Phi-4-Mini-推理：探索小型推理语言模型在数学中的极限

Haoran Xu‡ Baolin Peng‡ Hany Awadalla Dongdong Chen Yen-Chun Chen Mei Gao
Young Jin Kim Yunsheng Li Liliang Ren Yelong Shen Shuohang Wang Weijian Xu
Jianfeng Gao Weizhu Chen

Microsoft

摘要

思维链（CoT）通过训练大型语言模型（LLMs）显式生成中间推理步骤，显著增强了它们的正式推理能力。虽然 LLMs 很容易从这些技术中受益，但由于其有限的模型容量，在小型语言模型（SLMs）中提高推理能力仍然具有挑战性。Deepseek-R1(Luo et al., 2025) 的最近研究工作表明，来自 LLM 生成的合成数据的知识蒸馏可以显著提升 SLM 的推理能力。然而，详细的建模配方并未公开披露。在这项工作中，我们提出了一种系统性的训练方法，该方法包含四个步骤：（1）大规模中间阶段训练多样化的蒸馏长 CoT 数据，（2）基于高质量长 CoT 数据的监督微调，（3）利用精心策划的偏好数据集进行 Rollout DPO，以及（4）使用可验证奖励的强化学习（RL）。我们将该方法应用于 Phi-4-Mini ，一个紧凑的具有 38 亿参数的模型。结果得到的 Phi-4-迷你推理模型在数学推理任务上超过了更大规模的推理模型，例如，在 Math-500 上的表现优于 DeepSeek-R1-Distill-Qwen-7B 3.2 分，比 DeepSeek-R1-Distill-Llama-8B 高出 7.7 分。我们的结果验证了，通过大规模高质量 CoT 数据精心设计的训练配方能够有效解锁即使在资源受限的小模型中也具备强大的推理能力。

‡ 同等贡献。除第一和最后两位作者外，其余作者按字母顺序排列。

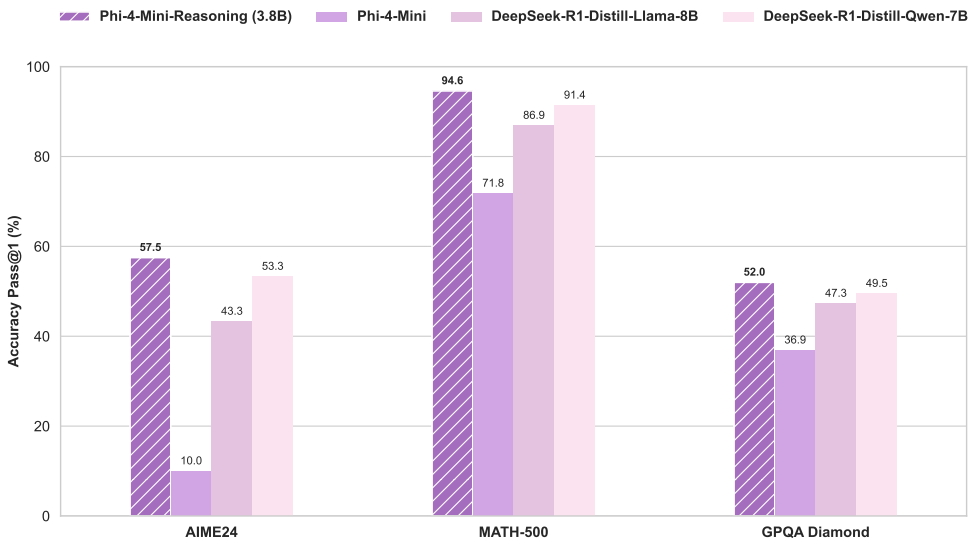


图 1: Phi-4-Mini-Reasoning 的数学基准性能。

1 介绍

大型语言模型 (LLMs) 在众多自然语言处理任务中展现了卓越的能力，但当面对复杂多步骤问题时，它们的推理能力往往会出现下降，此时简单地输出答案而不提供中间步骤会导致显著的性能差距 (Wei et al., 2022)。链式思维（CoT）方法通过明确提示模型生成最终答案前的一系列逻辑步骤来应对这一挑战，从而极大地提升了其推理能力 (Kojima et al., 2022; Wei et al., 2022)。在推断过程中纳入这种推理过程确立了测试时间缩放的范例，这进一步提高了复杂推理任务的表现 (Snell et al., 2024; Welleck et al., 2024; OpenAI, 2024)。

提高推理能力对大型 LLMs 来说本就较为容易，因为它们具有广泛的容量，而对于小型语言模型 (SLMs)，这仍然是一个挑战。幸运的是，Deepseek-R1(Guo et al., 2025) 表明，非 logits 级蒸馏——实际上是对更强大的模型生成的合成数据进行有效监督微调（SFT）SLMs——可以显著提高 SLM 推理性能。例如，这种做法可以使 Llama-8B(Grattafiori et al., 2024) 的 MATH-500(Lightman et al., 2023) 准确率从 44.4%提升到 89.1%。在此突破之后，包括 Bespoke-Stratos-7B(Labs, 2025) 和 OpenThinker-7B(OpenThoughts, 2025) 在内的诸多努力都旨在复制这些结果。尽管有这种热情，关于训练的主要焦点的辩论仍然存在。Deepscaler(Luo et al., 2025) 建议将 RL 的缩放方式类比于 GRPO(Shao et al., 2024) 以提升推理性能，而 S1 和 LIMO(Muennighoff et al., 2025; Ye et al., 2025b) 则强调了推理数据集的质量和多样性，揭示即使是少于 1K 示例的数据集也能增强推理表现。

与其专注于单独能够提升训练的技术，我们系统地探索了一种专门针对 SLMs 的训练范式，在这种范式中，有限的模型容量使得推理改进尤为具有挑战性。我们的方法包括两个阶段的知识蒸馏，接着是基于展开的偏好学习，该过程也重新使用了错误的 LLM 生成样本，并最终通过可验证奖励来进行强化学习。最初，我们采用知识蒸馏作为中期训练机制以嵌入基础推理能力。然后我们在微调阶段再次应用知识蒸馏以进一步提升模型泛化能力。在进行 LLM 展开采样用于蒸馏时，通常会丢弃一些错误的输出；

然而，我们将这些被丢弃的样本重新利用来创建定制化的偏好数据集，该数据集用于在蒸馏后的模型上实施偏好学习。最后，我们使用基于最终答案正确性的可验证奖励信号来进行强化学习微调。为了确保训练稳定，我们引入了几项针对性改进，包括提示优化、通过过采样和过滤进行的奖励再平衡以及探索过程中的温度退火。

我们通过使用紧凑的 38 亿参数模型 Phi-4-Mini (Microsoft et al., 2025) 进行验证，结果得到了 Phi-4-迷你推理，其性能几乎比其他两倍大小的推理模型（如 DeepSeek-R1-Distill-Qwen-7B 和 DeepSeek-R1-Distill-Llama-8B）高出近一倍。

2 背景

小型语言模型(SLMs)已经展示了强大的推理能力的巨大潜力。例如，Qwen-1.5B 仅通过从 DeepSeek-R1(Guo et al., 2025) 中蒸馏 80 万示例就能在 Math-500(Lightman et al., 2023) 基准测试中达到 83.9%的准确率。蒸馏已经作为增强 SLMs 推理能力的强大工具出现；然而，针对小型模型的最佳蒸馏策略仍然研究不足。最近的研究提供了补充见解：Luo et al. (2025) 指出通过强化学习逐渐增加生成长度可以进一步改进蒸馏后的模型，而 Muennighoff et al. (2025) 和 Ye et al. (2025b) 强调数据的多样性和质量，而不是数量本身，对于成功至关重要。尽管有这些进展，但对于 SLMs 的有效蒸馏配方仍缺乏全面理解。此外，盲目应用孤立的技术可能会导致性能下降。例如，直接将 S1K 数据集 (Muennighoff et al., 2025) 或 LIMO 数据集 (Ye et al., 2025b) 蒸馏到 Phi-4-Mini 上会导致推理性能显著下降。这一观察结果表明，由于其容量有限，SLM 需要比其较大的同类模型采用更加精心设计的数据和训练策略来发展强大的推理能力。详细的结果如表 1 所示。

模型	**AIME 2024**	**MATH-500**	**GPQA 钻石**
Phi-4-Mini	10.0	71.8	36.9
Phi-4-Mini + LIMO	6.7	57.8	24.8
Phi-4-Mini + S1K	3.0	47.0	26.3
Phi-4-Mini-Reasoning (with our full recipe)	**57.5**	**94.6**	**52.0**

表 1: Phi-4-Mini 在不同蒸馏设置下的 Pass@1 性能。简单地使用少量高质量数据会导致性能显著下降，突显了全面训练配方的必要性。

因此，我们的目标是开发一套全面且高效的训练 SLMs 的方案。我们首先观察到，非推理型 SLMs 需要一个初始中期训练阶段来吸收大量的推理轨迹，然后再应用任何额外的技术。然而，有几个关键问题仍然存在：需要多少中期训练数据？接下来应该采用哪些后续技术——如仔细蒸馏、偏好学习或强化学习？在这项工作中，我们系统地解决了这些问题，并提出了构建高性能推理 SLMs 的完整方案。

3 多阶段持续训练用于推理

这里，我们系统地介绍了我们的完整训练方案，而不是探索各个单独的组件。以一个预训练的语言模型作为基础，我们通过首先使用精心挑选的 CoT 推理数据集进行多阶段连续训练，然后再运行带有可验证奖励的强化学习来提高其形式化推理能力。

3.1 蒸馏作为中期训练

在第一阶段，我们将蒸馏视为中期训练的一部分。具体来说，我们在广泛的合成链式思维（CoT）数据语料库上对基础模型进行下一步标记预测训练，这些数据涵盖了来自不同领域的各种难度级别问题。这些问题的 CoT 风格答案由 Deepseek-R1 模型 (Guo et al., 2025) 采样得出，之后我们使用拒绝抽样来保留正确的答案。关于我们数据生成方法的更多细节，请参见第 4 节。我们将每个问题与其对应的正确 CoT 答案配对，并使用标准因果语言建模目标训练基础模型。我们在打包模式下训练模型，即在同一输入序列中打包多个短示例以提高训练效率。此次中期训练的目的是让小型基础模型具备一般的链式思维推理能力，而这些能力在模型中期训练时并未显式学习。我们发现允许中期训练迭代使用尽可能多的 CoT 训练数据直到验证集上的模型性能饱和是有效的。

3.2 蒸馏作为有监督微调

在学习了广泛且多样的推理链之后，我们的下一步是从中期训练数据集中选择一个紧凑但具有代表性的子集进行后续微调。微调是在非包装模式下进行的，在该模式中我们教导模型决定在哪里停止生成。由于高质量的数据可以显著提高模型性能和泛化能力，并使模型能够更好地回答复杂问题 (Xu et al., 2024a; Zhou et al., 2023; Ye et al., 2025b; Muennighoff et al., 2025)，我们构建了一个涵盖不同数学领域的综合数据集，难度水平超过“大学水平”。关于数据分类的更多细节请参见第 4 节。

3.3 展开偏好学习

在前两个阶段，模型仅接受生成的数据进行训练，过滤掉包含错误答案的回放。然而，被拒绝的回放是否完全没有价值？在这个阶段，我们使用被拒绝的回放来提升模型性能。如 Xu et al. (2024b) 所指出的，拒绝数据的质量对于偏好学习很重要。具体来说，与正确答案相比带有细微差别的不正确回答是构建有效偏好对的好候选。为了确保数据质量，我们保留了被 GPT-4o-mini(Achiam et al., 2023) 划分为“高中”级别或以上数学类的问题。然后通过将每个问题的正确答案指定为首选回放，错误答案指定为非优选回放来构建偏好数据集。最后，我们将直接偏好优化（DPO）(Rafailov et al., 2023) 应用于模型：

$$J_{\text{DPO}}(\pi_\theta;\pi_{\text{ref}}) = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\Big[\log\sigma\Big[\beta\log\frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \beta\log\frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)}\Big[\Big[, \tag{1}$$

其中 π_{ref} 是参考模型，y_w 和 y_l 分别是优选和非优选的展开。

3.4 具有可验证奖励的强化学习

尽管 DPO 利用精心策划的偏好对提升了模型的对齐和推理能力，但作为一种使用固定数据集的离线学习方法，DPO 存在局限性。为了通过在线学习提升模型的推理能力，我们在经过蒸馏和偏好训练的模型上执行 RL。接下来，我们将描述我们实验过的 RL 算法及 RL 训练配方。

近端策略优化（PPO） PPO (Schulman et al., 2017) 已成功应用于通过 RLHF 对 LLMs 进行微调。该算法采用剪辑代理目标来限制每次策略更新，使其保持接近之前的策略。这种剪辑机制避免了过大的重要性采样比率，既稳定了学习又提高了样本效率。PPO 力求最大化

$$J_{\mathrm{PPO}}(\theta) = \mathbb{E}_{q\sim\mathcal{D},\, o_{\le t}\sim\pi_{\theta_{\mathrm{old}}}(\cdot|q)}\Bigg[\min\Big[r_t(\theta)\,\widehat{A}_t,\ \mathrm{clip}\big[r_t(\theta),\,1-\epsilon,\,1+\epsilon\big[\,\widehat{A}_t\Big[\Bigg[, \tag{2}$$

其中

$$r_t(\theta) = \frac{\pi_\theta(o_t \mid q, o_{<t})}{\pi_{\theta_{\mathrm{old}}}(o_t \mid q, o_{<t})},$$

且 q 从数据分布 $\mathcal{D}$ 中采样，ϵ 控制剪切范围，$\widehat{A}_t$ 表示时间步 t 的优势估计。为了计算 $\widehat{A}_t$，PPO 使用了广义优势估计器（GAE）(Schulman et al., 2015)。给定一个价值函数 V 和一个奖励函数 R，该估计器是

$$\widehat{A}_t^{\mathrm{GAE}(\gamma,\lambda)} = \sum_{l=0}^{\infty} (\gamma\,\lambda)^l\,\delta_{t+l}, \tag{3}$$

其中时间差分项为

$$\delta_l = R_l + \gamma\,V(s_{l+1}) - V(s_l), \quad 0 \le \gamma, \lambda \le 1. \tag{4}$$

基于群组的相对策略优化（GRPO） GRPO (Shao et al., 2024) 通过比较一批 G 模型响应中的奖励来估计其基线，减少了批评家的成本并提高了模型训练的稳定性。具体来说，对于每个问题 q，它在旧策略 $\pi_{\theta_{\mathrm{old}}}$ 下采样一组候选响应 G $\{o_i\}_{i=1}^{G}$ ，然后计算它们的奖励 $\{R_i\}_{i=1}^{G}$。标准化优势计算为

$$A_i \;=\; \frac{R_i - \mathrm{mean}(R_1,\ldots,R_G)}{\mathrm{std}(R_1,\ldots,R_G)}, \tag{5}$$

GRPO 然后最大化一个裁剪的代理目标函数，在组上取平均值，并附加一个指向参考策略 π_{ref} 的 KL 惩罚：

$$J_{\mathrm{GRPO}}(\theta) = \mathbb{E}_{q\sim\mathcal{D},\,\{o_i\}\sim\pi_{\theta_{\mathrm{old}}}(\cdot|q)}\Bigg[\frac{1}{G}\sum_{i=1}^{G} \min\Big[r_i(\theta)\,A_i,\ \mathrm{clip}\big[r_i(\theta),\,1-\epsilon,\,1+\epsilon\big[\,A_i\Big[\; -\; \beta\,D_{KL}\big[\pi_\theta\|\pi_{\mathrm{ref}}\big[\Bigg[, \tag{6}$$

其中 ϵ 是裁剪参数，β 用于加权 KL 惩罚。

可验证奖励 强化学习与可验证奖励(RLVR)已被证明在训练用于各种推理任务的模型方面非常有效 (Guo et al., 2025; Ye et al., 2025a)。根据先前的研究 (Guo et al., 2025)，一个可验证任务的奖励被定义为模型最终答案准确性的函数。具体来说，

$$R(\hat{y}, y) = \begin{cases} +1, & \text{if verify}(\hat{y}, y), \\ -1, & \text{otherwise.} \end{cases} \tag{7}$$

其中 y 表示真实答案，$\hat{y}$ 表示模型的响应。

我们的 RL 配方 在我们应用 GRPO 训练基础模型的初步研究中，观察到了三个影响模型训练稳定性和有效性的问题。

1. **响应长度的高方差**尽管基础模型在中期训练后已经能够生成合理的 CoT 响应，但我们观察到在同一 GRPO 采样组内响应长度存在显著差异。对于相同的提示，正向奖励的响应范围从大约 12k 到 20k 个令牌不等。直接针对这种长度异质性响应优化标准 GRPO 目标模型会导致不稳定。Zhang et al. (2025) 报告了在数学和编码任务上训练模型时出现类似现象。

2. **消失的梯度在均匀奖励下** GRPO 对优势估计的依赖使其容易受到所有组内采样响应获得相同奖励而导致梯度消失问题的影响，从而导致回报方差为零。DAPO 框架 (Yu et al.) 通过过采样并过滤掉响应准确率为确切 0 或 1 的提示来解决此问题，从而保留非零优势信号。然而，我们在实验中发现，在将 DAPO 应用于我们的模型时需要解决以下两个问题：

   (i) 该模型对组内长度差异敏感：具有中等准确度（例如，0.1 或 0.9）的响应仍然会因响应长度的变化而引发不稳定的梯度幅度。

   (ii) 对于困难的数学任务，获得哪怕一个正向奖励样本（通过提示模型）需要将 GRPO 批次大小扩展到 128。这种正面和负面训练信号之间的不平衡阻碍了 RL 的收敛。

   我们假设这些问题在小型语言模型中更为突出，因为与大型语言模型相比，其强化学习稳定性更容易脆弱。

3. **探索-开发权衡**有效探索对于在强化学习中发现高回报策略至关重要。虽然使用 1.0 或更高的采样温度来鼓励探索，但通常会采用较低的温度（例如 0.6）来限制数学和编程任务中的输出方差。在我们的实验中，我们观察到这种训练期间使用的探索与评估时应用的利用设置之间的分歧导致了显著的性能差距。

为了解决上述挑战，我们引入了一组方法来提高 RL 训练的稳定性和有效性：

1. **提示优化**我们使用多个候选提示进行多轮采样，这些提示旨在使用蒸馏模型进行 RL 训练。然后仅保留那些生成的响应具有相对均匀标记长度的提示。该方法减轻了由于组内响应长度方差较高而在 GRPO 优化过程中产生的不稳定性。

2. **奖励再平衡通过过采样和过滤**受 DAPO (Yu et al.) 的启发，对于难以处理的提示，我们首先进行过采样以确保响应组中具有足够的多样性。然后通过保留所有正奖励响应并随机抽取相同数量的负奖励响应来重新平衡该组。为进一步减少长度变化，并避免过于简单的提示导致的不稳定，我们会过滤掉那些组级准确性超过某个阈值（例如 50%）的提示。

3. **温度退火**为了在模型训练过程中寻求探索与开发之间的最佳平衡，我们引入了温度退火策略。我们将采样温度初始化为 1.0，并在线性衰减的第一个 50%的训练步骤中将其降至 0.6。对于剩余的训练步骤，温度固定为 0.6。这一策略鼓励在强化学习早期阶段进行更广泛的探索，同时逐渐转向已知状态-动作子空间中的开发。

Data Resource	Size	Reasoning
AquaRAT (Ling et al., 2017)	98K	✘
Ape210K (Zhao et al., 2020)	210K	✘
MetaMathQA (Yu et al., 2023)	395K	✘
MathInstruct (Yue et al., 2023)	262K	✘
TAL-SCQ5K (TAL-SCQ5K, 2023)	5K	✘
OpenR1-Math (HuggingFace, 2025)	220K	✔
Bespoke-Stratos-17k (Labs, 2025)	17K	✔
OpenThoughts-114K (OpenThoughts, 2025)	114K	✔

表 2: 用于构建推理数据集的数据资源概述。对于非推理数据，我们仅使用来自 Deepseek R1 的问题和示例答案。

4 合成 CoT 数据生成

为了支持蒸馏和基于展开的偏好学习，我们构建了一个由大型语言模型生成的合成推理轨迹组成的大型推理数据集。具体来说，我们将多个公开数据集——例如 Bespoke (Labs, 2025), Openthoughts (OpenThoughts, 2025) 和 OpenR1-Math (HuggingFace, 2025)——与几个内部种子数据集汇总在一起。对于已经包含推理轨迹的数据集，我们直接使用提供的标注。对于缺乏此类轨迹的数据集，我们只保留数学问题，并使用 DeepSeek-R1 (671B) 生成新的思路答案。每个问题大约采样八次展开。表 2 提供了数据来源的概览。总共，我们在 160 万个样本中收集了约 1000 万次展开，包括来自公开数据集的贡献。对于可验证的数学问题，我们首先应用一个数学验证工具来评估答案的正确性。然而，由于自动验证有时无法有效验证复杂的解决方案——导致假阴性结果——我们还使用 GPT-4o-mini 对最初标记为错误的展开进行重新验证。为了保持数据集平衡，我们将每个数据样本标注上包括领域类别、难度级别和重复模式存在与否等属性。领域类别涵盖了广泛的范围，如代数、几何、理论、概率和微积分。难度级别分为小学、初中、高中、大学和研究生水平。中期训练阶段利用整个数据集，而后续的训练步骤则操作选定的数据子集。

5 实验

5.1 评估

我们在三个数学推理任务上评估了我们的模型：AIME24(MAA, 2024)、Math-500(Lightman et al., 2023) 和 GPQA Diamond(Rein et al.)。在评估过程中，生成参数设置为温度 0.6，top_p 为 0.95，并且最大序列长度为 32K。对于每个任务，我们进行了 3 次运行并报告了这些试验的平均性能。

5.2 基线

我们将我们的 Phi-4-迷你推理 模型与 o1-mini 和几个领先的开源小型推理模型进行了比较，包括 DeepSeek-R1-Distill-Llama-8B(Guo et al., 2025)、Bespoke-Stratos-7B(Labs, 2025) 以及 OpenThinker-7B(OpenThoughts, 2025)。

5.3 训练设置

对于前两个蒸馏阶段，我们使用批次大小为 128，学习率为 1e-5，总共 5 个训练周期，预热比例为 0.1。在第一阶段，序列长度设置为 16K 与打包策略，而在第二阶段序列长度延长至 20K 没有打包。对于 Rollout DPO 阶段，我们使用学习率为 5e-7 进行单个训练周期，序列长度为 16K。在 RL 阶段，采用学习率 5e-7 和序列长度 25k 来鼓励模型探索。

5.4 结果

总体结果呈现在表 3 中。尽管只有 38 亿个参数，菲-4-迷你推理 的表现优于所有开源基础模型，包括那些几乎是其两倍大小的模型。此外，我们提供了一个消融研究来展示每个训练阶段对 phi-4-微型推理 性能的贡献。

5.5 消融研究

在本节中，我们进行消融研究以了解我们的蒸馏训练对模型推理能力的影响，并比较我们的 RL 配方与 DAPO 的训练稳定性。

为了测量大语言模型的推理边界，我们使用 pass@k 指标。对于每个问题，我们从模型中抽取 k 个输出。如果在 k 个样本中有至少一个通过验证，则该问题的 pass@k 值为 1；否则为 0。数据集上的平均 pass@k 反映了模型能在 k 次尝试内解决的问题的比例。如图 2a 所示，我们的蒸馏管道作为向模型注入推理相关知识的有效方法。蒸馏阶段之后，pass@k 分数显著提高，表明蒸馏成功扩展了基础大语言模型的推理能力边界。这为后续的 RL 训练奠定了坚实的基础。在此基础上，RL 微调进一步提升了性能，平均额外提高了约 7 个点，并进一步完善了模型的能力。

我们还将我们的强化学习训练方法与 DAPO 进行了比较。如图 2b 所示，在我们的设置下，DAPO 表现不佳：随着训练的进行，AIME 数据集上的共识@16 指标持续下降。相比之下，我们的强化学习训练技术表现出更大的稳定性，并且始终能在基础模型上带来有意义的改进。

Model	AIME	MATH-500	GPQA Diamond
o1-mini*	63.6	90.0	60.0
DeepSeek-R1-Distill-Qwen-7B	53.3	91.4	49.5
DeepSeek-R1-Distill-Llama-8B	43.3	86.9	47.3
Bespoke-Stratos-7B*	20.0	82.0	37.8
OpenThinker-7B*	31.3	83.0	42.4
Llama-3.2-3B-Instruct	6.7	44.4	25.3
Phi-4-Mini	10.0	71.8	36.9
+ Distill Mid-training	30.0	82.9	42.6
+ Distill Fine-tuning	43.3	89.3	48.3
+ Roll-Out DPO	50.0	93.6	49.0
+ RL (Phi-4-迷你推理)	**57.5**	**94.6**	**52.0**

表 3: Pass@1 CoT 推理结果与更大的 7B 推理模型和 OpenAI 模型相比的 Phi-4-迷你推理 。星号 (*) 表示直接来自已发布的报告的结果，而其余结果是在我们的工作中复现的。

5.6 安全声明

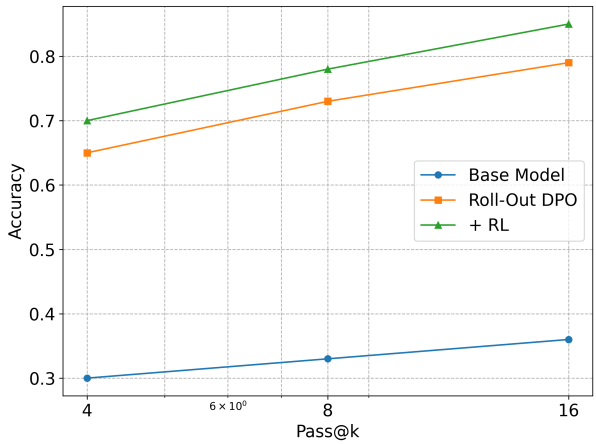
Phi-4-Mini-Reasoning 是根据微软的负责任 AI 原则开发的。该模型的响应潜在安全风险使用 Azure AI Foundry 的风险和安全性评估框架进行了评估，重点关注有害内容、直接越狱和模型基础性。Phi-4-Mini-Reasoning 模型卡片包含了有关我们对安全性和负责任 AI 考虑的方法的附加信息，开发者在使用该模型时应了解这些信息。

6 结论

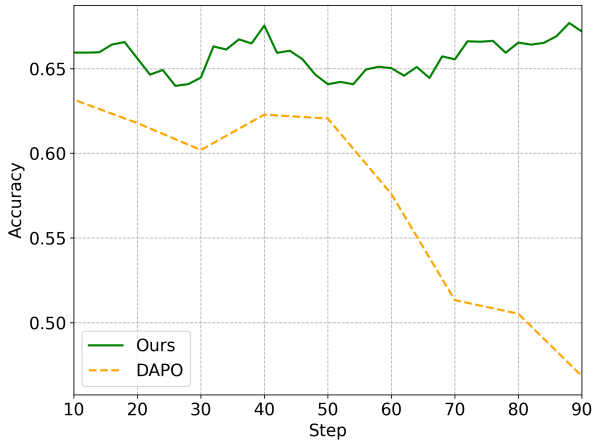
我们提出了一种多阶段训练范式，以增强小型语言模型（SLMs）的推理能力，结合了大规模蒸馏、展开偏好学习和具有可验证奖励的强化学习。应用于 Phi-4-Mini，我们的方法产生了菲-4-迷你推理 ，这是一个紧凑的 38 亿参数模型，在推理能力上几乎超过了其规模近两倍的开源模型。我们展示了精心协调的训练阶段序列对于解锁 SLMs 的强大推理能力至关重要。我们的结果显示，当使用精心选择的数据和训练策略进行训练时，小型模型可以匹配备受或甚至超过大型模型的能力。我们认为这项工作为在资源受限的情况下开发高效、高性能的模型提供了蓝图。

致谢

我们向 Amit Garg、Daniel Perez-Becker、Nguyen Bach、Tetyana Sych 和整个 GenAI 团队表示最深切的感谢，感谢他们对本工作的宝贵贡献。他们在模型审查、部署和产品化方面的支持对于完成这个项目至关重要。我们还非常感激图灵团队在技术合作和深入讨论中的持续帮助。



(a)



(b)

图 2: (a) 基础模型、Roll-Out DPO 模型和附加强化学习训练的模型在 AIME 2024 上的 Pass@k 曲线。Rollout DPO 显著提高了 Pass@k，扩展了模型的推理能力。进一步的强化学习训练带来了额外的提升。(b) DAPO 和我们的强化学习训练方法之间的比较，通过 AIME 2024 上的 cons@16 准确率进行评估。我们的强化学习训练方法展示了更好的稳定性。

参考文献

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 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, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

HuggingFace. 2025. Open r1: A fully open reproduction of deepseek-r1.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Bespoke Labs. 2025. Bespoke-stratos: The unreasonable effectiveness of reasoning distillation. https://www.bespokelabs.ai/blog/bespoke-stratos-the-unreasonable-effectiveness-of-reasoning-distillation. Accessed: 2025-01-22.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let's verify step by step. *arXiv preprint arXiv:2305.20050*.

Wang Ling, Dani Yogatama, Chris Dyer, and Phil Blunsom. 2017. Program induction by rationale generation: Learning to solve and explain algebraic word problems. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 158–167, Vancouver, Canada. Association for Computational Linguistics.

Michael Luo, Sijun Tan, Justin Wong, Xiaoxiang Shi, William Y. Tang, Manan Roongta, Colin Cai, Jeffrey Luo, Tianjun Zhang, Li Erran Li, Raluca Ada Popa, and Ion Stoica. 2025. Deepscaler: Surpassing o1-preview with a 1.5b model by scaling rl. https://pretty-radio-b75.notion.site/DeepScaleR-Surpassing-O1-Preview-with-a-1-5B-Model-by-Scaling-RL-19681902c1468005bed8ca303013a4e2. Notion Blog.

MAA. 2024. American invitational mathematics examination–aime. In American Invitational Mathematics Examination–AIME 2024.

Microsoft, Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, et al. 2025. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. *arXiv preprint arXiv:2503.01743*.

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. 2025. s1: Simple test-time scaling. *arXiv preprint arXiv:2501.19393*.

OpenAI. 2024. Learning to reason with llms. *https://openai.com/index/learning-to-reason-with-llms*.

OpenThoughts. 2025. Open Thoughts. https://open-thoughts.ai.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36:53728–53741.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. Gpqa: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*.

John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. 2015. High-dimensional continuous control using generalized advantage estimation. *arXiv preprint arXiv:1506.02438*.

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, Y Wu, et al. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. 2024. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*.

TAL-SCQ5K. 2023. TAL-SCQ5K: A high-quality mathematical competition dataset in english and chinese. Accessed: 2025-04-26.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Sean Welleck, Amanda Bertsch, Matthew Finlayson, Hailey Schoelkopf, Alex Xie, Graham Neubig, Ilia Kulikov, and Zaid Harchaoui. 2024. From decoding to meta-generation: Inference-time algorithms for large language models. *arXiv preprint arXiv:2406.16838*.

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2024a. A paradigm shift in machine translation: Boosting translation performance of large language models. In *The Twelfth International Conference on Learning Representations*.

Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Benjamin Van Durme, Kenton Murray, and Young Jin Kim. 2024b. Contrastive preference optimization: Pushing the boundaries of LLM performance in machine translation. In *Forty-first International Conference on Machine Learning*.

Guanghao Ye, Khiem Duc Pham, Xinzhi Zhang, Sivakanth Gopi, Baolin Peng, Beibin Li, Janardhan Kulkarni, and Huseyin A Inan. 2025a. On the emergence of thinking in llms i: Searching for the right intuition. *arXiv preprint arXiv:2502.06773*.

Yixin Ye, Zhen Huang, Yang Xiao, Ethan Chern, Shijie Xia, and Pengfei Liu. 2025b. Limo: Less is more for reasoning. *arXiv preprint arXiv:2502.03387*.

Longhui Yu, Weisen Jiang, Han Shi, Jincheng Yu, Zhengying Liu, Yu Zhang, James T Kwok, Zhenguo Li, Adrian Weller, and Weiyang Liu. 2023. Metamath: Bootstrap your own mathematical questions for large language models. *arXiv preprint arXiv:2309.12284*.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, et al. Dapo: An open-source llm reinforcement learning system at scale, 2025. *URL https://arxiv. org/abs/2503.14476*.

Xiang Yue, Xingwei Qu, Ge Zhang, Yao Fu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. 2023. Mammoth: Building math generalist models through hybrid instruction tuning. *arXiv preprint arXiv:2309.05653*.

Xiaojiang Zhang, Jinghui Wang, Zifei Cheng, Wenhao Zhuang, Zheng Lin, Minglei Zhang, Shaojie Wang, Yinghan Cui, Chao Wang, Junyi Peng, et al. 2025. Srpo: A cross-domain implementation of large-scale reinforcement learning on llm. *arXiv preprint arXiv:2504.14286*.

Wei Zhao, Mingyue Shang, Yang Liu, Liang Wang, and Jingming Liu. 2020. Ape210k: A large-scale and template-rich dataset of math word problems. *arXiv preprint arXiv:2009.11506*.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, LILI YU, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. 2023.

LIMA: Less is more for alignment. In *Thirty-seventh Conference on Neural Information Processing Systems*.