强化预训练

Qingxiu Dong[*†‡] **Li Dong**[*†]
Yao Tang[†] **Tianzhu Ye**[†§] **Yutao Sun**[†§] **Zhifang Sui**[‡] **Furu Wei**[†◇]
[†] Microsoft Research
[‡] Peking University
[§] Tsinghua University
https://aka.ms/GeneralAI

在这项工作中，我们介绍了强化预训练（RPT）作为大型语言模型和强化学习（RL）的一种新扩展范式。具体来说，我们将下一个词元预测重新框定为一个使用 RL 训练的推理任务，该任务为正确预测给定上下文的下一个词元提供可验证的奖励。RPT 提供了一种可扩展的方法来利用海量文本数据进行通用 RL，而不是依赖于特定领域的标注答案。通过激励下一个词元推理的能力，RPT 显著提高了预测下一个词元的语言模型精度。此外，RPT 为进一步强化微调提供了一个强大的预训练基础。扩展曲线显示，增加训练计算量能持续提高下一个词元预测的准确性。结果表明，RPT 是推进语言模型预训练的一个有效且有前途的扩展范式。

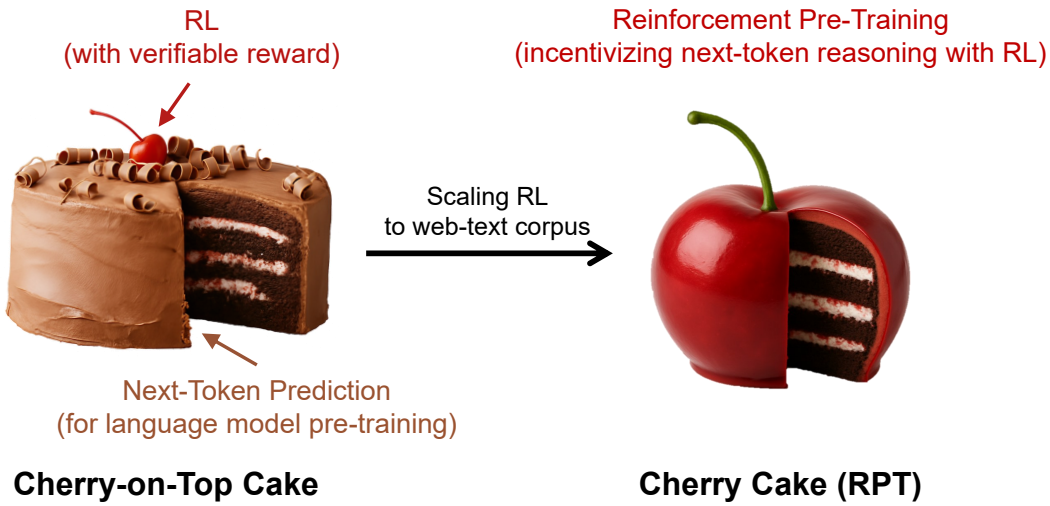

Figure 1: 强化预训练（RPT）将下一个标记的预测重新定义为一个推理任务，其中语言模型通过强化学习（RL）被激励去推理并正确预测下一个标记。所提出的方法允许将 RL 扩展到网络文本语料库。樱桃蛋糕的图像来自 LeCun 的幻灯片 [LeC16] 。

[*] Equal contribution. ◇ Contact person: fuwei@microsoft.com.

1 介绍

大型语言模型（LLM）在广泛的任务中展示了卓越的能力，这主要得益于在大量文本语料库上进行下一个标记预测目标的可扩展性。这种自监督范式已被证明是一种有效的通用预训练方法。同时，强化学习（RL）已成为一种强大的技术，用于微调 LLM，使其与人类偏好保持一致或增强特定技能，例如复杂推理 [OWJ+22, JKL+24, GYZ+25] 。

然而，在 LLM 训练中的强化学习当前应用面临可扩展性和通用性挑战。虽然从人类反馈中进行的强化学习 [OWJ+22] 在对齐方面有效，但依赖于昂贵的人类偏好数据，其学习的奖励模型可能容易受到奖励投机的影响，从而限制了可扩展性。另一种选择是使用可验证奖励的强化学习（RLVR）[LMP+25]，它利用基于规则的客观奖励，通常来自问答对。尽管这缓解了奖励投机的问题，RLVR 通常受到具有可验证答案的标注数据稀缺性的限制，将其应用限制在领域特定的微调，而不是通用预训练。

在这项工作中，我们引入了 reinforcement pre-training （RPT ），这是一种新颖的范式，桥接了可扩展的自监督预训练与强化学习的强大功能之间的差距。RPT 将基本的下一个令牌预测任务重新定义为下一个令牌推理过程。对于预训练语料库中的任何给定上下文，模型被激励在预测之前对后续的令牌进行推理。根据其预测与语料库自身的真实后续令牌的正确性，模型会获得一个可验证的内在奖励。这种方法将通常用于下一个令牌预测的庞大且未标注的文本数据转化为一个通用目的 RL 的大型数据集，而无需外部注释或领域特定的奖励函数。

这种方法提供了几个重要的优点。首先，RPT 本质上具有可扩展性和通用性：它利用了用于标准下一个标记预测的相同的大量未注释文本数据，将其转化为通用目的的 RL 的庞大数据集，而不需要外部注释。其次，使用直接的、基于规则的奖励信号（即预测的下一个标记的正确性）本质上最小化了与复杂的、学习到的奖励模型相关的奖励篡改风险。第三，通过明确鼓励下一个标记的推理模式，RPT 促进了更深层次的理解和泛化，而不仅仅是记住下一个标记。模型学会探讨和验证关于为什么某个标记应该跟随的假设，从而培养出更稳健的表示。最后，在预训练期间的内部推理过程有效地允许模型为每个预测步骤分配更多的“思考”或计算努力，这类似于在训练时为每个标记应用一种形式的推理时间扩展，从而直接提高了下一个标记预测的准确性。

我们的实验表明，RPT 显著提高了预测下一个标记的准确性。RPT 还为后续的强化微调提供了一个更加稳健的预训练基础，从而提升了最终任务的表现。扩展曲线显示，在 RPT 框架下增加训练计算能持续提升下一个标记预测的准确性，表明其作为一种可持续的扩展策略的潜力。这些结果将 reinforcement pre-training 作为推进大型语言模型预训练的高效且有前途的新范式。

我们的贡献总结如下：

- 我们介绍了 reinforcement pre-training （RPT ），一种新的扩展范式，将下一个标记的预测重新定义为一个通过强化学习训练的推理任务，利用直接从预训练语料库中获得的内在可验证奖励。
- RPT 提供了一种可扩展和通用的 RL 预训练方法，通过基于规则的奖励来最小化奖励欺骗，并通过鼓励下一个标记推理模式而不是死记硬背来促进泛化。
- RPT 显著提高了下一个标记预测的准确性，并表现出良好的扩展特性，其性能随着训练计算量的增加而持续提高。
- RPT 为随后的强化微调提供了更强大的预训练基础，并提高了在各种下游任务中的零样本性能。

2 初步

下一个标记预测（NTP） 下一个标记预测是现代大型语言模型的基本训练目标 [AAA+23] 。给定一个来自训练语料库的输入序列 $x_0 \cdots x_T$ ，模型经过训练以最大化以下目标：

$$\mathcal{J}_{\text{NTP}}(\theta) = \sum_{t=1}^{T} \log P(x_t \mid x_0, x_1, \ldots, x_{t-1}; \theta), \tag{1}$$

其中 θ 表示语言模型的参数。

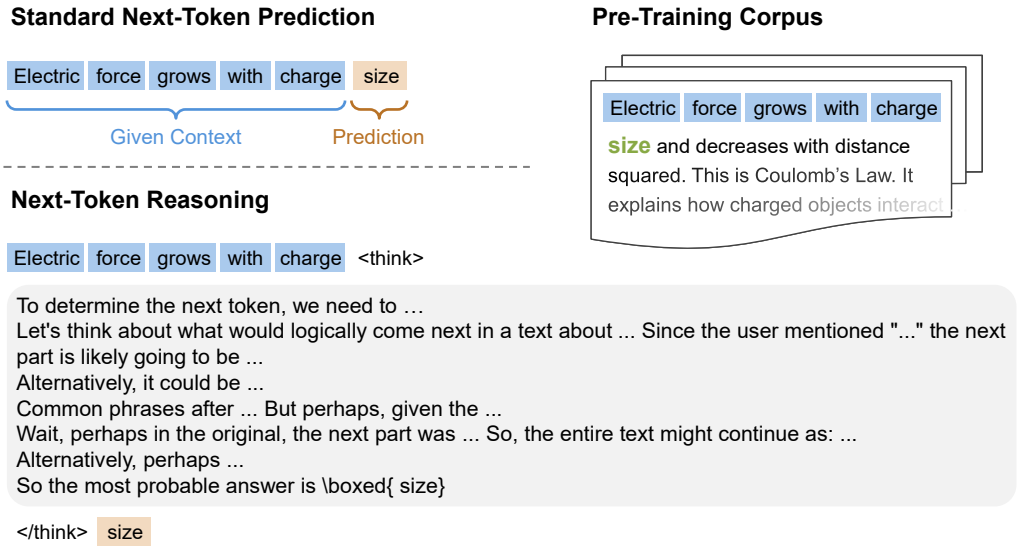

Figure 2: 标准下一个词元预测与下一个词元推理的比较。标准下一个词元预测直接估计预训练语料库中的下一个词元，而下一个词元推理在做出预测之前对多个词元进行推理。

具有可验证奖励的强化学习 (RLVR) RLVR 使用强化学习目标来增强具有可验证答案的特定技能 [LMP+25]。RLVR 需要一个标注的问题-答案对的数据集 $\mathcal{D} = \{(q, a)\}$。对于一个特定的对 XMathX_1，LLM π_θ 生成一个响应 $o \sim \pi_\theta(\cdot \mid q)$。一个确定性验证器 $\mathcal{V}$ 计算一个可验证的奖励 $r = \mathcal{V}(o, a)$，模型被训练以最大化期望奖励：

$$\mathcal{J}_{\text{RLVR}}(\theta) = \mathbb{E}_{(q,a)\sim\mathcal{D},\, o\sim\pi_\theta(\cdot|q)} \left[r(o, a) \right]. \tag{2}$$

3 强化预训练

3.1 预训练任务：下一个标记推理

我们为语言建模提出了下一个标记的推理任务。给定来自训练语料库的输入序列 $x_0 \cdots x_T$，对于每个位置 $t \in \{1, \ldots, T\}$，前缀 $x_{<t}$ 被视为上下文，真实的下一个标记是 x_t。在下一个标记的推理任务中，模型 π_θ 需要在为下一个标记生成预测 y_t 之前，生成一个思维链推理序列，表示为 c_t。整体模型响应为 $o_t = (c_t, y_t)$，$o_t \sim \pi_\theta(\cdot \mid x_{<t})$。

如 Figure 2 所示，进行下一步推理的长链思维过程可能涉及各种推理模式，如头脑风暴、自我批评和自我纠正。下一步推理任务将预训练语料库重构为大量推理问题，将预训练从学习表面上的词元级相关性转变为理解其背后的隐性知识，并使 RL 扩展成为可能。

3.2 用强化学习进行预训练

强化预训练（RPT）通过策略优化强化学习训练大型语言模型（LLMs）进行下一个标记的推理，如 Figure 3 所示。对于上下文 $x_{<t}$，我们提示语言模型 π_θ 生成 G 个响应（思考轨迹），$\{o_t^i\}_{i=1}^G$。每个响应 $o_t^i = (c_t^i, y_t^i)$ 由一个思维链推理序列 c_t^i 和一个最终预测序列 y_t^i 组成。

为了验证 y_t^i 的正确性，我们引入了一种前缀匹配奖励，这可以支持验证跨多个标记的预测或涉及超出词汇表的标记。令 $\overline{x}_{\geq t}$ 和 $\overline{y}_t^i$ 分别表示真实完成序列 $x_{\geq t}$ 和预测 y_t^i 的字节序列。用 l 表示 $\overline{y}_t^i$ 的字节长度。我们将真实完成序列中各标记的累计字节长度定义为有效边界，并将此集合表示为 $\mathcal{L}_{gt}$。形式上，i 处对于 $x_{<t}$ 的输出的奖励 r_t^i 定义为：

$$r_t^i = \begin{cases} 1 & \text{if } \overline{y}_t^i = \overline{x}_{\geq t}[1:l] \text{ and } l \in \mathcal{L}_{gt} \\ 0 & \text{otherwise} \end{cases}, \tag{3}$$

。如果预测的字节序列是真实完成序列的精确前缀，并且其长度 l 匹配任何有效的标记边界，则奖励为 1。

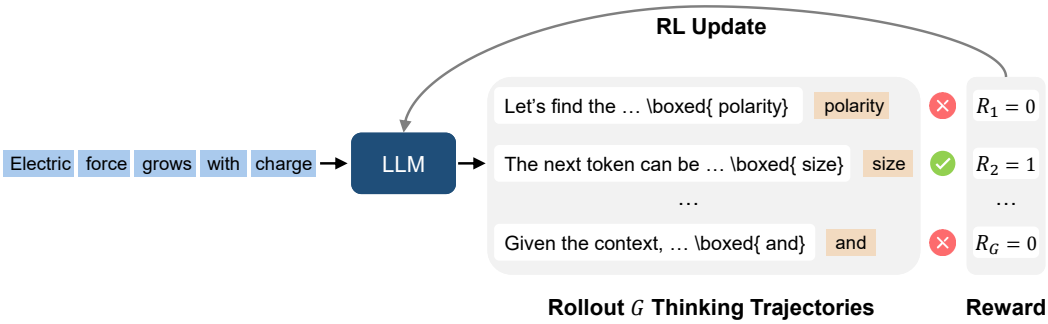

Figure 3: reinforcement pre-training 的一个例子。给定一个缺少后续部分的上下文，大型语言模型执行策略内展开，以生成 G 个不同的思维轨迹。每条轨迹包括一个中间推理步骤和一个下一个标记的最终预测。如果预测与真实标记匹配，则给予正奖励；否则，奖励为零。此奖励信号用于更新大型语言模型，鼓励引导至准确续写的轨迹。

令 $\mathcal{D}$ 为所有 $\{x_{<t}\}_{t=1}^{T}$ 的集合，该模型被训练以最大化期望奖励：

$$\mathcal{J}_{\mathrm{RPT}}(\theta) = \mathbb{E}_{(x_{<t}, x_{\geq t})\sim\mathcal{D},\ \{o_t^i\}_{i=1}^{G}\sim\pi_\theta(\cdot|x_{<t})}\left[r_t^i\right]. \tag{4}$$

3.3 预训练设置

我们使用 OmniMATH 数据集 [GSY+24] 进行 reinforcement pre-training 。OmniMATH 包含来自官方网站（如 AoPS Wiki [2] 和 AoPS 论坛 [3] ）的 4,428 个竞赛级数学问题及其解决方案。由于许多标记即使无需推理也很容易预测，我们在 reinforcement pre-training 之前进行标记级的数据过滤。特别是，我们使用 Deepseek-R1-Distill-Qwen-1.5B 模型作为一个小型代理模型。对于每个标记，我们计算代理模型在前 16 个下一个标记上的熵值。通过应用熵阈值，我们过滤掉低熵位置，优先对那些需要更大计算努力来预测的有挑战性的标记进行训练。

在所有实验中，我们使用 Deepseek-R1-Distill-Qwen-14B [GYZ+25] 作为基础模型。R1-Distill-Qwen-14B 是强化学习的良好起点，因为它具备基本的推理能力。我们使用 verl 库 [SZY+24] 实现我们的训练框架，并使用 vllm 进行推理。我们采用 GRPO 算法 [GYZ+25]，具体的超参数详见 Appendix B 。在训练期间，我们采取 8k 的训练长度，学习率为 1×10^{-6} ，零 KL 惩罚，批量大小为 256 个问题。对于每个问题，采样 G=8 个响应，在 rollout 过程中，我们使用 0.8 的温度。从每个响应中，我们直接提取最后 \框内的完整序列，在特殊标记 '</think>' 之后，作为模型预测的下一个标记。从 500 步开始，我们利用动态采样来提高训练效率 [YZZ+25] 。我们主要实验的总训练步数为 1,000。提示模板及其变体的讨论见 Appendix D 。

3.4 预训练模型的评估

一旦模型经过预训练，我们可以直接在下游任务上进行下一个标记预测和强化微调。我们使用这些设置来展示 reinforcement pre-training 提升了大型语言模型的语言建模能力和推理能力。

给定下一个标记推理目标，我们的模型可以自然地用于语言建模。我们报告下一个标记预测的准确性，以评估 RPT 的语言建模性能和扩展性。

我们以预训练到微调的方式对 RPT 模型进行持续的 RL 微调。由于 RPT 将预训练过程与 RL 对齐，因此在后训练期间预训练和 RL 之间的目标差距被最小化。我们评估 reinforcement pre-training 过程是否进一步增强了在最终任务中的后训练效果。

4 实验

我们在 OmniMATH 中保留的 200 个样本的验证集上评估语言建模性能。按照我们设置中描述的基于熵的数据过滤策略 (Section 3.3)，我们根据困难程度对验证集中的标记位置进行分

[2] https://artofproblemsolving.com/wiki/index.php
[3] https://artofproblemsolving.com/community/c13_contests

类。具体来说，我们使用 R1-Distill-Qwen-14B 计算每个标记位置的熵。然后，如果其熵分别超过 0.5, 1.0, 和 1.5 的阈值，我们将这些位置归为简单、中等或难的类别。为了进行比较，我们报告了两种不同方式评估的 R1-Distill-Qwen-14B 的性能：（1）标准的下一个标记预测，选择概率最高的标记；以及（2）下一个标记推理，在最终预测之前生成思路链。我们还包括了 Qwen2.5-14B 的结果，因为它是 R1-Distill-Qwen-14B 的基础模型。

	Easy	Medium	Hard
Standard next-token prediction			
Qwen2.5-14B	41.90	30.03	20.65
R1-Distill-Qwen-14B	41.60	29.46	20.43
Next-token reasoning			
R1-Distill-Qwen-14B	3.31	1.66	1.41
RPT-14B	**45.11**	**33.56**	**23.75**

Table 1: Next-token prediction accuracy across three test splits of varying difficulty. RPT outperforms both the standard next-token prediction baselines and the reasoning-based prediction baseline.

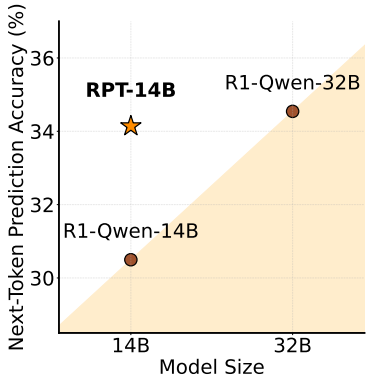

Figure 4: 不同难度数据的平均下一个标记预测准确性。R1-Qwen-14B/32B 分别表示 R1-Distill-Qwen-14B/32B。

如 Table 1 所示，RPT -14B 在所有难度级别上实现了一致更高的下一个令牌预测准确性，与 R1-Distill-Qwen-14B 相比。值得注意的是，它的性能与显著更大的模型（即 R1-Distill-Qwen-32B，Figure 4 ）相匹配。这些结果表明，reinforcement pre-training 在捕捉生成令牌所需的复杂推理信号方面是有效的，并且在提高 LLMs 的语言建模能力方面具有很大的潜力。

在这一节中，我们研究了 reinforcement pre-training 的缩放特性。在自然语言语料库上进行下一词预训练所取得的损失在经验上随着模型大小、训练标记数量和训练计算量呈现幂律衰减。接下来，我们具体分析 RPT 在训练计算量 C 方面的缩放行为。我们使用以下幂律形式来建模这种关系：其中，$P(C)$ 表示验证集上的下一词预测准确率。P^* ，α 和 A 是待估计的参数。

我们在不同的训练步骤（100, 200, 400, 800, 1000 和 1200）下评估 RPT 的下一个词预测准确性，并将其转换为相应的训练计算。为了评估数据难度的影响，我们考虑通过熵阈值 0.5（简单）、1.0（中等）和 1.5（困难）过滤的验证数据分割。更高的阈值意味着对于 LLM 而言输入更具挑战性。对于每个难度等级，我们根据 **??** 拟合结果。我们使用决定系数 R^2 来度量拟合的优度，该系数量化了缩放曲线与观察数据的拟合程度。

如 Figure 5 所示，随着训练计算量的增加，RPT 的下一个标记预测准确性稳步提升。在所有难度级别上，较高的 R^2 值表明拟合曲线准确地捕捉到了性能趋势。

4.1 使用 RPT 的强化微调

为了研究是否可以使用 RLVR 对 RPT 模型进行更有效的微调，我们从 Skywork-OR1 [HLL+25] 中随机抽取具有可验证答案的问题进行进一步训练。我们使用 256 个例子进行训练，200 个进行测试。遵循 Skywork-OR1 [HLL+25] 的数据过滤流程，我们使用 R1-Distill-Qwen-32B 来识别具有挑战性的训练实例。我们将训练批量大小和 PPO 小批量大小都设置为 64，并训练 15 个 epoch。在评估过程中，验证的最大 token 数量设置为 32,000，温度为 0.6。

如 Table 2 所示，强化预训练模型在使用 RLVR 进一步训练时达到了更高的上限。当模型在相同的数据上使用下一个 token 预测目标持续训练时，其推理能力显著下降。后续的 RLVR 仅能带来缓慢的性能提升。这些结果表明，在有限的数据下，reinforcement pre-training 可以快速将从下一个 token 推理中学到的加强推理模式转移到最终任务。

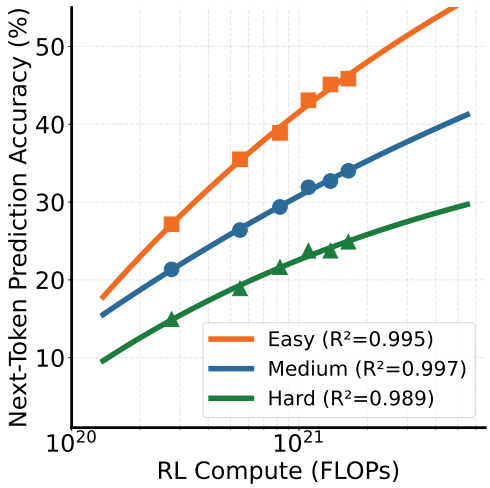

Figure 5: 在所有数据难度下，reinforcement pre-training 的下一个令牌预测精度随着训练计算的增加而持续提高。拟合曲线表现出较高的决定系数，表明预测值与观测值之间的一致性。

	Before RL	**After RL**
R1-Distill-Qwen-14B	51.2	52.7
+ Continual NTP training	10.7	13.0
RPT-14B	**56.3**	**58.3**

Table 2: 不同模型的强化微调性能。“持续的 NTP 训练”指的是在与 RPT -14B 相同语料库上，使用标准的下一个词预测目标进行的持续预训练。RPT 为后续的强化学习训练提供了更强的基础。

4.2 零样本在终端任务上的表现

我们评估了 RPT -14B 在终端任务上的零次性能。为了比较，我们评估了 R1-Distill-Qwen-14B 和 R1-Distill-Qwen-32B 的下一个词预测性能，以及 RPT -14B 的推理性能与 R1-Distill-Qwen-14B 的比较。

我们的评估涉及两个广受认可的基准：MMLU-Pro [HBB+20]，这是一个全面的多任务理解基准，用于评估 LLM 在各个领域的能力；SuperGPQA [DYM+25]，一个涵盖 285 个学科的大规模研究生水平推理问题的基准。在推理设置下，我们将最大令牌数设为 12,288，温度设为 0.8。根据之前的工作 [MLJ+25, ZLS+25b]，我们使用多项选择题格式进行评估，并报告准确率。

	SuperGPQA	MMLU-Pro
Standard next-token prediction mode		
R1-Distill-Qwen-14B	32.0	48.4
R1-Distill-Qwen-32B	37.2	56.5
Reasoning mode		
R1-Distill-Qwen14B	36.1	68.9
RPT-14B	39.0	71.1

Table 3: 零样本性能在通用领域终端任务中。RPT -14B 在推理模式下始终优于 14B 和 32B 基线。

如 Table 3 所示，RPT -14B 在所有基准测试中始终优于 R1-Distill-Qwen-14B（无论是使用标准的下一个标记预测还是作为推理模型进行评估）。值得注意的是，它还超过了明显更大的

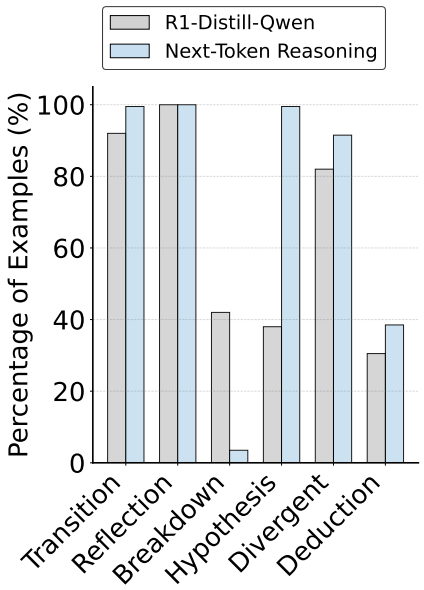

Figure 6: 用于问题解决的 R1-Distill-Qwen-14B 和用于下一个词推理的 RPT -14B 的推理模式统计。

R1-Distill-Qwen-32B（在下一个标记预测下），在 SuperGPQA 上提高了 7 分，在 MMLU-Pro 上提高了约 22 分。每个基准测试的详细每个主题的结果在 Appendix C 中提供。

4.3 下一个令牌推理模式分析

我们分析了下一个标记推理与明确问题解决之间推理模式的差异。根据先前的研究 [WYZ+25, GCD+25]，我们统计测量了包含推理指示性关键词（例如，“分解”，“或者”）的模型响应的比例。[4]

我们的分析比较了两个模型在 OmniMATH 数据集上的思维过程，即用于问题解决的 R1-Distill-Qwen-14B 和用于下一个词推理的 RPT -14B，基于每个模型的 200 个抽样响应。我们将推理模式分为六种类型：转换（切换策略）、反思（自我检查）、细分（分解问题）、假设（提出和验证假设）、发散思维（探索可能性）和演绎（逻辑推理）。

如 Figure 6 所示，RPT -14B 的下一个标记推理过程与 R1-Distill-Qwen-14B 的问题解决过程明显不同，前者对假设模式的使用增加了 161.8 %，而对演绎模式的使用增加了 26.2 %。相比之下，问题解决过程更加依赖于分解模式，这强调了下一个标记推理引发的推理过程在质上与结构化问题解决不同。

我们还提供了一个关于 Table 4 中的推理模式的例子。该例子揭示了模型参与的是一个深思熟虑的过程，而不是简单的模式匹配。它分析了更广泛的语义背景（“计算向量大小”）、识别关键短语（“复习一些……”），然后集思广益，权衡多种合理的延续。这涉及假设生成（“下一部分很可能是……”）、考虑替代方案（“或者，也可能是……”），以及对结构性提示（“使用标题标记”）甚至细粒度的标记级细节（“可以有一个空格”）的反思。这种多方面的推理，既包括高级语义理解又包括低级文本特征，展示了模型通过理性的探索来推断下一个标记的努力，符合 RPT 的目标，即培养超越表面相关性的深层理解。更多例子见于 Appendix F 。

5 相关工作

大规模语言模型的进步主要由两个扩展维度驱动：训练时计算 [KMH+20, HBM+22] 和测试时计算 [ZLS+25a] 。训练时扩展通过增加模型参数和训练数据显著提高了能力，使用下一个词预测作为预训练任务。与此同时，测试时扩展 [JKL+24] 则通过增加推理计算量以提升

[4] 关键词列在 Appendix E 中。

Thinking Process of RPT-14B
Okay, I need to figure out what the next token ... Let's think about what would logically come next in a text about calculating vector magnitude. 到目前为止的流程是 introducing the topic, explaining the formulas for 2D and 3D vectors, and now moving on to how to actually perform the calculation. Since the user mentioned "go over some..." 下一部分可能会是…… Alternatively, it could be ... 在"我们可以讨论一些"之后的常见短语可能是"例子"、"步骤"、"方法"等。在类似的背景中，在教育材料中， after explaining formulas, they often provide "steps" or "examples." But perhaps, given the user started a new section, 这可能是一个新段落... 等等，也许在原文中，接下来的部分是一个例子或解释. Let's think about standard educational content. So, the entire text might continue as: ... Alternatively, perhaps the next line is a new section. 等等，还有另一种方法：... However, the original text 可能在下一个单词前有一个空格…… Wait, perhaps looking at how the initial document is structured. It's using markdown with headers:...

Table 4: RPT 推理模式的例子。

大规模语言模型的推理能力。在超越现有扩展范式的同时，RPT 独特地整合了上述原理，将每个下一个词预测构架为一个推理任务。

用于大语言模型的强化学习 强化学习 (RL) 在大型语言模型的后训练阶段发挥了关键作用。从人类反馈中进行强化学习 [OWJ+22] 将预训练语言模型针对人类偏好数据进行微调，以提升一致性。除此之外，大规模 RL 也被引入以增强语言模型的推理能力 [JKL+24, GYZ+25]。[ZHS+24] 是最相关的工作，它鼓励语言模型生成有助于下一个标记预测的合理解释。基于帮助性的奖励往往会因为在生成的理由中重复目标标记而被滥用，这种捷径可能会损害模型。相比之下，我们使用下一个标记预测的正确性作为基于规则的奖励信号，以最小化奖励滥用。

6 结论与未来工作

我们引入了 reinforcement pre-training (RPT)，这是一种新颖的大型语言模型预训练范式。通过将下一个词的预测框定为一个可验证的推理任务，并应用基于正确性的奖励进行强化学习，RPT 允许 LLMs 在预训练期间利用扩展的计算来构建更强的基础推理能力。我们的实验表明，RPT 改进了下一个词的预测，增强了在零样本情况下的数学和一般推理基准测试中的表现，并为进一步的 RL 微调提供了更好的起点。RPT 通过从根本上重新思考预训练目标本身，为开发更强大和更具普遍智能的 LLMs 提供了一条有希望的新方向。

虽然这项对 RPT 的初步探索很有前景，但仍存在某些局限性。我们的实验主要是在一个 14B 参数模型上进行的。尽管 RPT 方法论被设计为通用的，但当前的预训练语料库主要由数学文档构成；未来的工作将探索其在更广泛的通用领域文本中的效果。此外，RPT 训练是从一个推理模型初始化的；研究从一个标准基础语言模型开始的 RPT 训练将为其基础影响提供进一步的洞察。

可以从以下几个方面推进这项工作。我们希望扩大训练语料库，包括数据规模和领域覆盖范围。在 reinforcement pre-training 期间可以利用大规模的通用互联网数据。我们还将增加训练计算量，以推动前沿发展。此外，我们可以为 reinforcement pre-training 建立扩展法则，以指导大语言模型的扩展。此外，我们对将混合思维 [JWH+25] 与 RPT 集成感兴趣，以通过自适应地触发下一个标记推理来实现细粒度的自适应思维。

7

致谢

我们对蒋雨婷在 GPU 集群维护方面的贡献表示感谢。我们还感谢迟泽文和王洋在 MI300 GPU 上的强化学习基础设施开发过程中提供的技术支持。我们的训练实现基于 verl [SZY+24] 。

References

[AAA+23] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[DYM+25] Xinrun Du, Yifan Yao, Kaijing Ma, Bingli Wang, Tianyu Zheng, King Zhu, Minghao Liu, Yiming Liang, Xiaolong Jin, Zhenlin Wei, et al. SuperGPQA: Scaling LLM evaluation across 285 graduate disciplines. *arXiv preprint arXiv:2502.14739*, 2025.

[GCD+25] Jiaxin Guo, Zewen Chi, Li Dong, Qingxiu Dong, Xun Wu, Shaohan Huang, and Furu Wei. Reward reasoning model, 2025.

[GSY+24] Bofei Gao, Feifan Song, Zhe Yang, Zefan Cai, Yibo Miao, Qingxiu Dong, Lei Li, Chenghao Ma, Liang Chen, Runxin Xu, Zhengyang Tang, Benyou Wang, Daoguang Zan, Shanghaoran Quan, Ge Zhang, Lei Sha, Yichang Zhang, Xuancheng Ren, Tianyu Liu, and Baobao Chang. Omni-MATH: A universal Olympiad level mathematic benchmark for large language models. *ArXiv*, abs/2410.07985, 2024.

[GYZ+25] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[HBB+20] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

[HBM+22] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

[HDW+25] Yaru Hao, Li Dong, Xun Wu, Shaohan Huang, Zewen Chi, and Furu Wei. On-policy RL with optimal reward baseline, 2025.

[HLL+25] Jujie He, Jiacai Liu, Chris Yuhao Liu, Rui Yan, Chaojie Wang, Peng Cheng, Xiaoyu Zhang, Fuxiang Zhang, Jiacheng Xu, Wei Shen, Siyuan Li, Liang Zeng, Tianwen Wei, Cheng Cheng, Bo An, Yang Liu, and Yahui Zhou. Skywork open reasoner 1 technical report. *arXiv preprint arXiv:2505.22312*, 2025.

[JKL+24] Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.

[JWH+25] Lingjie Jiang, Xun Wu, Shaohan Huang, Qingxiu Dong, Zewen Chi, Li Dong, Xingxing Zhang, Tengchao Lv, Lei Cui, and Furu Wei. Think only when you need with large hybrid-reasoning models, 2025.

[KMH+20] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *CoRR*, abs/2001.08361, 2020.

[LeC16] Yann LeCun. Predictive learning. Advances in Neural Information Processing Systems, 2016.

[LMP+25] Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V. Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Chris Wilhelm, Luca Soldaini, Noah A. Smith, Yizhong Wang,

Pradeep Dasigi, and Hannaneh Hajishirzi. Tulu 3: Pushing frontiers in open language model post-training, 2025.

[MLJ+25] Xueguang Ma, Qian Liu, Dongfu Jiang, Ge Zhang, Zejun Ma, and Wenhu Chen. General-reasoner: Advancing llm reasoning across all domains. *arXiv preprint arXiv:2505.14652*, 2025.

[OWJ+22] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

[SZY+24] Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. HybridFlow: A flexible and efficient RLHF framework. *arXiv preprint arXiv: 2409.19256*, 2024.

[WYZ+25] Yiping Wang, Qing Yang, Zhiyuan Zeng, Liliang Ren, Liyuan Liu, Baolin Peng, Hao Cheng, Xuehai He, Kuan Wang, Jianfeng Gao, Weizhu Chen, Shuohang Wang, Simon Shaolei Du, and Yelong Shen. Reinforcement learning for reasoning in large language models with one training example, 2025.

[YZZ+25] Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, et al. DAPO: An open-source LLM reinforcement learning system at scale. *ArXiv*, abs/2503.14476, 2025.

[ZHS+24] Eric Zelikman, Georges Harik, Yijia Shao, Varuna Jayasiri, Nick Haber, and Noah D Goodman. Quiet-star: Language models can teach themselves to think before speaking. *arXiv preprint arXiv:2403.09629*, 2024.

[ZLS+25a] Qiyuan Zhang, Fuyuan Lyu, Zexu Sun, Lei Wang, Weixu Zhang, Wenyue Hua, Haolun Wu, Zhihan Guo, Yufei Wang, Niklas Muennighoff, et al. A survey on test-time scaling in large language models: What, how, where, and how well? *arXiv preprint arXiv:2503.24235*, 2025.

[ZLS+25b] Xiangxin Zhou, Zichen Liu, Anya Sims, Haonan Wang, Tianyu Pang, Chongxuan Li, Liang Wang, Min Lin, and Chao Du. Reinforcing general reasoning without verifiers. *arXiv preprint arXiv:2505.21493*, 2025.

A 奖励的设计选择

我们还研究了几种替代的奖励函数，以评估它们对 reinforcement pre-training 的影响，除了 Section 3 中描述的奖励机制，即前缀匹配奖励。

一个变体是首个 token 匹配。在这种设置中，奖励仅反映模型预测的第一个 token y_t^i 是否匹配真实的下一个 token x_t ，忽略预测中第一个 token 之后的所有 token。另一种替代方案探讨了“密集奖励”方案。在这里，正确预测的下一个 token（即 $y_t^i[0] = x_t$ ）将获得满额奖励（例如，1）。对于错误预测（$y_t^i[0] \neq x_t$ ），奖励是一个正的较小值，特别是生成该特定错误 token 的语言模型概率 $P(y_t^i[0] \mid x_{<t}; \theta)$ 。这比二元奖励提供了更密集的反馈信号。第三种设计是有条件地应用这种密集奖励结构。如上所述，密集奖励（正确为 1，错误为 $P(y_t^i \mid x_{<t}; \theta)$ ）用于训练实例（针对给定前缀 $x_{<t}$ 的一组 rollouts），其中至少一个 G 采样的 rollouts 正确预测了下一个 token。如果一个组中的所有 G rollout 都是错误的，将应用不同的奖励方案（例如，对于所有的零奖励，或一个统一的小惩罚）。

我们的实验表明，替代的奖励设计通常达到了与前缀匹配奖励相当的性能。这表明 reinforcement pre-training 框架对于这些特定的奖励信号修改是相对鲁棒的，其核心优势可能对这些具体选择并不过于敏感，至少在所测试的变化范围内。

B Reinforcement Pre-Training 使用的超参数

Table 5 给出了 Section 4 中 reinforcement pre-training 的详细超参数。我们遵循精确的策略强化学习的设置 [HDW+25] ，并将熵损失系数设为 0 。

Params	Values
Actor gradient clip	0.2
Batch size	256
PPO mini batch size	256
Rollout number	8
Learning rate	10^{-6}
Adam β	(0.9, 0.999)
Weight decay	0.01
Sampling temperature	0.8
Max prompt length	4096
Max response length	8192
Entropy loss coefficient	0

Table 5: 在 Section 4 中用于 reinforcement pre-training 的超参数。

C 终端任务的详细结果

Table 6 和 Table 7 在通用终端任务基准上展示了详细的按类别性能。值得注意的是，R1-Distill-Qwen-14B 的性能通过两种不同的方式进行了评估：标准的下一个词预测和基于推理的答案预测（标识为‘+ think’）。RPT -14B 模型相较于 R1-Distill-Qwen-14B 和 R1-Distill-Qwen-32B 显示了更优越的性能。

	Agron.	Econ.	Educ.	Engin.	Hist.	Law	L. & A.	Manag.	Med.	Mil.	Sci.	Phil.	Sci.	Sociol.	Overall
Standard next-token prediction mode															
R1-Distill-Qwen-14B	30.0	38.0	32.0	31.0	24.5	26.0	28.5	39.0	35.5	36.0	37.0	24.0	30.1	32.0	
R1-Distill-Qwen-32B	32.5	39.5	43.0	34.0	29.5	31.0	28.5	41.5	43.5	49.0	44.5	29.5	38.5	37.2	
Reasoning mode															
R1-Distill-Qwen-14B	31.0	41.0	32.0	34.5	29.0	31.0	29.5	39.5	38.5	39.5	44.0	41.5	39.2	36.1	
RPT-14B	35.0	40.0	41.5	40.5	30.5	32.0	29.0	36.0	44.5	41.0	49.0	47.0	42.0	39.0	

Table 6: 在 SuperGPQA 上的详细零样本性能。

	Bio.	Bus.	Chem.	CS	Econ.	Engin.	Heal.	Hist.	Law	Math	Other	Phil.	Phys.	Psych.	Overall
Standard next-token prediction mode															
R1-Distill-Qwen-14B	72.5	42.5	34.0	46.5	58.0	44.0	57.5	54.0	37.0	36.5	50.0	48.5	34.5	62.0	48.4
R1-Distill-Qwen-32B	82.5	46.0	39.0	55.5	74.0	52.0	68.0	62.5	47.0	46.0	54.0	53.5	42.5	68.5	56.5
Reasoning mode															
R1-Distill-Qwen14B	85.0	65.5	74.5	75.0	81.5	52.0	70.0	61.5	42.0	86.0	65.0	62.5	80.0	64.5	68.9
RPT-14B	84.5	72.0	77.5	76.0	78.5	53.5	74.0	63.0	44.5	91.5	66.0	63.5	82.5	68.0	71.1

Table 7: 对 MMLU-Pro 的详细零样本表现。

Prompt Template	Random@1 (%)	Pass@8 (%)
v0	3.0	8.5
v1	5.7	11.0
v2	5.7	16.0
v3	5.3	11.0
v4	4.0	9.0
v5	4.4	12.5
v6	6.0	19.0

Table 8: 提示模板的影响。

Pattern Group	Keywords
Transition	alternatively, think differently
Reflection	wait, initial answer, original answer, looking back, thought process
Breakdown	break down, break this down
Hypothesis	probably, something like
Divergent Thinking	etc., or something, either, sometimes it refers, otherwise, exploring, options
Deduction	summarize, conclusion, conclude, finally, logically, consequently

Table 9: 模式组和关键词应用于 Section 4.3 。

D 提示模板的影响

我们探讨了各种提示模板对初始下一个标记推理性能的影响。Table 10 展示了七种模板变体。模板使用了不同的指令表达方式，并以各种格式包装了上下文。

如 Table 8 所示，清晰的提示显著提高了初始表现的正确性。注意，Section 4 中的 reinforcement pre-training 实验使用了“v0”提示模板。我们将基于其他模板变体的提示工程留待未来研究，这可能有助于提高最终表现。

E 推理模式分析的关键词

Table 9 展示了推理模式分析中应用的模式组和关键词。

F 案例研究

为了提供关于使用 RPT 训练的模型行为的定性见解，我们展示了一些在 Table 11 中的下一个令牌推理案例。

Version	Prompt Content
v0	Complete the given text under ‘ # # # Context’ by predicting the next token, and wrap it in ‘\boxed { } ’. Please reason step by step to find the most probable next token as the final answer, and enclose it in \boxed { } (note: the token may begin with a space, e.g., \boxed { para } or \boxed { = } ; do not use \text { }). # # # Context { prompt_content }
v1	Complete the given text under # # # Context by predicting the next token, and wrap it in \\boxed { } . Please reason step by step to find the most probable next token as the final prediction, and enclose it in \boxed { } (note: the token may begin with a space, e.g., \boxed { para } or \boxed { = } ; do not use \text { }). # # # Context ```{ prompt_content } ```.
v2	You are a helpful assistant, good at predicting the next token for a given context. Now, please complete the given text under # # # Context by predicting the next token, and wrap it in \\boxed { } . Please reason step by step to find the most probable next token, and enclose it in \boxed { } (note: the token may begin with a space, e.g., \boxed { para } or \boxed { += } ; do not use \text { }). # # # Context ```{ prompt_content } ```.
v3	Complete the given text under # # # Context by predicting the next token, list multiple potential tokens and select the most probable one as the final answer. Wrap your final answer in \boxed { } (note: the token may begin with a space, e.g., \boxed { para } or \boxed { = } ; do not use \text { }). # # # Context ```{ prompt_content } ```
v4	Complete the given text under # # # Context by predicting the next token, and wrap it in \boxed { } . Please reason step by step to find the most probable next token as the final answer, and enclose it in \boxed { } . Some examples: # # # Context \n \n ```...(some omitted)...Matching calculations with 1990 valid combinations indicates the minimum value of \(b \) that fits all pre-requisites and restrictions for triangle formation and symmetry generates the efficient outcome: \n \n \[\n \boxed { 1991^2 } \n \] \n \nIn``` The next token is \boxed { this } # # # Context \n \n ```...Thus $ 2^{ A } =\left(2^{ a } \right)^{ 2 } \left(2^{ 3 } \right)=``` The next token is \boxed { 9 } # # # Context \n \n ```..., numerical exploration shows``` The next token is \boxed { :\n } Now, the context is: # # # Context \n \n ```{ prompt_content } ```.
v5	Complete the given text under # # # Context by predicting the next token, and wrap it in \boxed { } . Please reason step by step to find the most probable next token as the final answer, and enclose it in \boxed { } (note: the token may begin with a space, e.g., \boxed { para } or \boxed { = } ; do not use \text { }). # # # Context ```{ prompt_content } ```.
v6	Complete the given text wrapped in ```and ```by predicting the next token, list multiple potential tokens and select the most probable one as the final prediction. Wrap your final prediction in \boxed { } (note: the token may begin with a space, e.g., \boxed { para } or \boxed { = } ; do not use \text { }). The context is: ```{ prompt_content } ```, now please predict the next token.

Table 10: 用于下一个令牌推理任务的七个提示模板。

Case 1
Okay, I need to figure out what the next token is in the given text under the Context section. The user provided a partially completed text about integers, comparing and ordering them ... 首先，我将分析所提供的背景 ... 等等，用户写了 6 x 4 = 24, 20 > 18 x，所以下一个符号可能是第二个分数 ... One way to find a common denominator is to multiply the two original denominators ... 或者，也可以是 ... Wait, but 18 x 24? That doesn't make sense. Alternatively, perhaps it's 18 x 4 = 72 or something... 另一种方法：可能下一个标记是 ... 让我看看： if the two fractions are 5/3 and 4/6. 5/3 is greater than 4/6 because 5/3 is 1.666..., and 4/6 is 0.666..., to compare them, find a common denominator... Alternatively, 可能下一个符号是反斜杠或 LaTeX 中的某个内容. Wait, but in the Markdown, the user presented it as: Writing Equivalent Fractions ... 但是等等， in the user's provided context, it's ... Maybe the next token is another multiplication symbol, but that seems unlikely. Alternatively, perhaps it's just the end of the thought. But given the context, 我认为下一个标记可能是一个空格后跟一个 "2" 或一个句号或其他东西 ... 等一下，也许另一个例子。 If the fractions were 5/6 and 4/3 ... So, after 'x' there's a space, then '4'.\n</think>\n\n所以，下一个标记是' 4'。\n \n \boxed { 4 }
Case 2
Okay, I need to figure out what the next token is in the given text under the Context section. Using the integral test, for a set ... 也许最初的答案只是一个不完整的句子。让我们再想一想。The question is to prove convergence of the series, the answer uses the integral test, which is the standard approach ... Alternatively, maybe the original answer was longer, but got cut off, 所以我需要预测接下来会发生什么。也许下一个标记是 ... 所以下一个标记是一个空格，然后是其他内容。 Alternatively, perhaps it's a continuation of the same sentence ... 因此, by the integral test, the series converges when $ p > 1 $. No, perhaps it's better to see what the original answer was supposed to be. The correct full proof would be: Using the integral test, for a set ... 但用户指令是提供给定文本之后的下一个标记 ... 或者, looking at the way the answer was written, perhaps ... I think the next token is ' $ '.\n</think>\n\n因此，下一个符号是' $ '。\n \n \boxed { $ }
Case 3
Okay, I need to figure out what the next token is in the given text under the Context section. The text so far is: Call the large circle's center ... Scaling by a factor... 或许接下来是半径的比率？或者是其他的东西。等等, the scaling factor is probably such that the smaller circle is scaled up to match the larger circle or something ... So, in coordinates, perhaps O1 is at (0,0), O2 is at (41,0). The small circle has radius 13, so the points where O1O2 intersects the small circle are at O2 plus or minus 13 units towards O1 ... 所以，点 A 在小圆上，比 O2 更远离 O1，对吧。 Because if you go from O2 towards O1, 13 units, that would be ... 因此，"将小圆按比例缩放" 的下一个标记很可能是数字 30/13。 ... Wait, I'm a bit confused. Alternatively, maybe it's presented as a fraction ... Scale the small circle by a factor of. 因此，实际的下一个标记是下一个单词或符号 ... Wait, but in the context given, the user just wrote ... 或者，也许在文本中写作 ... So, in the given text, the next token after 'Scale the small circle by a factor of' is space.\n</think>\n\n下一个标记是' '。\n \n \框 { }

Table 11: RPT -14B 的三种推理模式示例。