

注意差距：数据改写以实现稳定的离策略监督微调

Shiwan Zhao, Xuyang Zhao, Jiaming Zhou, Aobo Kong, Qicheng Li, Yong Qin*

College of Computer Science, Nankai University

zhaosw@gmail.com

{xychao, zhoujiaming, kongaobo}@mail.nankai.edu.cn

{liqicheng, qinyong}@nankai.edu.cn

摘要

大型语言模型的监督微调 (SFT) 可以被视为一个离策略学习问题，其中专家演示来自固定的行为策略，而训练旨在优化目标策略。重要性采样是校正这种分布不匹配的标准工具，但大的策略差距会导致方差高和训练不稳定。现有方法使用 KL 惩罚或剪切来缓解此问题，这些方法被动地约束更新而不是积极减小差距。我们提出了一种简单有效的数据重写框架，通过将正确的解决方案作为在策略数据保留，并用引导重新求解错误的解决方案来主动缩小政策差距，仅在需要时回退到专家演示。这在校正优化前使训练分布与目标策略一致，减少了重要性采样的方差并稳定了离策略微调。在五个数学推理基准上的实验展示了相比原始 SFT 和最先进的动态微调 (DFT) 方法的一致性和显著增益。数据和代码将在 <https://github.com/NKU-HLT/Off-Policy-SFT> 发布。

1 介绍

大型语言模型 (LLMs) 在链式思维 (CoT) 推理 (Wei et al., 2022) 方面取得了显著进展，这主要得益于一个结合了监督微调 (SFT) 与强化学习 (RL) 的后训练流水线 (Shao et al., 2024; Lambert et al., 2024; Guo et al., 2025; Liu et al., 2025b)。SFT 从高质量演示中提炼出任务特定的推理行为，使基础模型能够快速适应新任务。RL 通过基于奖励的目标优化在线策略

滚动过程来补充这一点，在具有挑战性的推理基准测试上提供了持续改进。在广泛采用的先 SFT 再 RL 范式中，SFT 为推理提供了一个强大的初始化，随后 RL 通过在线采样对其进行进一步细化。

尽管它们有紧密的联系，SFT 和 RL 展示了互补的优势和局限性 (Ma et al., 2025; Yan et al., 2025)。SFT 简单且高效，能够通过整合外部专家知识和推理模式来扩展模型的推理能力边界。然而，它完全依赖于离策略数据进行操作，因为专家演示来自固定的行为策略而非不断演化的模型策略，从而导致众所周知的策略差距，并引起高方差、训练不稳定性和过拟合问题。相比之下，RL 执行的是在策略优化，因此完全避免了策略差距的问题，但它面临较高的样本和计算复杂度，并且只能改进模型现有的推理行为而无法引入根本上新的能力。在这项工作中，我们专注于改进 SFT 本身，为独立微调提供更稳定的基础，也为未来涉及 RL 或混合方法的扩展奠定基础。

从离策略学习 (Precup et al., 2000) 的角度来看，重要性采样 (IS) 是校正行为策略和目标策略之间分布不匹配的标准工具。然而，当策略差距变大时，IS 权重变得高度偏斜，导致方差放大和不稳定优化。现有的补救措施，如 KL 惩罚、信任区域或裁剪比率 (Schulman et al., 2015, 2017)，通过被动约束更新来稳定优化，但未能主动减少数据分布本身的基础差距。

我们提出了一种简单而有效的数据重写框

*Corresponding Author.

架,该框架在优化开始前主动减少策略差距。对于每个问题,我们首先从目标模型中抽取多个响应。如果任何响应正确解决了问题,则将其保留为在线策略数据。否则,我们将以真实解决方案作为参考提示模型重新解决问题,生成更好地反映目标策略的摘要与转述数据。如果自解和重解都失败了,我们会回退到原始专家演示。受真正理解是在学习者用自己的话重新表达解决方案而不是逐字复制这一直觉启发,我们的摘要与转述策略(见图 2)转换了 SFT 数据集而不仅仅是限制优化动力学。此过程使训练分布更紧密地与目标策略对齐,减少了重要性采样方差并稳定了离线策略微调,剩余的不匹配通过 IS 加权进一步缓解。

五项数学推理基准测试实验表明,我们的方法在性能上始终优于传统的 SFT 和最先进的动态微调(DFT)方法(Wu et al., 2025)。特别是,在 Qwen2.5-Math-7B 模型上,我们的方法将平均准确率从 23.23%提高到了 30.33%,比 SFT 提高了,从 36.61%提高到了 42.03%,比 DFT 提高了。

我们的贡献有三方面:

- 我们将 SFT 公式化为一个离策略学习问题,并识别出政策差距是基于 IS 的优化中不稳定性的关键来源。
- 我们引入了一个数据重写框架,该框架主动减少了数据层面的策略差距,使得低方差和稳定的离线策略微调成为可能。
- 我们在多个模型和基准测试中验证了该方法,展示了相对于标准 SFT 和 DFT 基线的一致性增益。

2 相关工作

2.1 以数据为中心的 SFT 改进措施

SFT 的质量很大程度上取决于指令数据集的构建,之前的工作探讨了三个方面:扩展性、多样性和质量。Flan(Chung et al., 2024; Longpre et al., 2023)扩大了指令调优任务的数

量,并证明更大的任务集合显著提高了性能。LIMA(Zhou et al., 2023)表明仅在精心挑选的 1,000 个样本上对强大的预训练模型进行微调即可实现具有竞争力的结果,强调了数据质量和多样性的关键作用。

除缩放外,若干研究将训练样本转换为更好地与目标模型分布对齐。GRAPE(Zhang et al., 2025a)在 SFT 训练前从多个大语言模型中选择具有最高目标模型概率的响应。自我蒸馏微调(SDFT)(Yang et al., 2024b)使用模型本身生成提炼数据以弥合分布差距。自我到监督微调(S3FT)(Gupta et al., 2025)识别正确的模型响应并在这些响应上进行微调,同时对剩余样本进行改写或保留正确答案。我们的数据重写框架遵循这一思路但采用离策略视角:它通过重写在数据层面积极减小策略差距,并通过训练期间的重要性采样进一步减轻残余不匹配。

2.2 结合 SFT 和 RL

另一条研究路线将 SFT 的能力扩展与 RL 的在线策略鲁棒性相结合。交错或统一的训练目标联合优化监督信号和强化信号(Ma et al., 2025; Liu et al., 2025a; Zhang et al., 2025b),而动态加权策略平衡 SFT 模仿学习与基于 RL 的偏好优化。尽管 RL 采用在线策略操作,但这些方法中的 SFT 阶段仍然依赖于静态离线数据集,无法解决潜在的数据分布不匹配问题。我们的方法是正交的:它不是修改训练目标或交错线上和线下回放,而是通过数据重写直接在数据层面减少优化前的策略差距,为现有的基于目标的方法提供了一个互补视角。

2.3 离策略学习

一种平行的研究视角将 SFT 视为一个离策略学习问题,其中专家演示和不断演化的目标策略引起了一个分布不匹配的问题。另一条研究路线引入重要性采样或奖励校正来纠正这种不匹配(Wu et al., 2025; Qin and Springenberg, 2025)。还有一条研究路线专注于通过优化级别的技术(如裁剪或信任区域)减少重要性采样

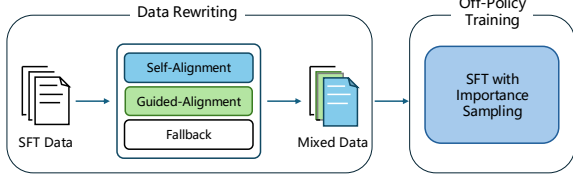


图 1: 整体框架包括 (一) 数据重写, 将 SFT 数据从离策略转换为更接近于在策略的分布, 以及 (二) 使用重要性采样的离策略训练, 进一步缓解剩余的策略差距。

的方差 (Zhu et al., 2025), 这些技术通过约束或重新加权更新来稳定训练。相比之下, 我们的数据重写框架在优化之前积极地使训练分布与目标策略对齐, 为这些优化级别方法提供了互补的视角。

3 方法

我们将监督微调 (SFT) 视为一个离策略学习问题, 并提出了一种统一框架 (图 1), 该框架结合了数据重写, 它在数据层面主动减少策略差距, 以及重要抽样 (IS), 它在优化过程中进一步校正剩余的不匹配。

3.1 SFT 作为离策略学习等价物

令 π_{sft} 表示生成 SFT 数据集的行为策略, 令 π_θ 表示由 θ 参数化的目标策略。SFT 的目标是在 π_θ 下最大化预期奖励:

$$J(\theta) = \mathbb{E}_{(x,y) \sim \pi_\theta} [r(x,y)], \quad (1)$$

其中 $r(x,y)$ 是特定任务的奖励信号 (例如, 正确性)。

然而, 由于训练数据来自 π_{sft} 而非 π_θ , SFT 成为一个离策略问题:

$$J(\theta) = \mathbb{E}_{(x,y) \sim \pi_{sft}} [w(x,y) r(x,y)], \quad (2)$$

其中重要性权重定义为 $w(x,y) = \frac{\pi_\theta(y|x)}{\pi_{sft}(y|x)}$ 。当分歧 $D(\pi_{sft} \parallel \pi_\theta)$ 较大时, 这些权重变得高度偏斜, 导致方差放大和不稳定优化。现有方法通过使用 KL 惩罚、信任区域或裁剪来缓解这一问题, 这被动地约束更新但不能主动减小潜在的策略差距。

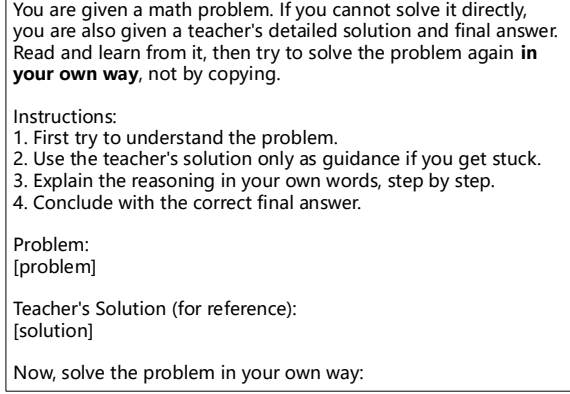


图 2: 摘要与转述提示提供了参考解决方案, 并要求模型用自己的话重新解决该问题。

3.2 数据重写作作为政策对齐

我们引入一个数据重写算子 $\mathcal{T}$, 它在训练前将 π_{sft} 转换为混合分布 π_{mix} :

$$\pi_{sft} \xrightarrow{\mathcal{T}} \pi_{mix} \quad \text{with} \quad D(\pi_{mix} \parallel \pi_\theta) < D(\pi_{sft} \parallel \pi_\theta). \quad (3)$$

操作符 $\mathcal{T}$ 应用一个三阶段对齐层次结构:

- **自对准:** 对每个输入 x , 我们从 π_θ 中采样多个响应。如果任何响应正确解决了问题, 我们随机保留一个正确的响应作为在线数据¹。
- **引导对齐:** 对于自对齐失败的输入, 我们提示 π_θ 使用参考解决方案生成摘要与转述响应, 这些响应是专家答案的同义转述, 而不是逐字复制 (参见图 2)。与自对齐一样, 我们采样多个响应, 并且如果其中任何一个正确, 则随机保留一个正确的响应作为重写数据。
- **备用方案:** 如果引导对齐也失败, 我们将回退到原始的专家演示。

所得数据集 $\mathcal{D}'$ 包含在线策略和重写示例的混合, 专家数据仅作为备选方案包含在内:

$$\mathcal{D}' = \mathcal{D}_{self} \cup \mathcal{D}_{retell} \cup \mathcal{D}_{expert}. \quad (4)$$

¹当存在多个正确响应时, 随机选择一个以公平地与标准 SFT 进行比较; 相同策略也适用于引导对齐阶段。

这一分层过程确保了 $\pi_{\text{mix}} = \mathcal{T}(\pi_{\text{sft}})$ 逐步将训练数据向目标策略转移，从而在优化开始前减少了策略差距。

3.3 重要性采样与对齐数据

尽管生成的数据集 $\mathcal{D}'$ 更加接近目标策略分布，但由于重写不完美和回退专家演示，仍可能存在残余偏差。即使重写完美，批量更新也可能引入策略差距，因为目标策略逐步更新²。为解决这一问题，我们在优化过程中应用重要性采样：

$$\mathcal{L}_{\text{IS}}(\theta) = \mathbb{E}_{(x, y') \sim \mathcal{D}'} \left[- \sum_{t=1}^{|y'|} w(x, y'_t) \cdot \log \pi_{\theta}(y'_t | x, y'_{<t}) \right], \quad (5)$$

其中

$$w(x, y'_t) = \text{sg} \left(\frac{\pi_{\theta}(y'_t | x, y'_{<t})}{\pi_{\text{mix}}(y'_t | x, y'_{<t})} \right)$$

是重要性权重， $\text{sg}(\cdot)$ 表示停止梯度算子，以防止梯度通过该权重本身流动。按照常见做法 (Wu et al., 2025; Zhang et al., 2025b)，我们近似分母 $\pi_{\text{mix}}(y'_t | x, y'_{<t}) \approx 1$ ，实际上将混合数据视为真实分布。

这种两级设计，通过重写结合数据级别对齐和通过 IS 结合优化级别校正，实现了低方差、无偏训练，并且作为动态微调 (DFT) (Wu et al., 2025) 等方法的补充。

4 实验

4.1 数据集和模型

遵循 DFT (Wu et al., 2025)，我们使用 NuminaMath CoT 数据集 (LI et al., 2024) 进行训练。原始数据集包含大约 860,000 道数学问题及其相应的解决方案。为了减少计算成本，我们随机抽取了 50,000 个实例，并在过滤掉过长的示例后保留了大约 48,000 个。

由于我们的方法需要模型生成数据重写的候选解决方案，我们实验了两种代表性基础架构：Qwen2.5-数学-7B (Yang et al., 2024a)，一个

没有明确指令调优的数学专业模型，和 llama-3.1-8B-Instruct (Grattafiori et al., 2024)，一个通用指令调优模型。这种比较使我们能够调查我们的方法是否可以使基本模型和指令调优模型都受益。

对于数据重写，我们在自我对齐和引导对齐阶段从目标模型中采样 10 个候选响应。两个模型的数据集统计信息见表 1。我们观察到 Qwen2.5-Math-7B 在自我对齐阶段解决了更多的问题，但在引导对齐阶段解决的问题较少，这可能是由于其相对于 Llama-3.1-8B-Instruct 较弱的指令跟随能力。

4.2 训练和评估详情

对于 SFT 训练，我们使用版本框架 (Sheng et al., 2025) 并采用 AdamW 优化器。学习率设置为 Qwen2.5-Math-7B 的 5×10^{-5} 和 Llama-3.1-8B-Instruct 的 7×10^{-6} 。训练进行一个周期，批量大小设为 256。我们采用余弦退火的学习率调度策略，并设置 0.1 的预热比例。

对于评估，我们遵循 DFT 并在五个广泛使用的基准上评估数学推理性能：Math500 (Hendrycks et al., 2021)、Minerva Math (Lewkowycz et al., 2022)、Olympiad-Bench (Investments, 2024)、AIME 2024 (American Institute of Mathematics, 2024) 和 AMC 2023 (Mathematical Association of America, 2023)。所有结果均报告为在温度设置为 1.0 的 16 次解码运行中的平均准确率。

4.3 主要结果

表 2 报告了五个数学推理基准测试的结果。我们的数据重写 (DR) 方法在 Qwen2.5-Math-7B 和 Llama-3.1-8B-Instruct 上始终改进了普通 SFT 和 DFT。

Qwen2.5-数学-7B。在这个数学专用模型上，标准 SFT 实现了平均准确率 23.23%，而 DFT 将其提升至 36.61%。引入 DR 带来了显著的增益，DR+SFT 将性能提高到 30.33%，而 DR+DFT 进一步将其推高至 42.03% 平均

²我们将每批数据的在线重写留作未来的工作。

	自对准	导向对齐	备用方案	总计
Qwen2.5-数学-7B	28,752	11,620	7,634	48,006
llama-3.1-8B-指令	26,947	16,335	4,719	48,001

表 1: 不同模型在对齐阶段的数据集统计（实例数量）。

	数学 500	密涅瓦	数学竞赛	AIME24	AMC23	平均值
Qwen2.5-数学-7B	39.90	14.43	17.16	7.50	29.38	21.67
+ SFT	52.61	19.13	17.32	2.06	25.00	23.23
+ DFT	68.70	31.92	32.31	6.68	43.44	36.61
+ DR + SFT (ours)	59.85	21.14	23.54	8.54	38.59	30.33
+ DR + DFT (ours)	70.40	34.85	36.12	14.58	54.22	42.03
Llama-3.1-8B-Instruct	36.18	16.01	9.52	0.83	14.53	15.41
+ SFT	28.71	11.23	6.26	0.41	10.31	11.39
+ DFT	46.5	24.11	15.65	3.95	22.50	22.54
+ DR + SFT (ours)	44.38	19.21	13.07	1.87	17.19	19.14
+ DR + DFT (ours)	47.91	24.72	16.52	4.99	26.09	24.05

表 2: 数学推理基准测试的平均准确率（%）。我们的方法结合 SFT 或 DFT，始终能比相应的基线方法提升性能。DR 代表数据重写。

准确率。值得注意的是，DR+DFT 在所有五个基准测试中都取得了最高分，特别是在 AIME2024（6.68% → 14.58%）和 AMC2023（43.44% → 54.22%）上有了显著提高。

Llama-3.1-8B-Instruct. 在此指令调整模型上，DR 将 SFT 从 11.39% 提升到 19.14%，并将 DFT 从 22.54% 提升到 24.05%。虽然收益保持一致，但其幅度小于 Qwen2.5-Math-7B 上的幅度，这表明指令调整模型在数据重写方面的受益较少。

为了更好地理解整体性能提升以及 Qwen2.5-Math-7B 和 LLaMA-3.1-8B-Instruct 之间的性能差距，我们分析了自我对齐、指导对齐和回退子集的平均对数概率（表 3，图 3）。分析表明，重写有效缩小了策略差距，因为重写后的响应始终实现了更高的平均对数概率（更不负面），这表明与目标政策更加一致。此外，三个子集的 SFT 数据揭示了问题难度的

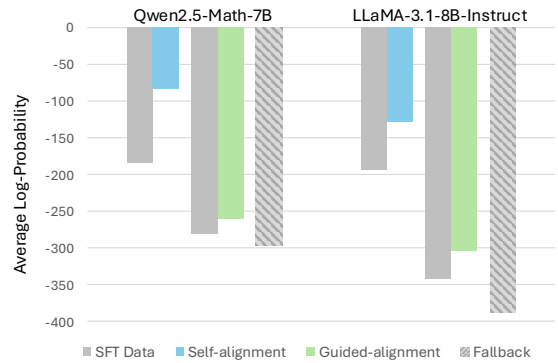


图 3: 两种模型在自对齐、引导对齐和回退子集上的平均对数概率。SFT 列表示每个阶段的原始 SFT 数据。在回退阶段，回退子集与 SFT 数据一致。

增加，从自我对齐到指导对齐再到回退，对数概率逐渐降低；摘要与转述策略仍然减轻了在较难问题上的策略差距，尽管效果不如自我解决。最后，指令微调限制了数据重写的效益：LLaMA-3.1-8B-Instruct 在整个子集中始终表现出比 Qwen2.5-Math-7B 更低的对数概率，这表

模型	自对准		导向对齐		备用方案
	SFT	Rewriting	SFT	Rewriting	SFT
Qwen2.5-数学-7B	-184.44	-83.36	-280.64	-259.44	-296.03
llama-3.1-8B-指令	-193.01	-127.67	-341.57	-303.44	-388.17

表 3: 模型响应在自我对齐、引导对齐和回退子集中的平均对数概率。较高值（不太负）表示响应更接近目标策略分布。

Method	Math500	密涅瓦	奥林匹克竞赛	AIME24	AMC23	平均值
DFT	68.70	31.92	32.31	6.68	43.44	36.61
DFT + Self-Alignment	67.86	32.35	32.79	10.19	50.94	38.83
DFT + Guided-Alignment	70.06	32.57	32.16	8.54	48.44	38.36
DFT + Full DR	70.40	34.85	36.12	14.58	54.22	42.03

表 4: 关于 Qwen2.5-Math-7B 的消融实验与 DFT。自我对齐替换仅正确的模型生成响应，引导对齐替换仅消化重述响应，而完全 DR 结合两者。

明指令微调使模型偏向于通用的指令跟随，并减少了缩小策略差距的空间。

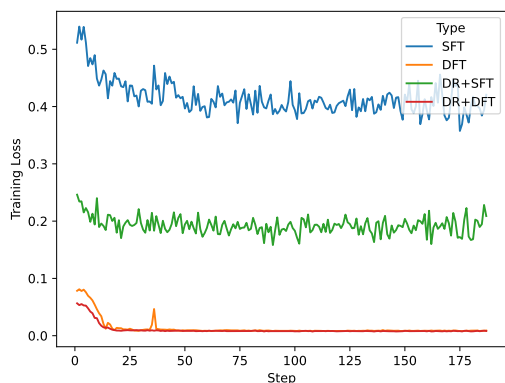
训练动态。图 4 展示了 Qwen2.5-Math-7B 和 LLaMA-3.1-8B-Instruct 两种模型在不同方法下的训练损失曲线。在这两个模型中，DFT 和 DR+DFT 比 SFT 和 DR+SFT 收敛得快得多，在最初的 40 – 50 步内就达到了接近零的训练损失，这突显了动态微调提供的强大监督信号和优化稳定性。数据重写在两个模型上都一致地降低了原始 SFT 的损失，证实了在优化前将训练分布与目标策略对齐可以减少方差并稳定训练。然而，对于两个模型而言，DR+SFT 停留在比基于 DFT 的方法更高的损失水平，这表明没有动态在线更新时残余分布不匹配仍然存在。值得注意的是，LLaMA-3.1-8B-Instruct 在所有方法上都表现出比 Qwen2.5-Math-7B 更高的训练损失，说明了训练数据与指令微调模型之间的分布差异更大，这解释了它从数据重写中获得的性能提升较小。最后，DR+DFT 结合了这两种方法的优点，实现了最低的最终损失和最稳定的收敛性，这也解释了其在表 2 中所有基准测试中的优越表现。

4.4 消融研究

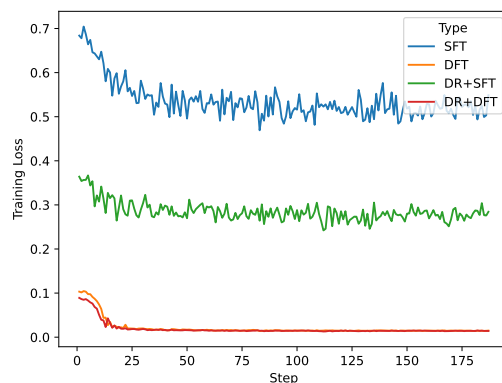
表 4 报告了在 Qwen2.5-Math-7B 上使用 DFT 基线的消融实验结果。将自对准纳入 DFT 产生了适度的收益，通过用正确的模型生成响应替换原始 SFT 数据，平均准确率从 36.61% 提高到 38.83%。添加了引导对齐，它只对较难的问题使用消化和重述策略进行改写，而保持自解题不变，也提升了性能至 38.36%，表明指导性改写在自我解决失败时有效地缓解了政策差距。扩展到完全 DR，该方法结合了自我对齐和指导对齐，实现了最佳结果，平均准确率提升至 42.03%，并在所有基准测试中均带来了一致的收益。这些发现强调了自我对齐与指导对齐之间的互补作用，以及在离策略优化之前积极调整训练数据以缩小政策差距并增强推理能力的重要性。

5 结论

我们提出了一种简单而有效的数据重写框架，用于大型语言模型的监督微调，将 SFT 表述为一个离策略学习问题。我们的方法在训练前主动减小了数据层面的策略差距，并通过优



(a) Qwen2.5-数学-7B



(b) llama-3.1-8B-Instruct

图 4: Qwen2.5-Math-7B 和 Llama-3.1-8B-Instruct 在 SFT、DFT 及其与数据重写 (DR) 组合下的训练损失曲线。DR+DFT 达到了最低的最终损失和最稳定的收敛。

化过程中的重要性采样进一步缓解了残余不匹配。在多个数学推理基准上的广泛实验表明，与传统的 SFT 和最先进的动态微调相比，性能持续提高，尤其是在基础模型上观察到了最显著的进步。这些发现强调了以数据为中心的策略对于稳定和增强大型语言模型离策略微调的价值。

限制

虽然我们的实验展示了数据重写在稳定非策略监督微调方面的有效性，但仍存在一些限制。首先，我们的评估仅限于一组有限的模型，主要是在中等参数规模下，因此评估其对更大和更多样化模型的应用性将留待未来的工作。其次，我们完全专注于数学推理基准测试；将该方法扩展到更广泛的领域，包括医疗保健和金融等行业设置，是重要的下一步。第三，我们的方法采用单轮离线重写策略，而更为复杂或在线的方法——例如，为减轻训练期间的策略变化而每批进行一次重写——可以进一步提高稳定性和性能。最后，探索更丰富的重写技术，如利用来自更先进模型的外部知识，代表了另一个有希望的方向。

References

American Institute of Mathematics. 2024. [AIME 2024](#)

Competition Mathematical Problems.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, and 1 others. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Sonam Gupta, Yatin Nandwani, Asaf Yehudai, Dinesh Khandelwal, Dinesh Raghu, and Sachindra Joshi. 2025. [Selective self-to-supervised fine-tuning for generalization in large language models](#). In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 6240–6249, Albuquerque, New Mexico. Association for Computational Linguistics.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*.

XTX Investments. 2024. Ai mathematical olympiad - progress prize 1. <https://kaggle.com/competitions/ai-mathematical-olympiad-prize>. Kaggle.

- Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, and 1 others. 2024. Tulu 3: Pushing frontiers in open language model post-training. *arXiv preprint arXiv:2411.15124*.
- Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, and 1 others. 2022. Solving quantitative reasoning problems with language models. *Advances in neural information processing systems*, 35:3843–3857.
- Jia LI, Edward Beeching, Lewis Tunstall, Ben Lipkin, Roman Soletskyi, Shengyi Costa Huang, Kashif Rasul, Longhui Yu, Albert Jiang, Ziju Shen, Zihan Qin, Bin Dong, Li Zhou, Yann Fleureau, Guillaume Lample, and Stanislas Polu. 2024. Numinamath. [<https://huggingface.co/AI-MO/NuminaMath-CoT>](https://github.com/project-numina/aimo-progress-prize/blob/main/report/numina_dataset.pdf).
- Mingyang Liu, Gabriele Farina, and Asuman Ozdaglar. 2025a. Uft: Unifying supervised and reinforcement fine-tuning. *arXiv preprint arXiv:2505.16984*.
- Zihan Liu, Zhuolin Yang, Yang Chen, Chankyu Lee, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. 2025b. Acereason-nemotron 1.1: Advancing math and code reasoning through sft and rl synergy. *arXiv preprint arXiv:2506.13284*.
- Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V Le, Barret Zoph, Jason Wei, and 1 others. 2023. The flan collection: Designing data and methods for effective instruction tuning. In *International Conference on Machine Learning*, pages 22631–22648. PMLR.
- Lu Ma, Hao Liang, Meiyi Qiang, Lexiang Tang, Xiaochen Ma, Zhen Hao Wong, Junbo Niu, Chengyu Shen, Runming He, Bin Cui, and 1 others. 2025. Learning what reinforcement learning can’t: Interleaved online fine-tuning for hardest questions. *arXiv preprint arXiv:2506.07527*.
- Mathematical Association of America. 2023. [AMC 2023 Competition Problems](#).
- Doina Precup, Richard S. Sutton, and Satinder P. Singh. 2000. Eligibility traces for off-policy policy evaluation. In *Proceedings of the Seventeenth International Conference on Machine Learning, ICML ’00*, page 759 – 766, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.
- Chongli Qin and Jost Tobias Springenberg. 2025. Supervised fine tuning on curated data is reinforcement learning (and can be improved). *arXiv preprint arXiv:2507.12856*.
- John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. 2015. Trust region policy optimization. In *International conference on machine learning*, pages 1889–1897. PMLR.
- John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.
- Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. 2025. Hybridflow: A flexible and efficient rlhf framework. In *Proceedings of the Twentieth European Conference on Computer Systems*, pages 1279–1297.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.
- Yongliang Wu, Yizhou Zhou, Zhou Ziheng, Yingzhe Peng, Xinyu Ye, Xinting Hu, Wenbo Zhu, Lu Qi, Ming-Hsuan Yang, and Xu Yang. 2025. On the generalization of sft: A reinforcement learning perspective with reward rectification. *arXiv preprint arXiv:2508.05629*.
- Jianhao Yan, Yafu Li, Zican Hu, Zhi Wang, Ganqu Cui, Xiaoye Qu, Yu Cheng, and Yue Zhang. 2025. Learning to reason under off-policy guidance. *arXiv preprint arXiv:2504.14945*.
- An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, and 1 others. 2024a. Qwen2. 5-math technical report: Toward mathematical expert model via self-improvement. *arXiv preprint arXiv:2409.12122*.
- Zhaorui Yang, Tianyu Pang, Haozhe Feng, Han Wang, Wei Chen, Minfeng Zhu, and Qian Liu. 2024b. [Self-distillation bridges distribution gap in language model fine-tuning](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1028–1043, Bangkok, Thailand. Association for Computational Linguistics.
- Dylan Zhang, Qirun Dai, and Hao Peng. 2025a. The best instruction-tuning data are those that fit. *arXiv preprint arXiv:2502.04194*.
- Wenhao Zhang, Yuexiang Xie, Yuchang Sun, Yanxi Chen, Guoyin Wang, Yaliang Li, Bolin Ding, and

Jingren Zhou. 2025b. On-policy rl meets off-policy experts: Harmonizing supervised fine-tuning and reinforcement learning via dynamic weighting. *arXiv preprint arXiv:2508.11408*.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, and 1 others. 2023. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36:55006–55021.

Wenhong Zhu, Ruobing Xie, Rui Wang, Xingwu Sun, Di Wang, and Pengfei Liu. 2025. Proximal supervised fine-tuning. *arXiv preprint arXiv:2508.17784*.