级别保持单个教师模型的监督（Ma 等人，2026年；Gao 等人，2026年）。而基于聚合的方法则结合多个教师模型针对同一轨迹提供的监督（Wang 等人，2026a）。我们的工作属于多教师在线策略蒸馏的更广泛范畴，但关注的是一个不同的问题：弱于学生模型的模型如何提供有用的监督。

5.3 弱模型监督

先前的研究已经证明了弱监督在其他学习场景中的价值：在这些场景中，较强的模型使用由较弱模型生成的标签进行训练（Burns 等人，2023年；Yuan 等人，2026年），同时，拒绝响应也被用作偏好学习中的负例（Wang 等人，2026b）。

Direct-OPD（Feng 等，2026）将弱模型在强化学习前后的检查点之间的对数概率对比，作为对更强学生模型的词元级监督。W2S-OPD（Yu 等，2026）通过将正向模型与负向模型之间的 logit 差异加到学生模型的 logits 上，构造代理教师。我们的方法支持使用成对的强化学习前后检查点以外的教师组合，而 cOPD 不会构造代理教师。相反，它将参考学生作为锚点，通过三方概率比较来选择词元级监督。

6 结论

在这项工作中，我们研究了表现不如学生模型的模型是否仍能提供有益的监督。我们的发现支持肯定的答案：弱教师可以与更强的教师模型并用，充当富有信息的对比参考。基于这一发现，

cOPD 通过双分支监督，有选择地强化有益的 token 偏好，并抑制有害偏好。在三项任务和多种模型配置上的实验表明，cOPD 能够提升学生性能，同时抑制不理想的响应模式。进一步证据表明，弱教师对比可能会鼓励模型探索独特且有效的响应。

我们认为，标准的基于 KL 的 OPD 并未明确验证教师模型的偏好是否能够改善任务结果，因此，匹配教师分布未必能带来最有益的训练方向。这促使我们从原始教师监督中选择有用的信号。我们将 cOPD 视为实现这一目标的一步：利用弱教师对比作为互补信息，以确定应当强化或抑制哪些令牌偏好。未来的工作可以将 cOPD 扩展到多个教师模型，将强教师的指导与不同弱教师提供的互补对比相结合，从而识别并抑制更广泛的失败模式。

人工智能使用声明

在本研究中，我们使用生成式人工智能工具协助实施所提出的方法。我们未使用生成式人工智能工具进行研究构思。作者人工审查了经人工智能辅助实现的内容，以确保其与预期方法一致，并在实验中检查了其行为。

我们还使用生成式人工智能工具协助润色和完善文字。这包括为初步草稿提供建议、提供不同的措辞选择和句式表达，以便更清晰地阐述我们的想法，以及进行语法检查。我们将人工智能建议的草稿视为参考，并逐句重写了稿件。最终措辞由作者确定。

We 已审阅所有经人工智能辅助完成的工作，并对本作品的最终内容承担责任，包括..... g 文本, 借助生成式 AI 生成的声明或成果。

可复现性声明

我们使用公开可用的大型语言模型和数据集进行训练与评估，并在附录 A 中列出了所有训练超参数。论文录用后，我们将公开实现代码。

参考文献

Rishabh Agarwal、Nino Vieillard、Yongchao Zhou、Piotr Stanczyk、Sabela Ramos Garea、Matthieu Geist 和 Olivier Bachem。语言模型的在线策略蒸馏：从自生成的错误中学习。载于《国际学习表征会议》，第 2024 卷，第 21246–21263 页，2024年。

Mohammadreza Armandpour、Fatih Ilhan、David Harrison、Ajay Jaiswal、Duc NM Hoang、Fartash Faghri、Yizhe Zhang、Minsik Cho 和 Mehrdad Farajtabar。揭示在线策略蒸馏：它何时有帮助、何时有害，以及原因。*arXiv* 预印本 *arXiv:2605.10889*，2026年。

Collin Burns、Pavel Izmailov、Jan Hendrik Kirchner、Bowen Baker、Leo Gao、Leopold Aschenbren-ner、Yining Chen、Adrien Ecoffet、Manas Joglekar、Jan Leike 等。弱到强泛化：利用弱监督激发强大能力。*arXiv* 预印本 *arXiv:2312.09390*，2023年。

DeepSeek-AI。Deepseek-r1：通过强化学习激励大语言模型的推理能力，2026年。URL https://arxiv.org/abs/2501.12948。

Shiyuan Feng、Huan-ang Gao、Haohan Chi、Hanlin Wu、Zhilong Zhang、Zheng Jiang、Bingxiang He、Wei-Ying Ma、Ya-Qin Zhang 和 Hao Zhou。通过直接在线策略蒸馏实现弱到强泛化。*arXiv* 预印本 *arXiv:2607.05394*，2026年。

Zixuan Fu、Bingxiang He、Yuxin Zuo、Haohuan Huang、Jinqian Zhang、Ruhang Xiao、Cheng Qian、Qinyu Luo、Huan-ang Gao、Yudong Wang 等。《重新思考大型语言模型的在线策略蒸馏 II：一个训练样本》。*arXiv* 预印本 *arXiv:2609.04172*，2026年。

Bofei Gao、Feifan Song、Zhe Yang、Zefan Cai、Yibo Miao、Qingxiu Dong、Lei Li、Chenghao Ma、Liang Chen、Zhengyang Tang 等。《Omni-Math：面向大型语言模型的通用奥林匹克级数学基准》。载于国际学习表征会议，第2025卷，第100540–100569页，2025年。Huan-ang Gao、Haohan Chi、Yong Yan、Shiyuan Feng、Hanlin Wu、Zheng Jiang、Bingxiang He、Wei-Ying Ma、Ya-Qin Zhang 和 Hao Zhou。《Open-MOPD：诊断并修复多教师在线策略蒸馏中的能力失衡》。*arXiv* 预印本 *arXiv:2608.19098*，2026年。Yuxian Gu、Li Dong、Furu Wei 和 Minlie Huang。《MiniLLM：大型语言模型的知识蒸馏》。载于国际学习表征会议，第2024卷，第32694–32717页，2024年。Bingxiang He、Zekai Qu、Zeyuan Liu、Yinghao Chen、Yuxin Zuo、Cheng Qian、Kaiyan Zhang、Weize Chen、Chaojun Xiao、Ganqu Cui 等。《JustRL：用简单的强化学习方案扩展一个15亿参数的大语言模型》。*arXiv* 预印本 *arXiv:2512.16649*，2025年。Woogyeol Jin、Taywon Min、Yongjin Yang、Dennis Wei、Yi Zhou、Swanand Ravindra Kadhe、Nathalie Baracaldo 和 Kimin Lee。《语言模型的熵感知在线策略蒸馏》。*arXiv* 预印本 *arXiv:2603.07079*，2026年。Yoon Kim 和 Alexander M Rush。《序列级知识蒸馏》。载于*2016*年自然语言处理经验方法会议论文集，第1317–1327页，2016年。Solomon Kullback 和 Richard A Leibler。《论信息与充分性》。数学统计年刊，22(1):79–86，1951年。Shizhen Li、Zhiyu Shen、Yuyin Lu、Yunhe Pang、Jielin Song、Yanghui Rao 和 Fu Lee Wang。《重新审视反向KL散度：自适应熵蒸馏》。*arXiv* 预印本 *arXiv:2608.14685*，2026a。Yaxuan Li、Yuxin Zuo、Bingxiang He、Jinqian Zhang、Chaojun Xiao、Cheng Qian、Tianyu Yu、Huan-ang Gao、Wenkai Yang、Zhiyuan Liu 等。《重新思考大型语言模型的在线策略蒸馏：现象学、机制与方案》。*arXiv* 预印本 *arXiv:2604.13016*，2026b。Xinyu Liu、Kechen Jiao、Chunyang Xiao、Runsong Zhao、Junhao Ruan、Bei Li、Jiahao Liu、Qifan Wang、Xin Chen、Jingang Wang 等。《在策略分歧较大的情况下，通过教师引导的策略优化进行在线策略推理蒸馏》。*arXiv* 预印本 *arXiv:2605.13230*，2026年。Wenhan Ma、Jianyu Wei、Liang Zhao、Hailin Zhang、Bangjun Xiao、Lei Li、Qibin Yang、Bofei Gao、Yudong Wang、Rang Li 等。《MOPD：用于大语言模型后训练中能力整合的多教师在线策略蒸馏》。*arXiv* 预印本 *arXiv:2606.30406*，2026年。Yifan Niu、Han Xiao、Dongyi Liu、Zelong Wang、Dihong Gong、Yasheng Wang 和 Jia Li。《打破分词器壁垒：跨模型系列的在线策略蒸馏》。*arXiv* 预印本 *arXiv:2606.09456*，2026年。Valentina Pyatkin、Saumya Malik、Victoria Graf、Hamish Ivison、Shengyi Huang、Pradeep Dasigi、Nathan Lambert 和 Hannaneh Hajishirzi。《可验证指令遵循的泛化》，2025年。Pranav Rajpurkar、Robin Jia 和 Percy Liang。《了解自己所不知道的：SQuAD 中的不可回答问题》。载于计算语言学协会第*56*届年会论文集（第*2*卷：短论文），第784–789页，2018年。Mingyang Song 和 Mao Zheng。《大型语言模型在线策略蒸馏综述》。*arXiv* 预印本 *arXiv:2604.00626*，2026年。Jianze Wang、Ying Liu、Jinlong Chen、Xuchun Hu、Qilong Zhang、Yu Cao、Jun Wang、Hua Yang、Yong Xie 和 Qianglong Chen。《MAD-OPD：通过多智能体辩论突破在线策略蒸馏的上限》，2026a。网址：https://arxiv.org/abs/2605.01347。Yifan Wang、Bolian Li、Junlin Wu、Zhaoxuan Tan、Zheli Liu、Ruqi Zhang、Ananth Grama 和 Qingkai Zeng。《Drift：从真实世界中大量的用户不满反馈中学习偏好》。载于国际学习表征会议，第2026卷，第120457–120479页，2026b。

Yuanyi Wang、Su Lu、Yanggan Gu、Pengkai Wang、Yifan Yang、Zhaoyi Yan、Congkai Xie、Jian- min Wu 和 Hongxia Yang。《并非所有分歧都可学习：在线策略蒸馏中的令牌可教性》。*arXiv* 预印本 *arXiv:2605.26844*，2026c。

Anyi Xu、Bangcai Lin、Bing Xue、Bingxuan Wang、Bingzheng Xu、Bochao Wu、Bowei Zhang、Chaofan Lin、Chen Dong、Chenchen Ling 等人。《Deepseek-v4：迈向高效的百万令牌上下文智能》。*arXiv* 预印本 *arXiv:2606.19348*，2026年。An Yang、Beichen Zhang、Binyuan Hui、Bofei Gao、Bowen Yu、Chengpeng Li、Dayiheng Liu、Jianhong Tu、Jingren Zhou、Junyang Lin、Keming Lu、Mingfeng Xue、Runji Lin、Tianyu Liu、Xingzhang Ren 和 Zhenru Zhang。《Qwen2.5-Math 技术报告：通过自我改进迈向数学专家模型》。*arXiv* 预印本 *arXiv:2409.12122*，2024年。

An Yang、Anfeng Li、Baosong Yang、Beichen Zhang、Binyuan Hui、Bo Zheng、Bowen Yu、ChangGao、ChengenHuang、Chenxu Lv 等。Qwen3 技术报告。*arXiv* 预印本 *arXiv:2505.09388*，2025年。

Fangxu Yu、Weijia Xu、Michael Xu、Tianyi Zhou 和 Zinan Lin。弱到强的在策略蒸馏，2026年。URL https://arxiv.org/abs/2607.26246。

Qiying Yu、Zheng Zhang、Ruofei Zhu、Yufeng Yuan、Xiaochen Zuo、Yu Yue、Weinan Dai、Tiantian Fan、Gaohong Liu、Lingjun Liu、Xin Liu、Haibin Lin、Zhiqi Lin、Bole Ma、Guang- ming Sheng、Yuxuan Tong、Chi Zhang、Mofan Zhang、Wang Zhang、Hang Zhu、Jinhua Zhu、Jiaze Chen、Jiangjie Chen、Chengyi Wang、Hongli Yu、Yuxuan Song、Xiangpeng Wei、Hao Zhou、Jingjing Liu、Wei-Ying Ma、Ya-Qin Zhang、Lin Yan、Mu Qiao、Yonghui Wu 和 Mingx- uan Wang。DAPO：面向大规模应用的开源大语言模型强化学习系统，2025年。URL https://arxiv.org/abs/2503.14476。

Yige Yuan、Teng Xiao、Shuchang Tao、Xue Wang、Jinyang Gao、Bolin Ding 和 Bingbing Xu。《从弱监督中激励强推理》。载于计算语言学协会欧洲分会第*19*届会议论文集（第*1*卷：长篇论文），第7138–7156页，2026年。

Aohan Zeng、Xin Lv、Zhenyu Hou、Zhengxiao Du、Qinkai Zheng、Bin Chen、Da Yin、Chendi Ge、Chenghua Huang、Chengxing Xie 等。《GLM-5：从氛围编程到智能体工程》。*arXiv*预印本 *arXiv:2602.15763*，2026年。

Yifan Zhang 和 Team Math-AI。2024年美国数学邀请赛（AIME），2024年。

Yifan Zhang 和 Team Math-AI。2025年美国数学邀请赛（AIME），2025年。

Yifan Zhang 和 Team Math-AI。《2026年美国数学邀请赛（AIME）》，2026年。

A 实现细节

表 3 总结了数学推理、指令遵循以及带弃答机制的问答任务的训练超参数。评估时，Pass@16 使用 0.7 的温度，Pass@1 使用 0.5 的温度，并采用 top-$p = 0.95$。对于指令遵循评估，我们会排除每种配置中任一参评模型产生无效评分器输出的提示；因此，在原有的 7,000 条留出提示中，Qwen3 和 Llama 分别得到包含 6,900 条和 6,906 条提示的共同评估子集。

B 消融研究

表4考察了 cOPD 的训练集选择和分支贡献。在完整的 DAPO-Math-17k 数据集上训练时，cOPD 的汇总 Pass@16（76.30%，而 RKL 和 JSD 均为 75.14%）和 Omni-Math Pass@1（24.14%，而 JSD 为 19.53%）均优于最强的单教师基线方法。在按教师划分的子集（包含 6,016 个提示）上训练时，汇总 Pass@16 达到 77.46%，Omni-Math Pass@1 达到

Tab表 3：三个任务的训练超参数。最大响应长度以 d in t词元为单位。

超参数	数学（DAPO）	IF	SQuAD 2.0
δ	0.05	0.05	0.05
τ	1.0	1.0	1.0
优化器	AdamW	AdamW	AdamW
学习率	1×10^{-6}	1×10^{-6}	1×10^{-6}
权重衰减	0.01	0.01	0.01
训练轮数	3	3	3
提示批次大小	64	64	64
每个提示的响应数	4	4	4
PPO小批次大小	64	64	64
训练采样温度	1.0	1.0	1.0
蒸馏支持集大小	16	16	16
最大响应长度	7,168	4,096	4,096

表4：数学推理中的训练数据与监督分支消融实验。筛选子集包含 6,016 条 DAPO-Math-17k 提示，与主实验保持一致。仅正向和仅负向变体分别只启用 m_v^+ 和 m_v^-。粗体和下划线标示各训练数据组中的最佳和次佳结果。

方法	AIME24 Pass@16 ↑	AIME25 Pass@16 ↑	AIME26 Pass@16 ↑	AMC23 Pass@16 ↑	Avg. Pass@16 ↑	Omni-Math Pass@1 ↑
完整 *DAPO-Math-17k*						
FKL	<u>76.67</u>	**60.00**	<u>36.67</u>	89.16	72.83	18.54
RKL	**80.00**	**60.00**	33.33	<u>93.98</u>	<u>75.14</u>	19.32
JSD	<u>76.67</u>	<u>53.33</u>	**43.33**	<u>93.98</u>	<u>75.14</u>	<u>19.53</u>
cOPD（我们的方法）	**80.00**	**60.00**	<u>36.67</u>	**95.18**	**76.30**	**24.14**
过滤后的 *DAPO-Math-17k* 子集（*6,016* 个提示）						
FKL	76.67	53.33	<u>43.33</u>	91.57	73.99	12.19
RKL	73.33	46.67	40.00	<u>93.98</u>	72.83	15.14
JSD	76.67	<u>56.67</u>	<u>43.33</u>	92.77	75.14	14.14
cOPD（仅 m_v^+）	**83.33**	**60.00**	<u>43.33</u>	92.77	<u>76.88</u>	20.95
cOPD（仅 m_v^-）	73.33	53.33	40.00	91.57	72.83	**26.76**
cOPD（我们的方法）	<u>80.00</u>	<u>56.67</u>	**46.67**	**95.18**	**77.46**	<u>23.82</u>

23.82%；尽管使用的学生训练提示更少，但这两项指标均超过了所有使用完整数据训练的单教师 OPD 基线。使用完整数据训练的 cOPD 仍具有竞争力，表明其性能并非仅限于筛选子集。在 cOPD 变体中，仅正向在所有四个 AIME/AMC Pass@16 基准测试上均优于仅负向，而仅负向在 Omni-Math Pass@1 上表现更好。结合两个分支的变体在这些变体中取得了最高的汇总 Pass@16，且 Omni-Math 得分居中，在两项评估指标上平衡了各自的优势。

C 指令遵循示例

我们展示了针对同一提示的简略响应。cOPD 和 RKL 示例分别对应图2中高亮的仅 cOPD 簇，以及圈出的字数限制违规。Markdown 标记按原样保留，[...] 表示省略的文本。

共享提示

法国的哪条大道是以维克多·雨果的名字重新命名的？请使用 Markdown 在回答中至少突出显示 7 个部分，例如，*突出显示的部分*。回答不得超过 606 个单词。字母 v 必须在您的响应中至少出现 13 次。

表5：cOPD 响应（满足所有指令约束）。

法国的这条大道——*维克多·雨果大道*——是以著名作家维克多·雨果的名字重新命名的。［...］*V* 是该大道名称中的重要元素，体现了人们对他的敬意。*V* 还出现在巴黎许多其他道路的名称中，例如 *V ́etheuil*、*Vieux-Paris* 和 *Ville-Victor*。［...］

表6：RKL 响应（违反字数限制）。

法国以维克多·雨果命名的大道是**维克多·雨果大道**。［...］
维克多·雨果的身影使这条大道成为法国文化与历史的象征。*重点标出的部分*［...］维克多·雨果的身影使这条大道成为法国文化与历史的象征。*重点标出的部分*［...］随着这位许多艺术家和作家参与其中，维克多·雨果大道
注：完整记录的响应包含大量重复内容，并以半句话结尾，原样保留如上。

表7：MAD-OPD 响应。

法国以维克多·雨果命名的大道是**维克多·雨果大道**。［...］*高亮部分* 维克多·雨果大道还经过**里沃利街**、**蓬佩街**和**蓬佩街**。［...］

表8：W2S-OPD 响应。

法国以维克多·雨果命名的大道是**维克多·雨果大道**。［...］*重点部分*这条大道还举办许多文化和艺术活动，例如定期举行的艺术展览和演出。［...］维克多·雨果大道中有13个字母**v**，是巴黎最重要的道路之一，与法国文化和历史密切相关。