摘要

The GPT-Academic program cannot find abstract section in this paper.

Open-Reasoner-Zero: An Open Source Approach to Scaling Up Reinforcement Learning on the Base Model

Jingcheng Hu [1,2*] , Yinmin Zhang [1] , Qi Han [1] , Daxin Jiang [1] , Xiangyu Zhang [1] ,
Heung-Yeung Shum2
[1] StepFun, [2] Tsinghua University

GitHub: https://github.com/Open-Reasoner-Zero/Open-Reasoner-Zero
HuggingFace: https://huggingface.co/Open-Reasoner-Zero

Abstract

We introduce Open-Reasoner-Zero, the first open source implementation of large-scale reasoning-oriented RL training focusing on scalability, simplicity and accessibility. Through extensive experiments, we demonstrate that a minimalist approach, vanilla PPO with GAE ($\lambda = 1, \gamma = 1$) and straightforward rule-based reward function, without any KL regularization, is sufficient to scale up both response length and benchmark performance on reasoning tasks, similar to the phenomenon observed in DeepSeek-R1-Zero. Notably, our implementation outperforms DeepSeek-R1-Zero-Qwen-32B on the GPQA Diamond benchmark, while only requiring 1/30 of the training steps. In the spirit of open source, we release our source code, parameter settings, training data, and model weights. *Work done during internship at StepFun. Contents

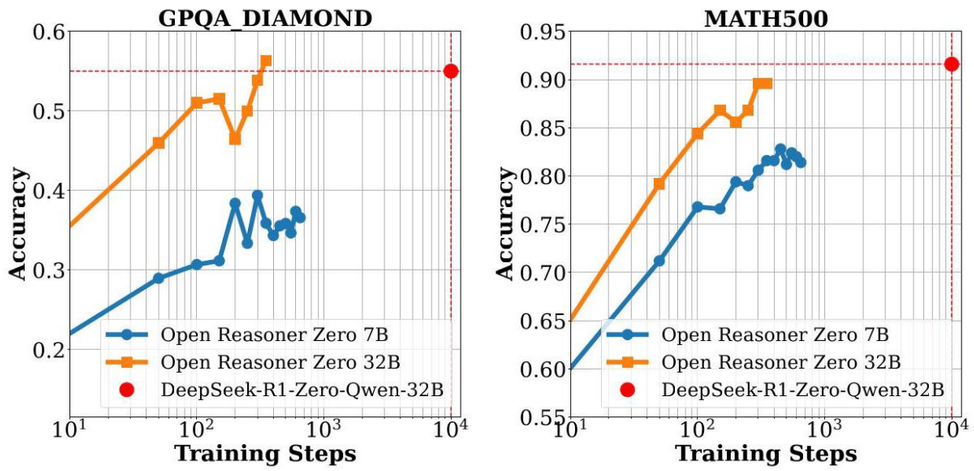

Figure 1: Evaluation performance of Open-Reasoner-Zero-{7B, 32B}. We report the average accuracy on the benchmarks for each question with 16 responses. Notably, Open-Reasoner-Zero- 32B outperforms DeepSeek-R1-Zero-Qwen-32B on the GPQA Diamond benchmark while only requiring 1/30 of the training

摘要

* 警告：该 PDF 由 GPT-Academic 开源项目调用大语言模型 +Latex 翻译插件一键生成，版权归原文作者所有。翻译内容可靠性无保障，请仔细鉴别并以原文为准。项目 Github 地址 https://github.com/binary-husky/gpt_academic/。项目在线体验地址 https://auth.gpt-academic.top/。当前大语言模型：Qwen2.5-72B-Instruct，当前语言模型温度设定：0.3。为了防止大语言模型的意外谬误产生扩散影响，禁止移除或修改此警告。
GPT-Academic 程序无法在本文中找到摘要部分。

Open-Reasoner-Zero: An Open Source Approach to Scaling Up Reinforcement Learning on the Base Model

Jingcheng Hu [1,2*] , Yinmin Zhang [1] , Qi Han [1] , Daxin Jiang [1] , Xiangyu Zhang [1] ,
Heung-Yeung Shum2

[1] StepFun, [2] 清华大学

GitHub: https://github.com/Open-Reasoner-Zero/Open-Reasoner-Zero
HuggingFace: https://huggingface.co/Open-Reasoner-Zero

Abstract

我们介绍了 Open-Reasoner-Zero，这是第一个开源的大规模推理导向的强化学习训练实现，专注于可扩展性、简单性和易用性。通过广泛的实验，我们证明了一种极简的方法，即使用 GAE ($\lambda = 1, \gamma = 1$) 的纯 PPO 和简单的基于规则的奖励函数，无需任何 KL 正则化，就足以在推理任务上扩展响应长度和基准性能，类似于在 DeepSeek-R1-Zero 中观察到的现象。值得注意的是，我们的实现不仅在 GPQA Diamond 基准上超过了 DeepSeek-R1-Zero-Qwen-32B，而且仅需其 1/30 的训练步骤。秉承开源精神，我们发布了源代码、参数设置、训练数据和模型权重。* 实习期间在 StepFun 完成的工作。内容

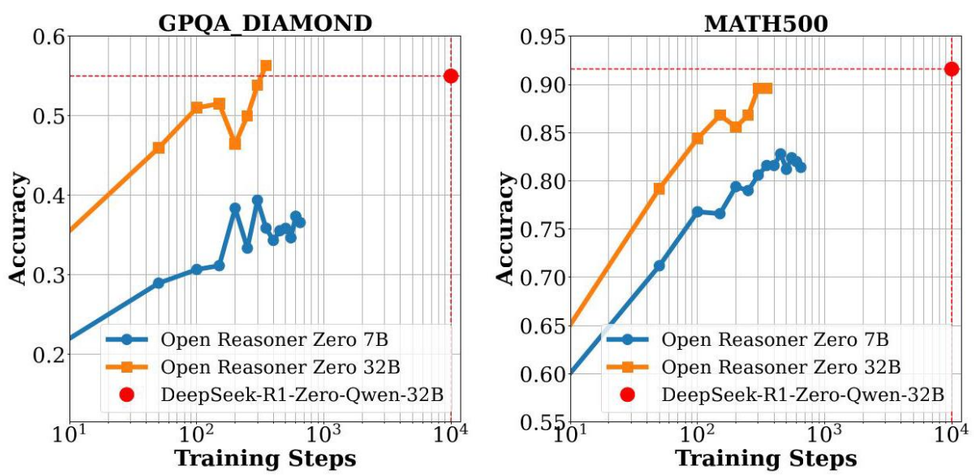

图 1：Open-Reasoner-Zero-{7B, 32B} 的评估性能。我们报告了每个问题在 16 个响应下的基准测试平均准确率。值得注意的是，Open-Reasoner-Zero-32B 在 GPQA Diamond 基准测试中优于 DeepSeek-R1-Zero-Qwen-32B，而仅需 1/30 的训练步骤。我们将在预印本发布前继续扩大这些强化学习设置的规模，因为目前还没有出现饱和的迹象。

1. Introduction

Large-scale reinforcement learning (RL) training of language models on reasoning tasks has emerged as a promising paradigm for mastering complex problem-solving skills. Recent breakthroughs, particularly OpenAI's o1 [1] and DeepSeek's R1-Zero [2], have demonstrated remarkable training time scaling phenomenon: as the training computation scales up, both the model's benchmark performance and response length consistently and steadily increase without any sign of saturation. Inspired by these advancements, we aim to explore this new scaling phenomenon by conducting large-scale RL training directly on base models, an approach we refer to as Reasoner-Zero training.

In this work, we introduce Open-Reasoner-Zero (ORZ), the first open-source implementation of large-scale reasoning-oriented RL training on large language models (LLMs) with our best practices, designed to be robust, scalable and simple-to-follow. Under Reasoner-Zero paradigm, LLMs are trained to master diverse reasoning skills under verifiable rewards, spanning arithmetic, logic, coding and commonsense reasoning (e.g., scientific problems, numerical reasoning, natural language understanding and even creative writing). While DeepSeek's R1-Zero outlined their training pipeline briefly, we provide a comprehensive study of our training strategy, with in-depth insights into overcoming common challenges such as training instability, stagnating response length, benchmark performance plateaus, and reward design. Our goal is to democratize advanced RL training techniques accessible to the broader research community.

Our proposed Open-Reasoner-Zero-32B outperforms the DeepSeek-R1-Zero-Qwen-32B, with the same Qwen-32B base model, on GPQA Diamond benchmark, yet requires 1/30 iterations. We have conducted tens of thousands of iterations to explore our best practical setting. Through extensive ablation studies, we summarize some key findings and lessons learned from our exploration. Specifically, vanilla PPO using GAE ($\lambda = 1$ and $\gamma = 1$) and without any KL-related regularization, combined with a straightforward rule-based reward function, is sufficient to achieve steady scalability in both response length and benchmark performances across varying model sizes and training data scales. Open-Reasoner-Zero's stable scaling resonates well with the bitter lesson [3]: the most significant performance improvements stem

1. Introduction

大规模强化学习（RL）在推理任务上训练语言模型已成为掌握复杂问题解决技能的有希望的范式。最近的突破，特别是 OpenAI 的 o1 [1] 和 DeepSeek 的 R1-Zero [2]，展示了显著的训练时间扩展现象：随着训练计算的扩展，模型的基准性能和响应长度持续且稳定地增加，没有任何饱和的迹象。受这些进展的启发，我们旨在通过直接在基础模型上进行大规模 RL 训练来探索这种新的扩展现象，我们称之为 Reasoner-Zero 训练。

在本工作中，我们介绍了 Open-Reasoner-Zero (ORZ)，这是第一个大规模面向推理的 RL 训练在大型语言模型（LLMs）上的开源实现，结合了我们的最佳实践，旨在稳健、可扩展且易于遵循。在 Reasoner-Zero 范式下，LLMs 被训练以在可验证的奖励下掌握多样化的推理技能，涵盖算术、逻辑、编码和常识推理（例如，科学问题、数值推理、自然语言理解和甚至创意写作）。虽然 DeepSeek 的 R1-Zero 简要概述了他们的训练管道，我们提供了我们训练策略的全面研究，深入探讨了克服常见挑战（如训练不稳定性、响应长度停滞、基准性能平台期和奖励设计）的见解。我们的目标是使先进的 RL 训练技术为更广泛的科研社区所用。

我们提出的 Open-Reasoner-Zero-32B 在 GPQA Diamond 基准上超越了 DeepSeek-R1-Zero-Qwen-32B，尽管使用了相同的 Qwen-32B 基础模型，但仅需 1/30 的迭代次数。我们进行了数万次迭代以探索最佳实践设置。通过广泛的消融研究，我们总结了一些关键发现和从探索中获得的经验教训。具体而言，使用 GAE ($\lambda = 1$ 和$\gamma = 1$) 的纯 PPO，不使用任何与 KL 相关的正则化，结合一个简单的基于规则的奖励函数，足以在不同模型大小和训练数据规模下实现响应长度和基准性能的稳定扩展。Open-Reasoner-Zero 的稳定扩展与苦涩的教训 [3] 相呼应：最显著的性能提升来源于训练数据量、模型大小和训练迭代次数的规模，而不是设计选择的复杂性。最关键的是如何设计一个简单而有效的 RL 算法来扩大训练过程。

我们很高兴与科研社区分享这一规模扩展 RL 的突破，以及我们从中获得的经验教训，使每个人不仅能够利用最终成果（例如，API 或模型权重），还能亲身体验并参与这一 AI 发展中的变革时刻。为了帮助促进可重复性和进一步研究直接在 LLMs 上进行大规模 RL 训练，我们承诺发布所有训练资源，包括代码、参数、数据和模型权重。

我们的主要贡献如下：

1. 我们提供了一个完全开源的大规模 RL 训练直接在基础 LLM 上的实现，我们称之为 Open-Reasoner-Zero。

2. 我们分享了在扩展过程中经历的令人沮丧的失败和令人兴奋的突破中的经验教训。

from the scale of training data, model size, and training iterations, rather than the complexity of design choices. The most critical thing is how to design a simple and effective RL algorithm to scale up the training process.

We are excited to share this breakthrough of scale-up RL, along with the lessons learnt with the research community, enabling everyone to not only utilize the end results (e.g., APIs or model weights), but to experience and participate in this transformative moment in AI development themselves. To help facilitate reproducibility and further research in large-scale RL directly training on LLMs, we are committed to release all of our training resources, including code, parameters, data, and model weights.

Our primary contributions are as follow:

1. We provide a fully open-source implementation of large-scale RL training directly on a base LLM, a strategy we refer to as Open-Reasoner-Zero.

2. We share empirical insights and lessons learned from frustrating failures and exciting breakthroughs during our scaling-up journey.

3. We release comprehensive training code, parameter settings, data, and model weights to the research community.

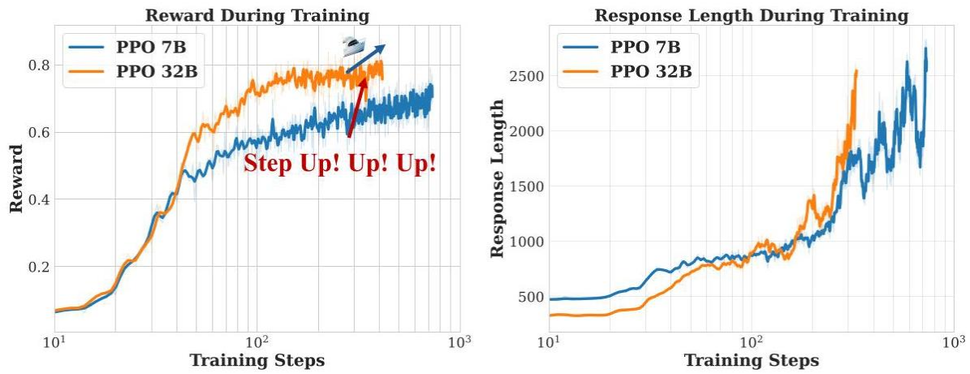

Figure 2: Train Time Scale up on Reward and Response Length of Open-Reasoner-Zero-7B, 32B. The training is still ongoing, and shows no sign of collapse.

2. Scale-up Reinforcement Learning from a Base Model

In this section, we describe the strategy and critical components for scale-up reasoning-oriented reinforcement learning (RL) directly from a base model. First, we introduce the basic yet critical settings for our scale-up RL training from a base model, including data curation, reward function, and detailed settings of the Proximal Policy Optimization (PPO) [4] algorithm. We then discuss key insights derived from our comprehensive ablation experiments that enable successful scale-up RL training.

2.1. Basic Settings

We conduct our experiments utilizing the Qwen2.5-{7B, 32B} as our base model [5], and directly starting the large-scale RL training without any fine-tuning (e.g., distillation or SFT) [6, 7]. Building upon the Qwen2.5-{7B, 32B} base model, we scale up the standard PPO algorithm [4] for reasoning-oriented

3. 我们向科研社区发布了全面的训练代码、参数设置、数据和模型权重。

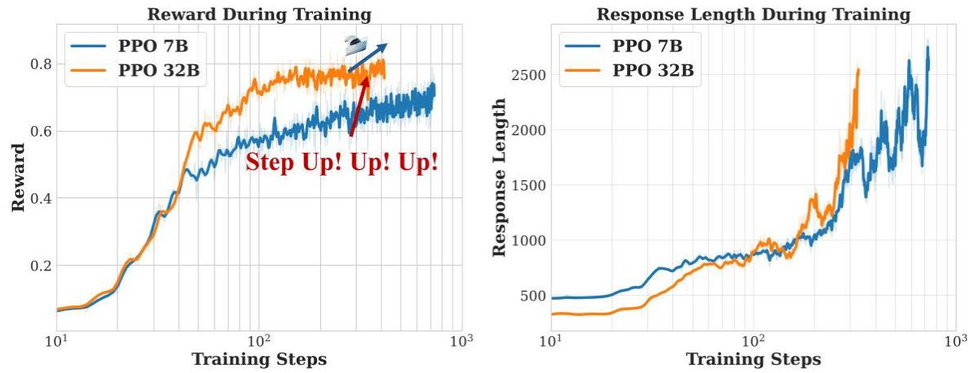

图 2: Open-Reasoner-Zero-7B, 32B 的奖励和响应长度随训练时间的增加。训练仍在进行中，且没有显示出崩溃的迹象。

2. Scale-up Reinforcement Learning from a Base Model

在本节中，我们描述了直接从基础模型进行规模扩展的以推理为导向的强化学习（RL）的策略和关键组件。首先，我们介绍从基础模型进行规模扩展的 RL 训练的基本但关键的设置，包括数据整理、奖励函数以及近端策略优化（PPO）[4] 算法的详细设置。然后，我们讨论了从全面的消融实验中得出的关键见解，这些见解使成功的规模扩展 RL 训练成为可能。

2.1. Basic Settings

我们利用 Qwen2.5-{7B, 32B} 作为基础模型 [5] 进行实验，并且在没有任何微调（例如，蒸馏或 SFT）的情况下直接开始大规模的强化学习训练 [6, 7]。基于 Qwen2.5-{7B, 32B} 基础模型，我们扩展了标准的 PPO 算法 [4]，用于以推理为导向的强化学习训练，同时仔细考虑了可扩展性和鲁棒性。我们的训练数据包括数万个精心策划的问题和答案对，涵盖 STEM、数学和推理任务，专门设计用于增强模型在多样和复杂问题解决场景中的能力。受 DeepSeek-R1 [2] 的启发，我们设计了提示模板，以激发模型利用推理计算，逐步掌握复杂任务的推理能力，如表 1 所示。此外，我们基于 OpenRLHF [8] 开发了一个高效且易于使用的大规模强化学习训练框架，通过引入更灵活的训练器，实现 GPU 共置生成，并支持卸载和回载训练。在接下来的章节中，我们将提供从基础模型扩展强化学习训练的详细设置。

2.1.1. Dataset

在本节中，我们介绍我们精心策划的数据集，详细说明其来源描述、清理过程和未来方向的扩展见解。高质量的训练数据对于可扩展的 Reasoner-Zero 训练至关重要。我们在数据配方中确定了三个关键方面：数量、多样性和质量。遵循这些关键方面，我们通过全面的收集和清理过程策划了我们的数据集：用户和助手之间的对话。用户提出问题，助手解答。助手首先在心中思考推理过程，然后向用户提供答案。推理过程和答案分别被包含在 <think> </think> 和 <answer> </answer> 标签中，即 <think> 这里是推理过程 </think> <answer> 这里是答案 </answer>。用户：您必须将答案放在 <answer> </answer> 标签内，即 <answer> 这里是答案 </answer>。您的最终答案将通过\boxed{} 标签自动提取。{{prompt}}

Assistant: <think>

RL training, with careful consideration of scalability and robustness. Our training data comprises tens of thousands of carefully curated question and answer pairs consisting of STEM, Math, and Reasoning tasks, designed specifically for enhancing models' capability in diverse and complex problem-solving scenarios. Inspired by DeepSeek-R1 [2], we design our prompt template to elicit the model to utilize inference computation, gradually mastering the reasoning ability for complex tasks, as shown in Table 1. Furthermore, we develop an efficient and easy-to-use large-scale RL training framework based on OpenRLHF [8], by introducing a more flexible trainer, enabling GPU collocation generation, and training with offload and backload support. In the following sections, we provide detailed settings for our scale-up RL training from a base model.

2.1.1. Dataset

In this section, we introduce our carefully curated dataset, detailing its source description, cleaning process, and scaling insights for future directions. High-quality training data are crucial for scalable Reasoner-Zero training. We identify three key aspects in our data recipe: quantity, diversity, and quality. Following these key aspects, we curate our dataset through a comprehensive collection and cleaning process: A conversation between User and Assistant. The user asks a question, and the Assistant solves it. The assistant first thinks about the reasoning process in the mind and then provides the user with the answer. The reasoning process and answer are enclosed within <think> </think> and <answer> </answer> tags, respectively, i.e., <think> reasoning process here </think> <answer> answer> answer here </answer>. User: You must put your answer inside <answer> </answer> tags, i.e., <answer> answer> answer here </answer>. And your final answer will be extracted automatically by the \boxed{} tag. {{prompt}}
Assistant: <think>

Table 1: Template for Open-Reasoner-Zero. prompt will be replaced with the specific reasoning question during training.

- We collect public data from various sources, including AIME (up to 2023), MATH, Numina-Math collection [9], Tulu3 MATH [10], and other open-source datasets. Based on source and problem difficulty, we retrieve AMC, AIME, Math, Olympiads, and AoPS forum components as our difficult level prompts to ensure appropriate difficulty levels.
- We synthesize additional reasoning tasks using programmatic approaches to augment the dataset.
- We exclude problems that are challenging to evaluate with our rule-based reward function, such as multiple-choice and proof-oriented problems, ensuring accurate and consistent reward computation during training.
- We implement a model-based filtering strategy based on heuristic evaluation of problem difficulty. Specifically, we use LLM to assess the pass rate of each problem, removing samples with either too high or zero pass rates.
- We apply N-gram and embedding similarity-based filtering to deduplicate samples and maintain data diversity.

Table 1: Template for Open-Reasoner-Zero. prompt will be replaced with the specific reasoning question during training.

- 我们从各种来源收集公开数据，包括 AIME（截至 2023 年）、MATH、Numina-Math 集合 [?]、Tulu3 MATH[?] 和其他开源数据集。根据来源和问题难度，我们检索 AMC、AIME、Math、奥林匹克数学和 AoPS 论坛的组件作为我们的难度提示，以确保适当的难度水平。
- 我们使用程序化方法合成额外的推理任务以扩充数据集。
- 我们排除了使用基于规则的奖励函数难以评估的问题，例如选择题和证明题，以确保在训练过程中奖励计算的准确性和一致性。
- 我们基于问题难度的启发式评估实施了一种基于模型的过滤策略。具体来说，我们使用大型语言模型（LLM）来评估每个问题的通过率，移除通过率过高或为零的样本。
- 我们应用 N-gram 和嵌入相似性过滤来去重样本并保持数据多样性。

最终整理的数据集包含大约 57,000 个样本，涵盖了 STEM、数学和推理领域。该数据集专门设计用于增强模型在复杂问题解决任务中的能力，精心平衡了数量、多样性和质量。更多详细讨论见附录。未来，我们计划通过与研究社区合作，鼓励研究人员自愿贡献各种领域的额外数据，从高级数学和推理任务到竞赛编程和软件工程任务，以扩展我们的数据集。

2.1.2. Reward Function

与 DeepSeek-R1-Zero [2] 不同，我们的规模扩展强化学习（RL）训练采用了一个简单的基于规则的奖励函数，该函数仅检查答案的正确性，而不提供任何额外的格式奖励。具体来说，这个奖励函数设计为在训练期间提取'<answer>' 和'</answer>' 标签之间的内容，并将其与参考答案进行比较。为了保持规模扩展 RL 的清晰和简洁，我们实施了一个二元奖励方案——对于与参考答案完全匹配的情况奖励 1，其他所有情况奖励 0。为了确保评估的严格性和一致性，我们采用了广泛使用的 Math-Verify 库，其使用方法如图 3 所示。

令人惊讶的是，我们发现，即使未对齐的基础模型在高概率下也能产生格式良好的响应。在训练的早期阶段，基础模型可以快速学习并强化由我们的简单基于规则的奖励函数激励的正确推理和回答格式。

```
from math_verify import verify, parse
verify(parse(ground_truth), parse(model_output))
```

图 3：使用 Math-Verify 库验证生成答案数学正确性的代码片段。

更重要的是，我们的初步实验表明，复杂的奖励函数不仅没有必要，还可能留下奖励操纵的潜在空间。

https://github.com/huggingface/Math-Verify

The final curated data consists of approximately 57k samples spanning STEM, mathematics, and reasoning domains. This collection is specifically designed to enhance models' capabilities in complex problem-solving tasks, carefully balancing quantity, diversity, and quality. More detailed discussions are provided in the appendix. In the future, we aim to expand our dataset by collaborating with the research community to encourage researchers to voluntarily contribute additional data across various domains, from advanced mathematics and reasoning tasks to competitive programming and software engineering tasks.

2.1.2. Reward Function

Unlike DeepSeek-R1-Zero [2], our scale-up RL training employs a simple minimalist rule-based reward function that solely checks answer correctness, without any additional format rewards. Specifically, this reward function is designed to extract the content between '<answer>' and '</answer>' tags during training and compare it with the reference answer. To maintain clarity and simplicity in scale-up RL, we implement a binary reward scheme - awarding a reward of 1 for exact matches with the reference answer, and 0 for all other cases. To ensure rigorous and consitent assessment in evaluation, we adopt the widely-used Math-Verify library and its usage as shown in Figure 3.

Surprisingly, we found that with our designed prompt, even unaligned base model can yield well-formatted responses in high probability. During early training stages, the base model can quickly learn and reinforce the correct format for reasoning and answering incentivized

```
from math_verify import verify, parse
verify(parse(ground_truth), parse(model_output))
```

Figure 3: The code snippet for verifying the mathematical correctness of generated answers using the Math-Verify library.

by our simple rule-based reward function alone, as shown in Figure 4. More importantly, our preliminary experiments revealed that complicated reward functions were not only unnecessary, but could leave potential room for reward hacking.

2.1.3.RL Algorithm

We adopt the Proximal Policy Optimization (PPO) algorithm [4] as the RL algorithm for our scale-up training, unlike GRPO used in DeepSeek-R1-Zero. Specifically, for each question q (i.e., prompt), the model generates a group of responses $\{o_1, o_2, \ldots, o_n\}$ and receives corresponding rewards $\{r_1, r_2, \ldots, r_n\}$ based on the rule-based reward function, where n represents the number of sampled trajectories (i.e., rollout size per prompt). For each response o_i at time step t (i.e., token t), let s_t denote the state at time t which comprises the question and all previously generated tokens, and a_t denote the token generated at that step. We compute the advantage estimation $\widehat{A}_t$ for each token using Generalized Advantage Estimation (GAE) [11]. Generally, GAE provides a trade-off between bias and variance in the advantage

https://github.com/huggingface/Math-Verify

2.1.3.RL Algorithm

我们采用近端策略优化（PPO）算法 [4] 作为我们扩展训练的强化学习（RL）算法，不同于 DeepSeek-R1-Zero 中使用的 GRPO。具体来说，对于每个问题 q（即提示），模型生成一组响应 $\{o_1, o_2, \ldots, o_n\}$，并根据基于规则的奖励函数接收相应的奖励 $\{r_1, r_2, \ldots, r_n\}$，其中 n 表示采样轨迹的数量（即每个提示的 rollout 大小）。对于每个在时间步 t（即 token t）的响应 o_i，令 s_t 表示包含问题和所有先前生成的 token 的时间步 t 的状态，a_t 表示该步骤生成的 token。我们使用广义优势估计（GAE）[11] 为每个 token 计算优势估计 $\widehat{A}_t$。通常，GAE 通过参数 λ 控制的指数加权平均结合多个 n 步优势估计，提供优势估计中的偏差和方差之间的权衡。优势计算为 $\widehat{A}_t = \delta_t + (\gamma\lambda)\,\delta_{t+1} + \ldots + (\gamma\lambda)^{T-t-1}\delta_{T-1}$，其中 $\delta_t = r_t + \gamma V\left(s_{t+1}\right) - V\left(s_t\right)$ 是 TD（时间差分）残差，γ 是确定未来奖励相对于即时奖励的价值的折扣因子。PPO 算法通过优化以下目标函数来更新策略模型参数 θ 以最大化预期奖励，并更新价值模型参数 ϕ 以最小化价值损失：

$$\mathcal{J}_{\mathrm{PPO}}(\theta) = \mathbb{E}_{t, s_t, a_t \sim \pi_{\theta_{\text{ old}}}}\left[\min\left(\frac{\pi_\theta\left(a_t \mid s_t\right)}{\pi_{\theta_{\text{ old}}}\left(a_t \mid s_t\right)}\widehat{A}_t, \operatorname{clip}\left(\frac{\pi_\theta\left(a_t \mid s_t\right)}{\pi_{\theta_{\text{ old}}}\left(a_t \mid s_t\right)}, 1-\epsilon, 1+\epsilon\right)\widehat{A}_t\right)\right], \tag{1}$$

$$\mathcal{J}_{\text{value }}(\phi) = \frac{1}{2}\mathbb{E}_{t, s_t, a_t \sim \pi_{\theta_{\text{ old}}}}\left[\left(V_\phi\left(s_t\right) - R_t\right)^2\right], \tag{2}$$

其中 ϵ 是裁剪参数，π_θ 是当前策略，$\pi_{\theta_{\text{ old}}}$ 是更新前的旧策略，V_ϕ 是价值函数，$R_t = \sum_{k=0}^{T-t} \gamma^k r_{t+k}$ 是折扣回报。我们使用精心调整的超参数实例化 PPO 算法：GAE 参数 $\lambda = 1.0$ ，折扣因子 $\gamma = 1.0$ ，裁剪参数 $\epsilon = 0.2$ 。

2.2.Key Findings

在本研究中，我们探讨了以推理为导向的强化学习（RL）训练的最佳实践，重点是稳定性和可扩展性。我们在 Reasoner-Zero 训练的设计空间中进行了广泛的实验。以下是我们的实验得出的关键发现：

- RL 算法关键实现：我们的实证研究表明，原始的 PPO 在不同的模型规模和训练时长下提供了非常稳定和强大的训练过程，而无需额外的修改。通过广泛的实验，我们确定 GAE 参数在 PPO 中对推理任务起着关键作用。具体来说，尽管在传统的 RL 场景中通常认为设置 $\lambda = 1.0$ 和 $\gamma = 1.0$ 是次优的，但在规模扩展的 RL 训练中，这种设置达到了理想平衡。

estimation by combining multiple n-step advantage estimates through an exponentially weighted average controlled by the parameter λ . The advantage is computed as $\widehat{A}_t = \delta_t + (\gamma\lambda)\,\delta_{t+1} + \ldots + (\gamma\lambda)^{T-t-1}\delta_{T-1}$, where $\delta_t = r_t + \gamma V\left(s_{t+1}\right) - V\left(s_t\right)$ is the TD (temporal difference) residual and γ is the discount factor that determines how much future rewards are valued relative to immediate rewards. The PPO algorithm updates the policy model parameters θ to maximize the expected reward and value model parameters ϕ to minimize the value loss by optimizing the following objective function:

$$\mathcal{J}_{\mathrm{PPO}}(\theta) = \mathbb{E}_{t,s_t,a_t\sim\pi_{\theta_{\text{ old}}}}\left[\min\left(\frac{\pi_\theta\left(a_t \mid s_t\right)}{\pi_{\theta_{\text{ old}}}\left(a_t \mid s_t\right)}\widehat{A}_t, \operatorname{clip}\left(\frac{\pi_\theta\left(a_t \mid s_t\right)}{\pi_{\theta_{\text{ old}}}\left(a_t \mid s_t\right)}, 1-\epsilon, 1+\epsilon\right)\widehat{A}_t\right)\right], \tag{1}$$

$$\mathcal{J}_{\text{ value }}(\phi) = \frac{1}{2}\mathbb{E}_{t,s_t,a_t\sim\pi_{\theta_{\text{ old}}}}\left[\left(V_\phi\left(s_t\right) - R_t\right)^2\right], \tag{2}$$

where ϵ is the clipping parameter, π_θ is the current policy, $\pi_{\theta_{\text{ old}}}$ is the old policy before the update, V_ϕ is the value function, and $R_t = \sum_{k=0}^{T-t}\gamma^k r_{t+k}$ is the discounted return. We instantiate the PPO algorithm with carefully tuned hyperparameters: GAE parameter $\lambda = 1.0$, discount factor $\gamma = 1.0$, and clipping parameter $\epsilon = 0.2$.

2.2.Key Findings

In this study, we explore best practices for reasoning-oriented RL training with an emphasis on stability and scalability. We conduct extensive experiments across the design space of Reasoner-Zero training. Here are the key findings from our experiments:

- RL Algorithm Key Implementations: Our empirical studies demonstrate that vanilla PPO provides a remarkably stable and robust training process across different model scales and training duration without requiring additional modifications. Through extensive experiments, we identified that the GAE parameters play a critical role in PPO for reasoning tasks. Specifically, setting $\lambda = 1.0$ and $\gamma = 1.0$, while typically considered suboptimal in traditional RL scenarios, achieves the ideal balance for scale-up RL training.

图 4：遵循推理格式的响应百分比。结果表明，即使基础模型仅使用简单的基于规则的奖励函数，也能迅速采用结构化的推理模式。我们的研究结果表明，复杂的奖励函数对于训练 Reasoner-Zero 模型并非必要。

- 最小奖励函数设计：我们证明了一个简单的基于规则的奖励函数不仅是足够的，而且是最优的，因为最小设计没有为潜在的奖励操纵留下空间。值得注意的是，即使是未对齐的基础模型也能迅速适应所需的格式，这表明这是一个无需复杂奖励工程的简单任务。

- 损失函数：我们实现了稳定的训练，而无需依赖任何基于 KL 的正则化技术（例如，KL 形状的奖励和损失），这与事实上的 RLHF 社区 [12] 和 Reasoner 模型 [13, 2] 不同。这还为进一步大规模 RL 提供了有希望的潜力。

- 扩大训练数据：我们发现，扩大数据量和多样性对于 Reasoner-Zero 的训练至关重要。虽然在有限的学术数据集（如 MATH）上训练会导致性能迅速达到瓶颈，但我们精心策划的大型多样化数据集使得训练和测试集在性能上都能持续提升，没有出现饱和的迹象。

3. Experiments

在本节中，我们展示了 Open-Reasoner-Zero 模型的全面实验结果和分析。我们首先介绍训练设置和超参数，然后进行深入的训练结果分析和消融研究。接下来，我们从两个方面探讨初步结果：利用我们训练的推理模型进行蒸馏，以及采用 Open-Reasoner-Zero 训练管道对蒸馏后的模型进行进一步训练以增强其推理能力，方法类似于 DeepSeek-R1 [2]。最后，我们讨论评估结果，并对训练过程进行详细分析。

3.1. Training Details and Hyperparameters

我们使用 Qwen-2.5 基础模型（7B 和 32B 变体）初始化我们的策略和评论家网络，其中值头从 $\mathcal{U}\left(-\sqrt{5},\sqrt{5}\right)$ 随机初始化，且没有偏置项。在训练过程中，策略和评论家网络不共享权重。对于策略和评论家网络，我们使用 AdamW 优化器，$\beta = [0.9, 0.95]$，且不使用权重衰减。策略和评论家网络的学习率分别设置为 1×10^{-6} 和 5×10^{-6}。学习率调度器均为常数学习率，线性预热 50 个优化器步。我们在训练过程中使用样本打包。

Figure 4: Percentage of responses following the reasoning format. Results demonstrate rapid adoption of structured reasoning patterns even by the base model using only a simple rule-based reward function. Our findings suggest that complicated reward functions are unnecessary for training Reasoner-Zero models.

- Minimal Reward Function Design: We show that a simple rule-based reward function is not only sufficient but optimal, as minimal design leaves no room for potential reward hacking. Notably, even unaligned base models quickly adpots to desired format, suggesting this is a straightforward task without requiring complex reward engineering.
- Loss Function: We achieve stable training without relying on any KL-based regularization techniques (e.g., KL shaped rewards and loss), different from the de facto RLHF community [12] and Reasoner model [13, 2]. This also offers promising potential for further large-scaling RL.
- Scale up Training Data: We identify that scaling up data quantity and diversity is crucial for Reasoner-Zero training. While training on limited academic datasets like MATH leads to quick performance plateaus, our curated large-scale diverse dataset enables continuous scaling without signs of saturation on both training and test sets.

3. Experiments

In this section, we present comprehensive experimental results and analysis of our Open-Reasoner-Zero models. We begin by the training setup and hyperparameters, followed by an in-depth analysis of traning results, and ablation studies. We then investigate the preliminary results on two fronts: leveraging our trained reasoner model for distillation, and employing the Open-Reasoner-Zero training pipeline on the distilled model to further enhance its resasoning capabilities, following an approach similar to DeepSeek-R1 [2]. Finally, we discuss the evaluation results and provide a detailed analysis of the training process.

3.1. Training Details and Hyperparameters

We initialize both our policy and critic networks with Qwen-2.5 base models (7B and 32B variants), where value head is random initialized from $\mathcal{U}\left(-\sqrt{5}, \sqrt{5}\right)$ with no bias term. The policy and critic do not share weights during training. For both policy and critic networks, we employ AdamW optimizer with $\beta = [0.9, 0.95]$ without weight decay. The learning rates are set to 1×10^{-6} and 5×10^{-6} for the policy and critic networks, respectively. The learning rate scheduler are both constant learning rate with linear warm-up of 50 optimizer steps. We employ sample packing during training.

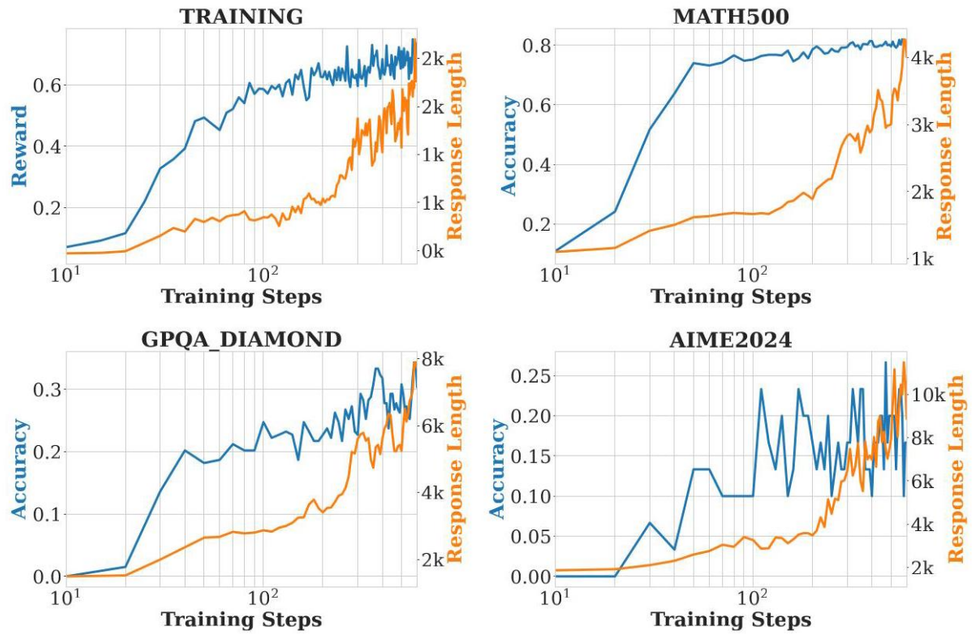

图 5：Open-Reasoner-Zero 7B 模型的训练和评估奖励及平均响应长度的比较。所有基准测试在某个点上都经历了奖励和响应长度的突然增加，这种现象类似于涌现行为。

每个生成步骤包含从数据集中采样的 128 个唯一提示，并为每个提示生成 64 个响应，温度和 top-p 均设置为 1.0。为了保持训练的稳定性，我们对策略网络实施严格的在线优化，每个生成步骤对应一个优化步骤。批评网络对离线更新的敏感度较低，因此在 12 个小批量中处理经验，每个迭代有效地执行 12 个优化步骤。我们在训练中应用了批级优势归一化。

值得注意的是，我们的训练过程在没有任何与 KL 相关的正则化项或熵奖励的情况下稳定运行，这表明纯 PPO 可以在没有这些常用稳定技术的情况下实现稳定的训练。

为了全面评估我们模型的推理能力，我们在涵盖数学推理、编程和一般问题解决的多样化基准上进行了实验。这些包括 GPQA DIAMOND [14]、AIME2024、AIME2025 [15]、MATH500 [16] 和 LIVECODEBENCH [17] 数据集。对于每个基准，我们报告每个问题 16 个样本的平均准确率作为主要评估指标。此外，我们还通过在 MMLU [18] 和 MMLU_PRO [19] 基准上的评估来评估模型的通用能力，以全面了解它们在各种任务中的表现。

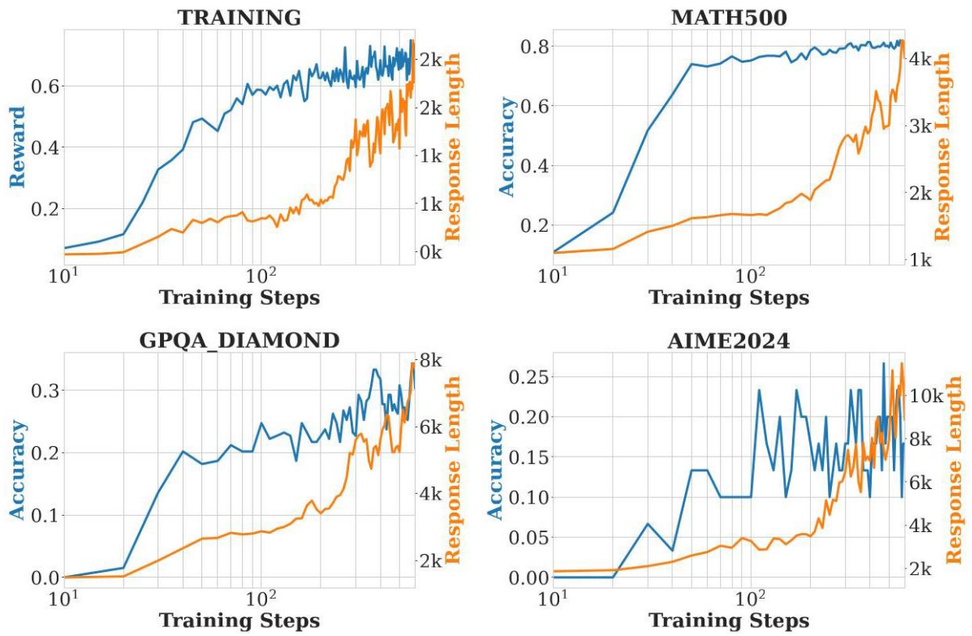

Figure 5: Comparison of training and evaluation reward and average response length for the Open-Reasoner-Zero 7B model. All of benchmarks experience a sudden increase in reward and response length at a certain point, a phenomenon like emergent behavior.

Each generation step contains 128 unique prompts sampled from the dataset, and generating 64 responses per prompt with temperature and top-p both set to 1.0 . To maintain training stability, we implement strict on-policy optimization for the policy network, where each generation corresponds to exactly one optimization step. The critic network, being less sensitive to off-policy updates, processes the experiences in 12 mini-batches, effectively performing 12 optimization steps per iteration. We apply batch level advantage normalization in the training.

Notably, our training process operates stably without any KL-related regularization terms or entropy bonuses, demonstrating that vanilla PPO can achieve stable training without these commonly used stabilization techniques.

To comprehensively evaluate our models' reasoning capabilities, we conduct experiments on diverse benchmarks spanning mathematical reasoning, coding, and general problem solving. These include GPQA DIAMOND [14], AIME2024, AIME2025 [15], MATH500 [16], and LIVE-CODEBENCH [17] datasets. For each benchmark, we report the average accuracy across 16 samples per question as our primary evaluation metric. Moreover, we also assess the models' general capabilities through evaluations on MMLU [18] and MMLU_PRO [19] benchmarks to provide a comprehensive understanding of their performance across diverse tasks.

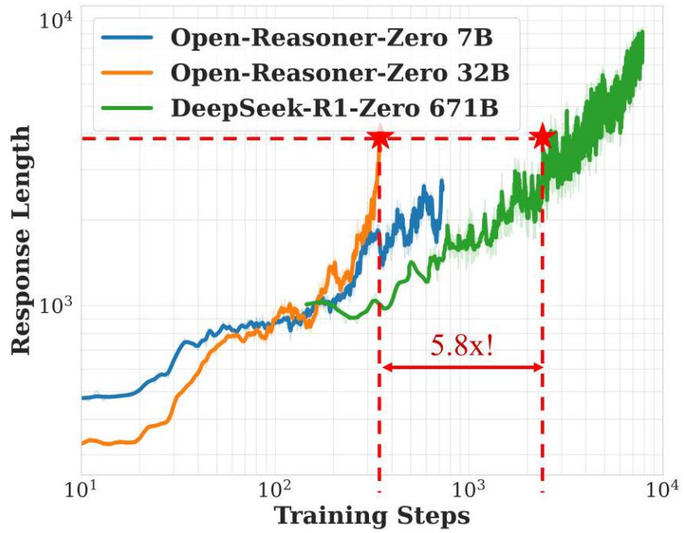

图 6：我们实验中的训练响应长度与训练步数的关系。我们的模型在整个训练过程中表现出响应长度的持续改进，类似于 DeepSeek-R1-Zero（671B MoE）。值得注意的是，我们的 Open-Reasoner-Zero-32B 模型在训练步数少 5.8 倍的情况下，达到了与 DeepSeek-R1-Zero（671B MoE）相当的响应长度。DeepSeek R1 Zero 的响应长度数据是从他们论文中的图 3 估算的。

3.2. Training Results

在本节中，我们展示了实验训练结果的关键发现。我们从多个角度评估了训练过程，包括训练集中的奖励、平均响应长度和生成质量指标。这些指标提供了模型性能和学习动态的全面视图。

训练曲线。图 2 显示了我们实验中 Open-Reasoner-Zero 7B 和 32B 的训练奖励和平均响应长度曲线，而图 5 显示了 Open-Reasoner-Zero 7B 在训练集和评估集上的奖励/准确性和平均响应长度曲线。训练奖励曲线和响应长度曲线分别表示每个生成步骤中生成响应的平均奖励和平均长度。我们观察到，这两个模型在所有基准测试中，这些指标在整个训练过程中持续改进，值得注意的是：Open-Reasoner-Zero 表现出一种引人注目的“跃进时刻”现象，即响应指标在训练过程中突然增加，揭示了新兴的推理能力。

响应长度扩展与 DeepSeek-R1-Zero 的比较。如图 6 所示，我们观察到响应长度在整个训练过程中持续增加，没有饱和的迹象，这与 DeepSeek-R1-Zero 的行为相似。值得注意的是，虽然模型大小和训练步骤都对响应长度的改进有贡献，但我们的 Open-Reasoner-Zero-32B 模型仅用 1/5.8 的训练步骤就达到了与 DeepSeek-R1-Zero（671B MoE）相当的响应长度。这种显著的训练效率展示了我们极简方法的有效性。

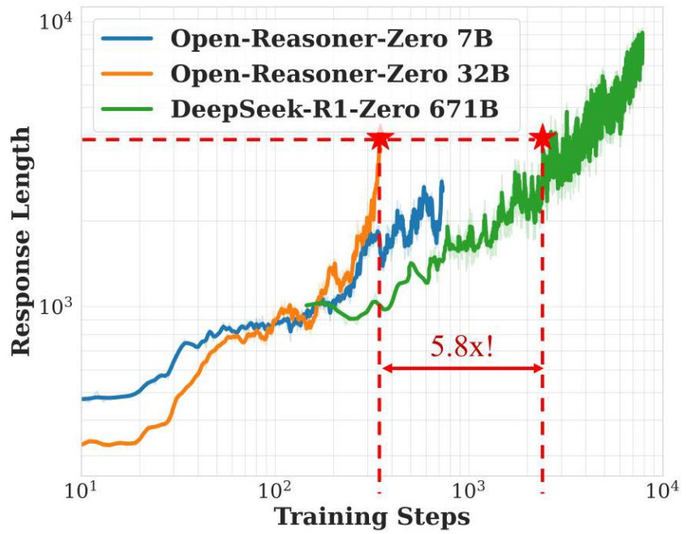

Figure 6: Training response length v.s training steps of our experiments. Our models demonstrate continuous improvements in response length throughout training, similar to DeepSeek-R1-Zero (671B MoE). Notably, our Open-Reasoner-Zero-32B model achieves comparable response lengths to DeepSeek-R1-Zero (671B MoE) with 5.8x fewer training steps. DeepSeek R1 Zero's response length data is estimated from Figure 3 in their paper.

3.2. Training Results

In this section, we present the key findings from our experimental training results. We evaluate the training process from multiple perspectives, including reward in training sets, average response lengths, and generation quality metrics. These metrics provide a holistic view of model performance and learning dynamics.

Training Curves. Figure 2 shows the training reward and average response length curves of our experiments for both Open-Reasoner-Zero 7B and 32B, while Figure 5 shows the reward/accuracy and average response length curves of our experienments for Open-Reasoner-Zero 7B on training and evaluation sets. The training reward curve and response length curve represent the average reward of the generated responses and the average length of the generated responses at each generation step, respectively. We observe consistent improvements in these metrics throughout training across both models and all benchmarks, with notable observations: Open-Reasoner-Zero exhibits an intriguing "step moment" phenomenon, where response metrics suddenly increase during training, revealing emergent reasoning capabilities.

Response Length Scale-up vs DeepSeek-R1-Zero. As shown in Figure 6, we observe a consistent increase in response length throughout training with no signs of saturation, mirroring the behavior seen in DeepSeek-R1-Zero. Notably, while both model size and training steps contribute to response length improvements, our Open-Reasoner-Zero-32B model achieves comparable response lengths to DeepSeek-R1-Zero (671B MoE) in just 1/5.8 of the training steps. This remarkable training efficiency demonstrates the effectiveness of our minimalist approach

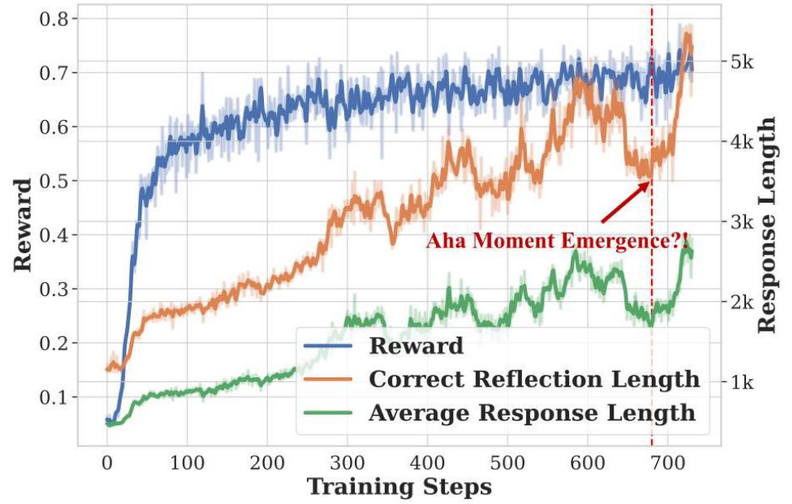

图 7：生成响应中的反射模式。在整个训练过程中，平均正确反射长度始终超过平均响应长度。一个特别值得注意的现象出现在大约第 680 步，我们观察到三个指标同时加速：训练集中的奖励、平均正确反射长度和平均响应长度。

to large-scale RL training.

质量分析。在这里，我们提供了我们 Open-Reasoner-Zero 模型生成响应的一些定性分析。为了分析模型的反思能力并观察类似于 DeepSeek-R1-Zero 的“顿悟”时刻，我们识别了五种代表性反思模式（“wait,”，‘”recheck”’，‘”retry”’，‘”alternatively,”’，和‘‘”however,”‘），遵循类似于 [20] 的方法。我们将包含这些模式的响应计为“反思响应”，并确定正确反思的平均长度（包含反思模式且达到正确答案的响应长度）。如图 7 所示，正确反思的平均长度在整个训练过程中始终超过平均响应长度，表明包含反思模式的响应利用了更多的“思考时间”来达到正确答案，类似于 OpenAI o1 中描述的测试时间尺度。一个特别值得注意的现象出现在大约第 680 步，我们观察到三个指标同时加速：奖励、正确反思的平均长度和平均响应长度。通过手动检查第 680 步前后的模型输出，我们观察到后者响应中的反思模式在定性上更加明显。这种新兴行为值得进一步研究，我们目前正在进行详细分析以理解这一现象的潜在机制。对于全面的定量和定性分析，请参阅我们在 Notion 2 上提供的详细文档。

3.3. Ablation Study

我们对关键训练策略和超参数进行了消融研究，这些策略和超参数使得从基础模型直接成功扩展强化学习训练成为可能。更全面的消融研究可在附录中找到。

[2] 全面的定量和定性分析。

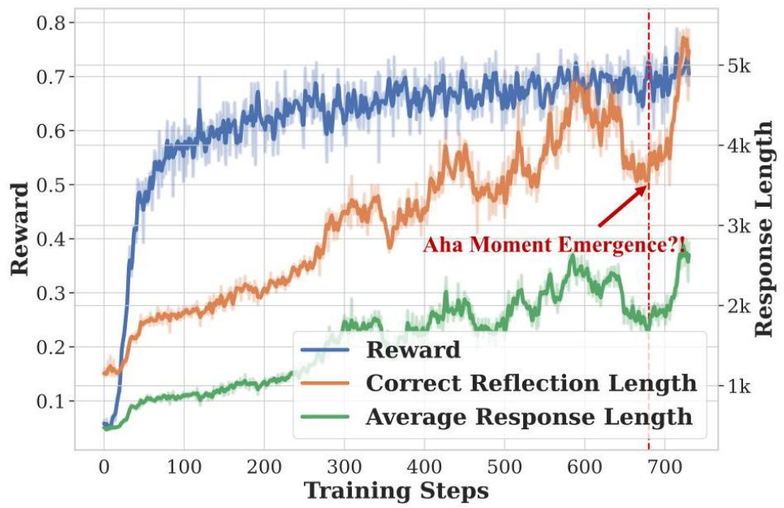

Figure 7: Reflection patterns in generated responses. The Average Correct Reflection Length consistently exceeds the Average Response Length throughout the training process. A particularly noteworthy phenomenon emerges around step 680, where we observe a simultaneous acceleration in three metrics: Reward in training set, Average Correct Reflection Length, and Average Response Length.

to large-scale RL training.

Quality Analysis. Here we provide some qualitative analysis of the generated responses from our Open-Reasoner-Zero models. To analyze the model's reflection capabilities and observe the Aha moment like DeepSeek-R1-Zero, we identify five representative reflection patterns ("wait,", ‘"recheck"’, ‘"retry"’, ‘"alternatively,"’, and ‘‘"however," ‘), following a methodology similar to [20]. We count the number of responses containing any of these patterns as 'reflection responses', and identify the average correct reflection length (the length of responses containing reflection patterns that achieve correct answers). As shown in Figure 7, the average correct reflection length consistently exceeds the average response length throughout the training process, indicating that responses containing reflection patterns utilize more "thinking time" to achieve correct answers, similar to the test-time scale described in OpenAI o1. A particularly noteworthy phenomenon emerges around step 680, where we observe a simultaneous acceleration in three metrics: the reward, average correct reflection length, and average response length. Through manual inspection of model outputs before and after step 680, we observed qualitatively more pronounced reflection patterns in the latter responses. This emergent behavior warrants further investigation, and we are currently conducting detailed analyses to understand the underlying mechanisms of this phenomenon. For comprehensive quantitative and qualitative analyses, please refer to our detailed documentation available at Notion 2

3.3. Ablation Study

We present ablation studies over key training strategies and hyperparameters that enable successful scaling of RL training directly from a base model. More comprehensive ablation studies are available in

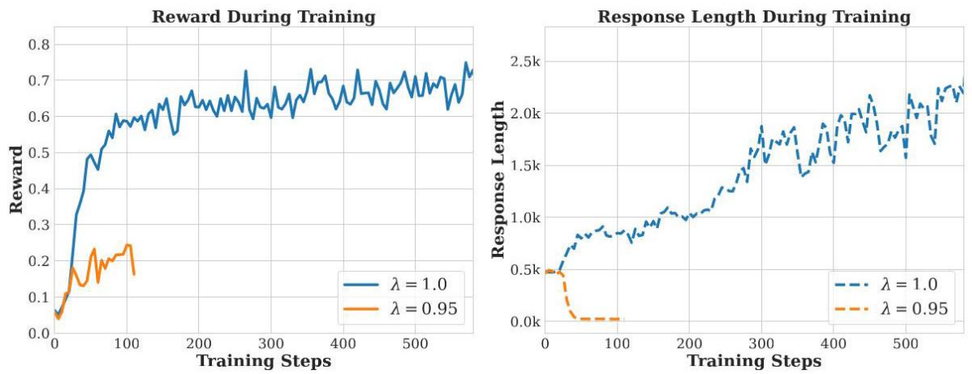

图 8: 不同 GAE λ 值的比较。GAE $\lambda = 1.0$ 在训练奖励和响应长度方面都表现出比 $\lambda = 0.95$ 更好的稳定性和性能。

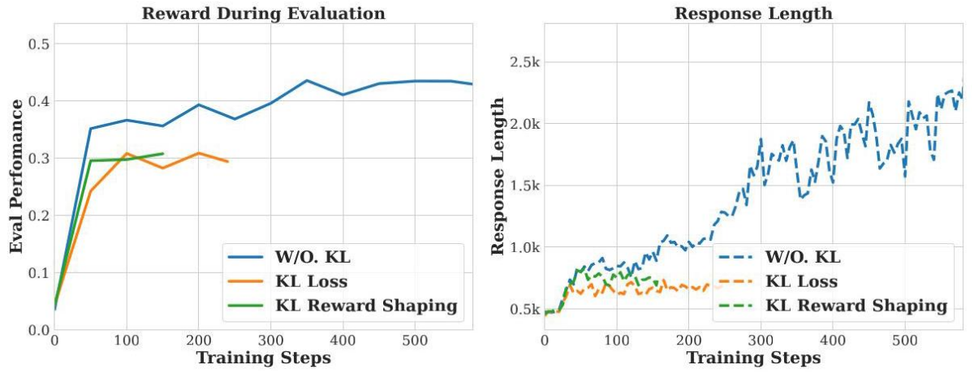

图 9：应用 KL 相关正则化的比较。值得注意的是，没有 KL 约束的训练在平均基准性能和长度缩放属性方面优于使用 KL 损失和 KL 惩罚训练的模型。性能评估使用 pass@1 指标在 MATH500、AIME2024 和 GPQA DIAMOND 基准上进行。

GAE 分析。我们比较了不同的 GAE λ 组合。从实验结果来看，GAE $\lambda = 1.0$ 在训练稳定性和最终性能方面表现最佳。具体来说，在训练奖励中，GAE $\lambda = 1.0$ 曲线在早期迅速上升并保持稳定，最终收敛到约 0.8；而 GAE $\lambda = 0.95$ 曲线上升缓慢且波动。在响应长度方面，GAE Lambda=1.0 曲线在训练过程中保持合理水平；而 GAE $\lambda = 0.95$ 曲线表现出不稳定趋势，导致 PPO 学习不稳定。这些结果表明，GAE $\lambda = 1.0$ 可以更好地平衡训练稳定性和生成质量。此外，折扣因子 (γ) 设置为 1.0 对大规模 RL 训练也有显著影响。小于 1.0 会导致对长期奖励的惩罚，导致响应长度减少，难以提高最终性能。

appendix.

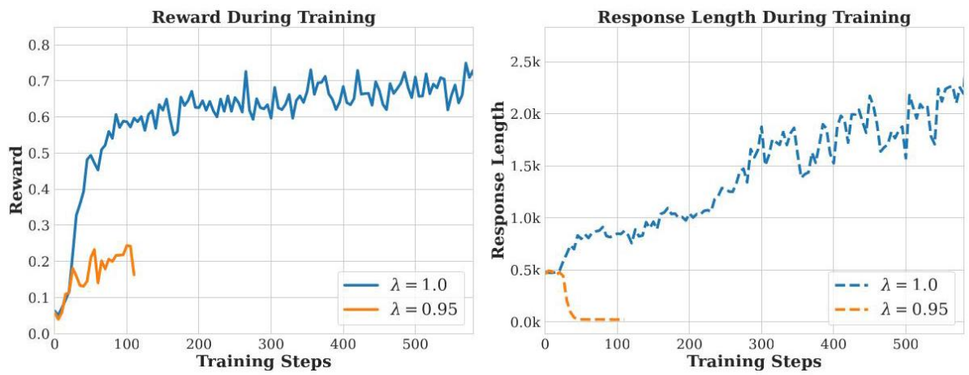

Figure 8: Comparison of different GAE λ values. GAE $\lambda = 1.0$ shows better stability and performance compared to $\lambda = 0.95$ for both training reward and response length.

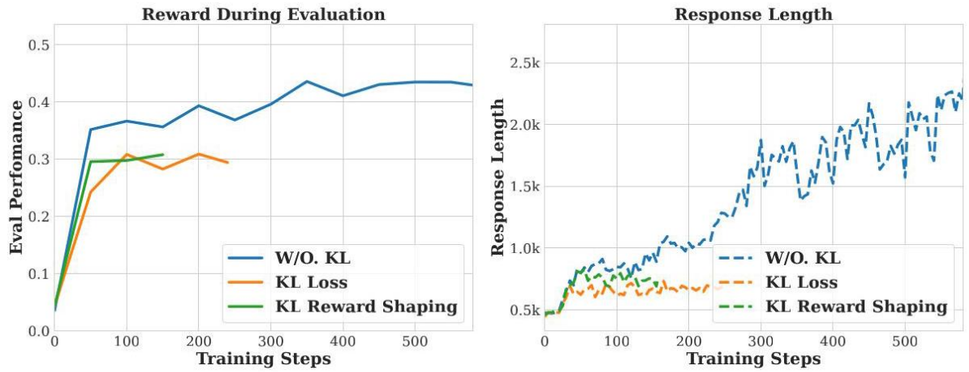

Figure 9: Comparisons to applying KL-related regularizations. Notably, training without KL constraints demonstrates superior average benchmark performance and length scaling property, compared to models trained with KL Loss and KL Penalty. Performance is evaluated on MATH500, AIME2024, and GPQA DIAMOND benchmarks using pass@1 metric.

GAE Analysis. We compare different GAE λ combinations. From the experimental results, we find that GAE $\lambda = 1.0$ performs best in terms of training stability and final performance. Specifically, in the training reward, the GAE $\lambda = 1.0$ curve rises quickly in the early stage and remains stable, finally converging to about 0.8 ; while the GAE $\lambda = 0.95$ curve rises slowly and fluctuates. In the Response Length, the GAE Lambda=1.0 curve maintains a reasonable level during the training process; while the GAE $\lambda = 0.95$ curve shows an unstable trend, leading to PPO learning instability. These results indicate that GAE $\lambda = 1.0$ can better balance the training stability and generation quality. Moreover, discount factor (γ) set to 1.0 also has a significant impact on the scale-up RL training. Less than 1.0 will result in

[2] Notion Comprehensive quantitative and qualitative analyses.

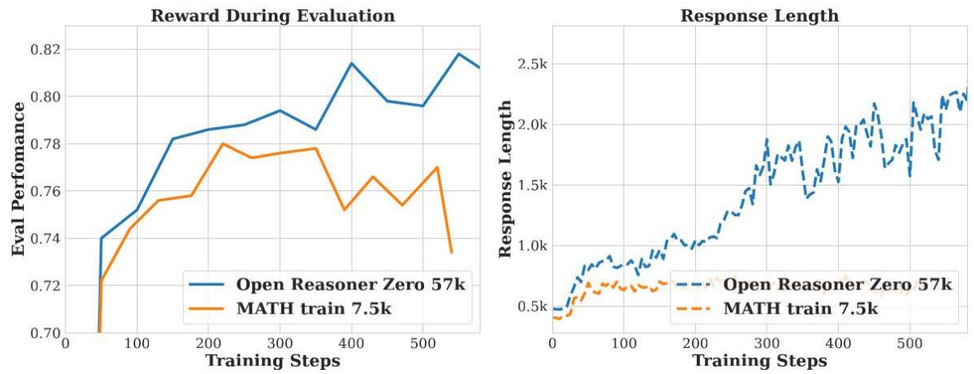

图 10：数据规模消融研究。从 math train 7.5k 到 Open-Reasoner-Zero 57k 的训练数据，我们观察到训练奖励和响应长度在训练集和评估集上都有一致的增加，这表明数据规模在训练性能中起着关键作用。性能在 MATH500 基准上使用 pass@1 指标进行评估。

KL 约束分析。我们评估了不同组合的 KL 损失和 KL 惩罚对 Open-Reasoner-Zero 7B 模型的影响，分析它们对评估指标和训练集响应长度的影响。这一分析尤为重要，因为奖励塑形可能会对训练奖励产生额外的影响。我们的实验结果表明，移除 KL 损失和 KL 惩罚可以达到最佳的训练稳定性和最终性能。KL 损失和 KL 惩罚机制不仅会减慢训练过程，还会消耗本可以更好地用于奖励优化的计算资源。此外，消除这些组件可以减少超参数调整的负担和实现的复杂性，这对于有效扩展 RL 训练至关重要。

数据规模。我们比较了不同规模的数据进行训练，范围从 7.5k 到 30k 样本。如图 10 所示，更大的数据规模在训练奖励和响应长度上始终表现出更好的性能，无论是在训练集还是评估集上。这一结果表明，数据规模在训练性能中起着关键作用，增加训练数据规模可以有效提高模型的推理能力。关于数据量、质量和多样性的更全面的消融研究可在附录中找到。

3.4. Evaluation Results

在本节中，我们列出了我们的主要实验结果。在我们的 32B 实验中，Open-Reasoner-Zero 在训练效率和模型性能方面都表现出显著的改进，如图 1 11 所示。该模型在所有基准测试中都实现了更优的响应长度和准确性，特别是在 GPQA DIAMOND 基准测试中显著优于 DeepSeek-R1-Zero-Qwen2.5-32B，而仅需 $\frac{1}{30}$ 的训练步骤。

如图 5 所示，Open-Reasoner-Zero 7B 模型在不同基准测试中表现出有趣的动态学习特性。准确性通常在训练过程中稳步提高，而响应长度则表现出更显著的增长模式。值得注意的是，在评估过程中我们观察到一个有趣的突现现象，即在某个时间点，奖励和响应长度突然出现阶跃函数式的增长，我们称之为“阶跃时刻”。这表明模型在训练过程中逐渐掌握了更详细和全面的推理能力。这种模式在 GPQA DIAMOND 和 AIME2024 中尤为明显，响应长度在后期训练步骤中显著增加。

penalty for long-term reward, leading to a decrease in response length decrease and struggling to improve the final performance.

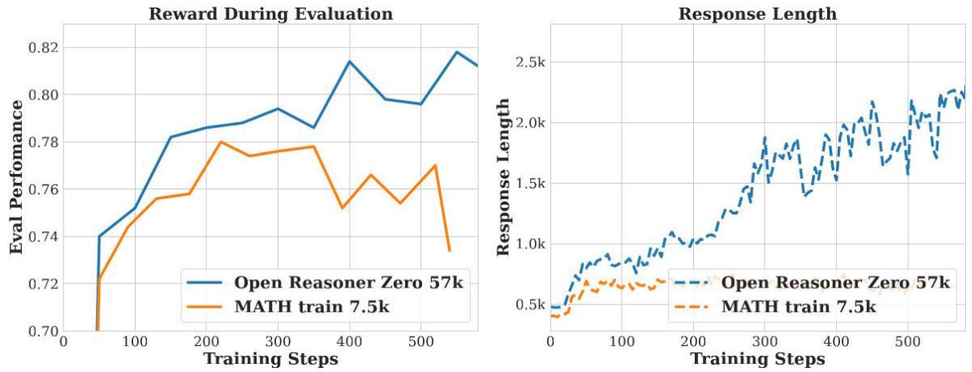

Figure 10: Data scale ablation study. Training data from math train 7.5k to Open-Reasoner-Zero 57k, we observe a consistent increase in both training reward and response length for training and evaluation set, indicating that data scale plays a crucial role in training performance. Performance is evaluated on MATH500 benchmark using pass@1 metric.

KL Constrains Analysis. We evaluate different combinations of KL Loss and KL Penalty for the Open-Reasoner-Zero 7B model, analyzing their impact on evaluation metrics and response length of the training set. This analysis is particularly important since reward shaping can introduce additional effects on training rewards. Our experimental results demonstrate that removing both KL Loss and KL Penalty yields optimal training stability and final performance. Both KL Loss and KL Penalty mechanisms not only slow down the training process but also consume computational resources that could be better utilized for reward optimization. Furthermore, eliminating these components reduces hyperparameter tuning burden and implementation complexity, which is crucial for scaling up RL training effectively.

Data Scale. We compare different data scales for training, ranging from 7.5k to 30k samples. As shown in Figure 10, larger data scales consistently lead to better performance in both training reward and response length for both training and evaluation sets. This result suggests that data scale plays a crucial role in training performance, and increasing the training data scale can effectively improve the model's reasoning capabilities. More comprehensive ablation studies including data quantity, quality, and diversity are available in the appendix.

3.4. Evaluation Results

In this section, we list our main experimental results. In our 32B experiments, Open-Reasoner-Zero demonstrates significant improvements in both training efficiency and model performance, as shown in Figure 1 11. The model achieves superior response length and accuracy across all benchmarks, notably outperforming DeepSeek-R1-Zero-Qwen2.5-32B on the GPQA DIAMOND benchmark while requiring only 1/30 of the training steps.

As shown in Figure 5, Open-Reasoner-Zero 7B model demonstrates interesting learning dynamics across different benchmarks. The accuracy generally shows a steady increase during training, while the response length exhibits more dramatic growth patterns. Notably, we observe an interesting emergent

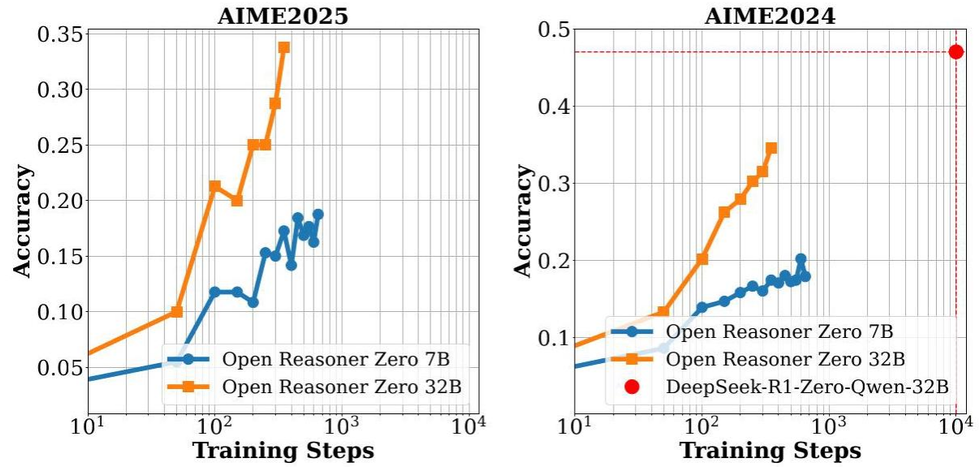

图 11: Open-Reasoner-Zero- $\{7\,\mathrm{B}, 32\,\mathrm{B}\}$ 的评估性能。我们报告了每个问题在基准数据集上的平均准确率，每个问题有 16 个回答。我们正在继续扩大这些强化学习设置，直到本预印本发布，因为目前还没有出现饱和的迹象。

Model	MMLU	MMLU_PRO
Qwen2.5-32B Base	83.3	55.1
Qwen2.5-32B Instruct	83.2	69.2
Open-Reasoner-Zero-32B	85.1	72.7

表 2：Open-Reasoner-Zero 模型在 MMLU 和 MMLU_PRO 基准上的泛化性能。通过仅在推理导向任务上扩大 RL 训练，Open-Reasoner-Zero 在这两个基准上都取得了优异的性能，超过了 Qwen2.5 Instruct，而无需任何额外的指令调优。这证明了我们的训练管道在增强模型泛化能力方面的显著有效性。

我们随后展示了模型在知识和指令跟随基准 MMLU_PRO 和 IFEval 上的泛化能力。如表 2 所示，Open-Reasoner-Zero 32B 模型展示了强大的泛化能力，在 MMLU 和 MMLU_PRO 上显著优于 Qwen2.5 Instruct 32B，这仅通过在推理导向任务上扩大 RL 训练规模实现，而无需任何额外的指令调优。

4. Conclusion and Discussions

在本工作中，我们介绍了 Open-Reasoner-Zero (ORZ)，这是第一个大规模推理导向的强化学习训练的开源实现，重点关注可扩展性、简单性和易用性。通过广泛的实验，我们的最佳实践表明，使用 GAE $(\lambda = 1, \gamma = 1)$ 的纯 PPO 和简单的基于规则的奖励函数，无需任何 KL 正则化，就足以在响应长度和推理任务的基准性能上进行扩展，展现出令人惊讶的泛化能力，取得了与 DeepSeek-R1-Zero 管道相当的竞争力。我们提供了成功进行大规模强化学习训练所需的关键组件和设置的全面分析，以及关于扩展 PPO 的重要见解。通过发布我们的完整训练资源，我们旨在使更广泛的人群能够参与这一 AI 发展的关键时刻。我们认为，我们正处于这一新扩展趋势的早期阶段，我们非常兴奋能够与社区分享我们的发现和经验。

回顾过去的惨痛教训：从长远来看，唯一重要的是能够随着计算和数据的增加而有效扩展。这一基本见解继续指导我们的研究方向。未来，我们计划进一步探索以下方向，以持续扩展推理导向的强化学习：

- 数据缩放：我们将研究如何通过增加训练数据的数量、质量和多样性来有效扩大规模。通过开源我们自己的训练数据集，我们希望鼓励研究社区贡献和分享更多的训练数据。

phenomenon during evaluation where both the reward and response length exhibit sudden, step-function-like increases at a certain point, which we refer to as the "step moment". This suggests that the models learn to master more detailed and comprehensive reasoning capacities as training progresses. This pattern is particularly pronounced in GPQA DIAMOND and AIME2024, where response lengths increase substantially in the later training steps.

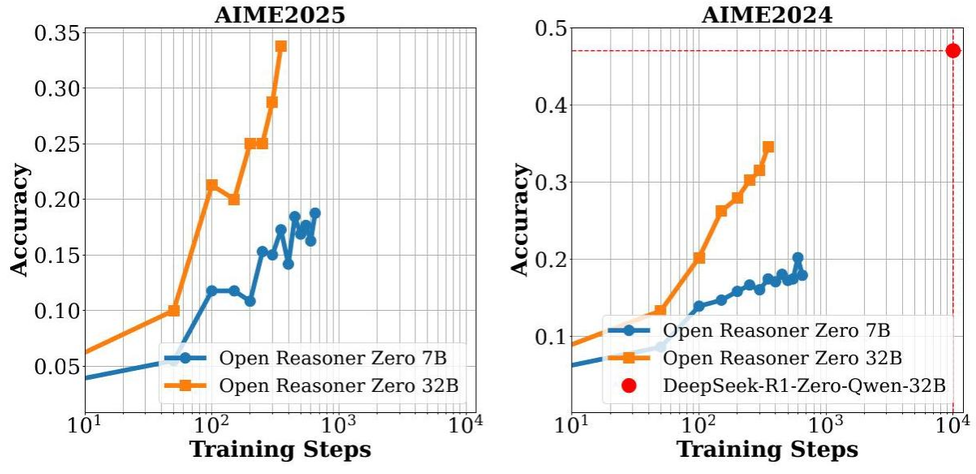

Figure 11: Evaluation performance of Open-Reasoner-Zero- $\{7\,\mathrm{B}, 32\,\mathrm{B}\}$. We report the average accuracy on the benchmark dataset for each question with 16 responses. We are continuing to scale up these RL settings until this preprint is released, as there is no sign of saturation yet.

Model	MMLU	MMLU_PRO
Qwen2.5-32B Base	83.3	55.1
Qwen2.5-32B Instruct	83.2	69.2
Open-Reasoner-Zero-32B	85.1	72.7

Table 2: Generalization performance of Open-Reasoner-Zero models on MMLU and MMLU_PRO benchmarks. Through solely scaling up RL training on reasoning-oriented tasks, Open-Reasoner-Zero achieves superior performance on both benchmarks, surpassing Qwen2.5 Instruct without any additional instruction tuning. This demonstrates the remarkable effectiveness of our training pipeline in enhancing model generalization capabilities.

We then present the generalization capabilities of our models on knowledge and instruction following benchmarks, MMLU_PRO and IFEval. As shown in Table 2, Open-Reasoner-Zero 32B models demonstrate strong generalization capabilities significantly outperforming Qwen2.5 Instruct 32B on MMLU, MMLU_PRO through pure scale-up RL training on reasoning-oriented tasks, without any additional instruction tuning.

4. Conclusion and Discussions

In this work, we present Open-Reasoner-Zero (ORZ), the first open-source implementation of large-scale reasoning-oriented RL training, focusing on scalability, simplicity, and accessibility. Through extensive experiments, our best practice demonstrates that vanilla PPO with GAE ($\lambda = 1, \gamma = 1$) and

- 模型扩展：我们将探讨如何扩展模型架构以提高推理能力。我们将研究多模态模型如何能够实现跨不同模态的更丰富的推理，以及更长的序列长度如何能够支持更复杂的多步骤推理。
- 测试时间扩展：我们将探讨如何扩展测试时间计算。我们将研究多轮交互如何增强上下文推理能力，价值模型如何评估推理轨迹，以及多代理场景如何导致更复杂的推理策略。
- 场景扩展：我们将探讨如何提升推理的复杂性以应对一般场景。我们的重点将放在将推理能力推广到日益多样的任务上，这些任务涵盖了创意写作、科学发现和社会互动领域。

5. Acknowledgements

这项工作得到了 StepFun 提供的计算资源和基础设施的支持。我们感谢 StepFun 和清华大学的同事们提供的宝贵反馈和贡献。

References

[1] OpenAI. Learning to reason with llms. https://openai.com/index/learning-to-r eason-with-llms/, 2025.

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

[3] Richard Sutton. The bitter lesson. Incomplete Ideas (blog), 13(1):38, 2019.

[4] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347, 2017.

[5] Hui Binyuan, Yang Jian, Cui Zeyu, Yang Jiaxi, Liu Dayiheng, Zhang Lei, Liu Tianyu, Zhang Jiajun, Yu Bowen, Lu Keming, et al. Qwen2. 5-coder technical report. arXiv preprint arXiv:2409.12186, 2024.

[6] Luo Michael, Tan Sijun, Wong Justin, Shi Xiaoxiang, Tang William, Roongta Manan, Cai Colin, Luo Jeffrey, Zhang Tianjun, Li Erran, Popa Raluca Ada, and Stoica Ion. Deepscaler: Surpassing o1-preview with a 1.5b model by scaling rl. https://pretty-radio-b75.n otion.site/DeepScaleR-Surpassing-01-Preview-with-a-1-5B-Model-by-S caling-RL-19681902c1468005bed8ca303013a4e2, 2025. Notion Blog.

[7] Lyu Chengqi, Gao Songyang, Gu Yuzhe, Zhang Wenwei, Gao Jianfei, Liu Kuikun, Wang Ziyi, Li Shuaibin, Zhao Qian, Huang Haian, Cao Weihan, Liu Jiangning, Liu Hongwei, Liu Junnan, Zhang Songyang, Lin Dahua, and Chen Kai. Exploring the limit of outcome reward for learning mathematical reasoning, 2025.

[8] Hu Jian, Wu Xibin, Zhu Zilin, Xianyu, Wang Weixun, Zhang Dehao, and Cao Yu. Open-rlhf: An easy-to-use, scalable and high-performance rlhf framework. arXiv preprint arXiv:2405.11143, 2024.

[9] 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. Numinamath. [https://huggin gface.co/AI-MO/NuminaMath-CoT](https://github.com/project-numina/ai mo-progress-prize/blob/main/report/numina_dataset.pdf),2024.

[10] 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.

straightforward rule-based reward function, without any KL regularization, is sufficient to scale up in both response length and benchmark performance on reasoning tasks with surprising generalization capabilities, achieving competitive results compared to DeepSeek-R1-Zero pipeline. We provide comprehensive analysis of the key components and settings required for successful large-scale RL training, along with critical insights into scaling up PPO. By releasing our complete training resources, we aim to enable broader participation in this pivotal moment of AI development. We believe we are at an early stage of this new scaling trend, and we are excited to share our findings and experiences with the community.

Recall a bitter lesson from the past: the only thing that matters in the long run is what scales up effectively with increased computation and data. This fundamental insight continues to guide our research direction. In the future, we plan to further explore the following directions for continuously scaling up reasoning-oriented RL:

- Data Scaling: We will investigate how to effectively scale up by increasing the quantity, quality and diversity of training data. By open sourcing our own training dataset, we hope to encourage the research community to contribute and share more training data.

- Model Scaling: We will explore how to scale up model architectures to improve reasoning abilities. We will investigate how multimodal models can enable richer reasoning across different modalities, and how extended sequence lengths can allow for more complex multi-step reasoning.

- Test Time Scaling: We will explore how to scale up test time computation. We will investigate how multi-turn interactions can enhance contextual reasoning abilities, how value model can assess reasoning trajectories, and how multi-agent scenarios can lead to more sophisticated reasoning strategies.

- Scenario Scaling: We will explore how to scale up the complexity of reasoning for general scenarios. Our focus will be on generalizing reasoning capabilities to increasingly diverse tasks spanning creative writing, scientific discovery, and social interaction domains.

5. Acknowledgements

This work was supported by computing resources and infrastructure provided by StepFun We are grateful for our colleagues from StepFun and Tsinghua University for their valuable feedback and contributions.

References

[1] OpenAI. Learning to reason with llms. https://openai.com/index/learning-to-r eason-with-llms/, 2025.

[2] 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.

[3] Richard Sutton. The bitter lesson. Incomplete Ideas (blog), 13(1):38, 2019.

[4] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347, 2017.

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

[12] 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.[13] Kimi 团队. Kimi k1.5: 使用大型语言模型扩展强化学习. 2025.

[14] David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, 和 Samuel R Bowman. GPQ.: 一个研究生级别的 Google 证明 QA 基准. arXiv 预印本 arXiv:2311.12022, 2023.

[15] Mislav Balunović, Jasper Dekoninck, Martin Vechev Ivo Petrov, Nikola Jovanović. MathArena: 在未受污染的数学竞赛中评估大型语言模型, 2025 年 2 月.

[16] Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, 和 Jacob Steinhardt. 使用数学数据集测量数学问题解决能力. Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, 2021.

[17] Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, 和 Ion Stoica. LiveCodeBench: 大型语言模型代码评估的全面且无污染方法. arXiv 预印本 arXiv:2403.07974, 2024.

[18] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, 和 Jacob Steinhardt. 测量大规模多任务语言理解. In International Conference on Learning Representations.

[19] Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, 和 Wenhu Chen. MMLU-Pro: 一个更强大和更具挑战性的多任务语言理解基准. In Advances in Neural Information Processing Systems, NeurIPS 2024, 2024.

[20] Edward Yeo, Yuxuan Tong, Morry Niu, Graham Neubig, 和 Xiang Yue. 揭示大型语言模型中的长链推理. arXiv 预印本 arXiv:2502.03373, 2025.

[21] Loubna Ben Allal, Lewis Tunstall, Anton Lozhkov, Elie Bakouch, Guilherme Penedo, 和 Gabriel Martín Blázquez Hynek Kydlicek. Open R1: 在未受污染的数学竞赛中评估大型语言模型, 2025 年 2 月.

[22] 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.

[23] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, 等. LLaMA 3 模型群. arXiv 预印本 arXiv:2407.21783, 2024.

[24] Yingqian Min, Zhipeng Chen, Jinhao Jiang, Jie Chen, Jia Deng, Yiwen Hu, Yiru Tang, Jiapeng Wang, Xiaoxue Cheng, Huatong Song, Wayne Xin Zhao, Zheng Liu, Zhongyuan Wang, 和 Ji-Rong Wen. 模仿、探索和自我改进: 慢思考推理系统的复现报告. arXiv 预印本 arXiv:2412.09413, 2024.

[5] Binyuan Hui, Jian Yang, Zeyu Cui, Jiaxi Yang, Dayiheng Liu, Lei Zhang, Tianyu Liu, Jiajun Zhang, Bowen Yu, Keming Lu, et al. Qwen2. 5-coder technical report. arXiv preprint arXiv:2409.12186, 2024.

[6] Michael Luo, Sijun Tan, Justin Wong, Xiaoxiang Shi, William Tang, Manan Roongta, Colin Cai, Jeffrey Luo, Tianjun Zhang, Erran Li, Raluca Ada Popa, and Ion Stoica. Deepscaler: Surpassing o1-preview with a 1.5b model by scaling rl. https://pretty-radio-b75.n otion.site/DeepScaleR-Surpassing-01-Preview-with-a-1-5B-Model-by-S caling-RL-19681902c1468005bed8ca303013a4e2, 2025. Notion Blog.

[7] Chengqi Lyu, Songyang Gao, Yuzhe Gu, Wenwei Zhang, Jianfei Gao, Kuikun Liu, Ziyi Wang, Shuaibin Li, Qian Zhao, Haian Huang, Weihan Cao, Jiangning Liu, Hongwei Liu, Junnan Liu, Songyang Zhang, Dahua Lin, and Kai Chen. Exploring the limit of outcome reward for learning mathematical reasoning, 2025.

[8] Jian Hu, Xibin Wu, Zilin Zhu, Xianyu, Weixun Wang, Dehao Zhang, and Yu Cao. Open-rlhf: An easy-to-use, scalable and high-performance rlhf framework. arXiv preprint arXiv:2405.11143, 2024.

[9] 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. Numinamath. [https://huggin gface.co/AI-MO/NuminaMath-CoT](https://github.com/project-numina/ai mo-progress-prize/blob/main/report/numina_dataset.pdf),2024.

[10] 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.

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

[12] 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.

[13] Kimi Team. Kimi k1.5: Scaling reinforcement learning with llms. 2025.

[14] David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. Gpq.: A graduate-level google-proof q&a benchmark. arXiv preprint arXiv:2311.12022, 2023.

[15] Mislav Balunović, Jasper Dekoninck, and Martin Vechev Ivo Petrov, Nikola Jovanović. Matharena: Evaluating llms on uncontaminated math competitions, February 2025.

[16] Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, 2021.

[17] Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. Livecodebench: Holistic and contamination free evaluation of large language models for code. arXiv preprint arXiv:2403.07974, 2024.

[18] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In International Conference on Learning Representations.

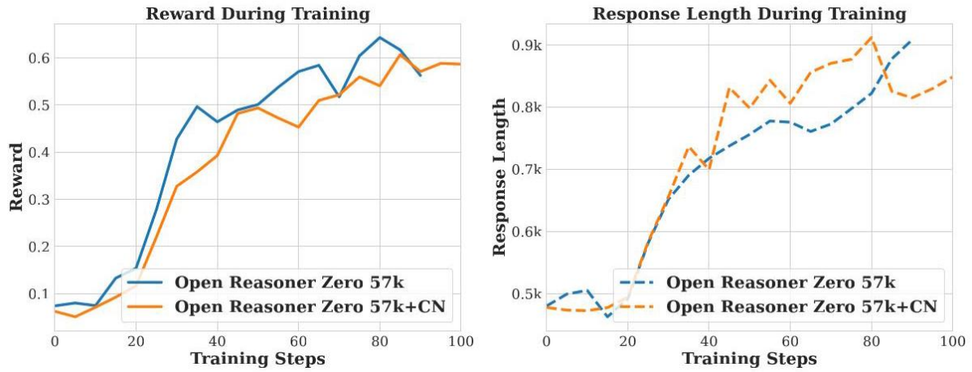

图 12: 数据整理消融研究。CN 代表中文数据，EN 代表英文数据。我们的结果表明，仅使用英文数据集可以带来更稳定的训练过程和最终模型性能。

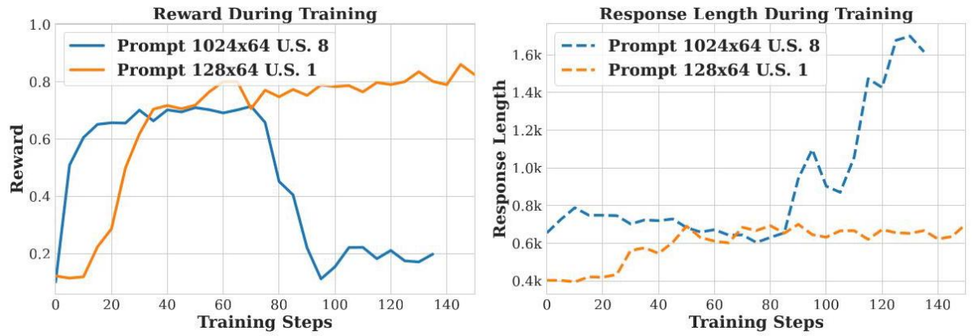

图 13：不同 Prompt、Rollout、Batch Size 组合的比较。U.S. 代表每次生成步骤中的模型参数更新步数。在策略更新在每次样本收集中的表现优于离策略更新，无论是在训练奖励还是响应长度上。

A. More Ablation Studies

在本节中，我们展示了在探索扩大强化学习（RL）训练规模过程中进行的额外消融研究。值得注意的是，我们的消融实验是在努力扩大 RL 训练规模时进行的，其中一些实验采用了不同的基本训练策略，以探索训练过程的各个方面。

数据整理的更多消融研究。基于我们对数据质量问题的分析，我们进行了全面的消融研究，以评估不同的数据整理策略如何影响模型训练的稳定性和性能。受 OpenR1 研究 [21] 的启发，该研究发现 SFT 在中文子集上的性能下降是由于问题模式较为简单，我们尝试了两种数据整理方法：仅使用英文数据与同时使用英文和中文数据。我们的结果表明，仅使用英文数据集可以带来更好的训练稳定性和最终模型性能。尽管我们的大多数实验使用了包括中文内容在内的完整数据集，但我们仍公开发布了最终的 Open-Reasoner-Zero 57k 数据集，因为它在各种任务中提供了更广泛的应用性。

[19] Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. In Advances in Neural Information Processing Systems, NeurIPS 2024, 2024.

[20] Edward Yeo, Yuxuan Tong, Morry Niu, Graham Neubig, and Xiang Yue. Demystifying long chain-of-thought reasoning in llms. arXiv preprint arXiv:2502.03373, 2025.

[21] Loubna Ben Allal, Lewis Tunstall, Anton Lozhkov, Elie Bakouch, Guilherme Penedo, and Gabriel Martín Blázquez Hynek Kydlicek. Open r1: Evaluating llms on uncontaminated math competitions, February 2025.

[22] 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, and Zhenru Zhang. Qwen2.5-math technical report: Toward mathematical expert model via self-improvement. arXiv preprint arXiv:2409.12122, 2024.

[23] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. arXiv preprint arXiv:2407.21783, 2024.

[24] Yingqian Min, Zhipeng Chen, Jinhao Jiang, Jie Chen, Jia Deng, Yiwen Hu, Yiru Tang, Jiapeng Wang, Xiaoxue Cheng, Huatong Song, Wayne Xin Zhao, Zheng Liu, Zhongyuan Wang, and Ji-Rong Wen. Imitate, explore, and self-improve: A reproduction report on slow-thinking reasoning systems. arXiv preprint arXiv:2412.09413, 2024.

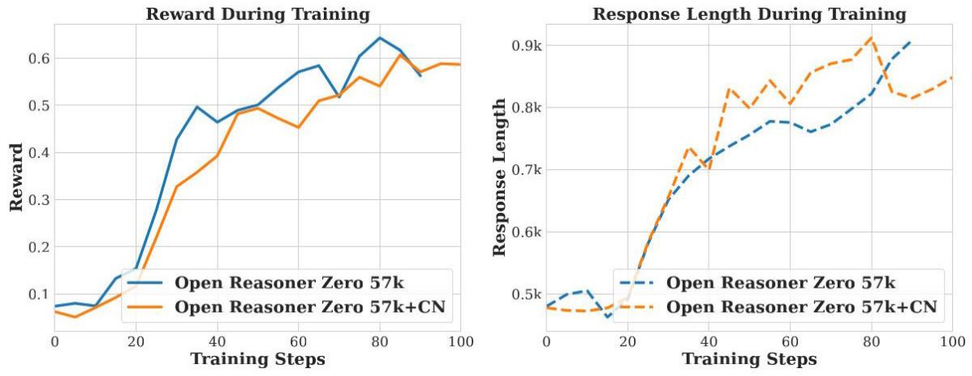

Figure 12: Data Curation Ablation Study. CN represents Chinese data and EN represents English data. Our results demonstrate that the English-only dataset yields superior training stability and final model performance.

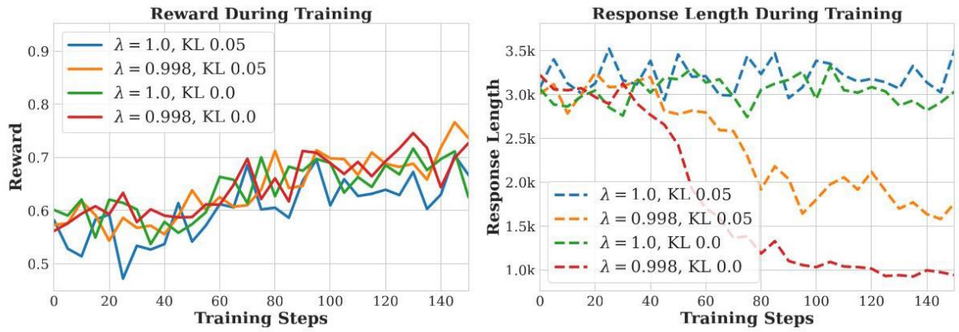

Figure 14: Comparison of different KL Loss, KL Penalty, and GAE λ values.

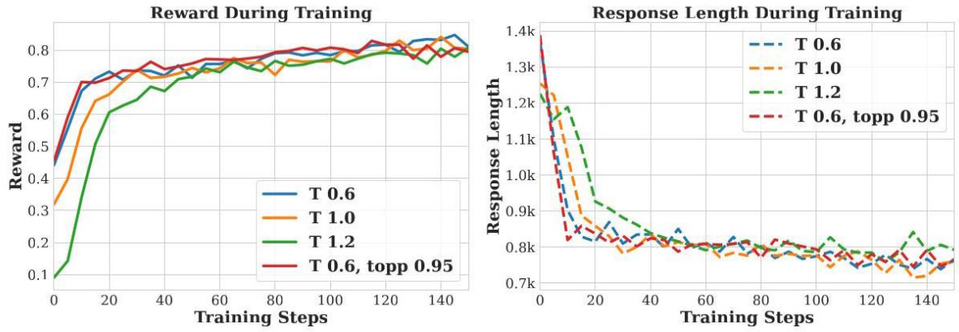

图 15：不同采样策略的比较。T 表示温度，topp 表示 top-p 采样。

采样策略。我们比较了不同的采样策略，包括温度 T = 0.6 , 1.0, 1.2 和 T=0.6 topp=0.95，与我们的主要设置相比，使用了不同的模型初始化、训练数据集和训练超参数。具体来说，这里我们使用 Qwen2.5-Math-7B[22] 作为初始化，并使用 MATH 训练集作为训练数据。至于训练超参数，这里我们采用了 1024 个独特的提示和每个提示在每次生成时的 8 个响应。我们将经验处理成最多 8 个 mini-batch，用于策略和批评者训练。从实验结果来看，我们发现大多数基本采样策略相比改变温度或 topp 表现更好。考虑到训练配方的可扩展性，我们最终选择了最基础的采样策略，即 T 和 topp 均等于 1.0。

更多关于 KL 损失、KL 惩罚和 GAE λ 的分析。我们分析了 LLaMA3.1-Instruct-SFT 模型 [23] 的 GAE Lambda 和 KL 损失。该模型在 STILL-2 数据 [24] 上进行训练。从实验结果来看，我们发现 GAE $\lambda = 1.0$ 和不使用 KL 损失的组合在训练稳定性和最终性能方面表现最佳。如图所示，这种配置在训练奖励和响应长度方面表现出最稳定的性能。此外，我们早期的实验还发现，引入 KL 惩罚（类似于 RLHF 中的奖励整形）显著影响了模型的推理能力。基于这些发现和可扩展性的考虑，我们最终选择了不使用 KL 约束的训练策略。

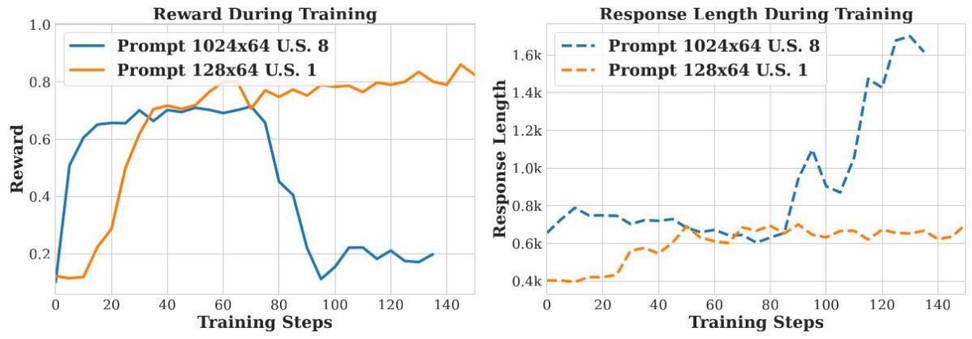

Figure 13: Comparison of different Prompt, Rollout, Batch Size combinations. U.S. represents Update steps of model parameters in each generation steps. On policy update for each sample collection performance better than off policy update on both training reward and response length.

A. More Ablation Studies

In this section, we present additional ablation studies conducted during our exploration of scaling up RL training. Notably, our ablation experiments were conducted during our efforts to scale up RL training, with some experiment employing different basic training strategies to explore various aspects of the training process.

More Ablations over Data Curation. Based on our analysis of data quality issues, we conduct comprehensive ablation studies to evaluate how different data curation strategies affect model training stability and performance. Motivated by OpenR1's finding [21] that SFT performance degradation on Chinese subsets was due to simpler question patterns, we experiment with two data curation approaches: using English-only data versus using both English and Chinese data. Our results demonstrate that the English-only dataset yields superior training stability and final model performance. While most of our experiments utilize the full dataset including Chinese content, we make the final Open-Reasoner-Zero 57k dataset publicly available as it provides broader applicability across diverse tasks.

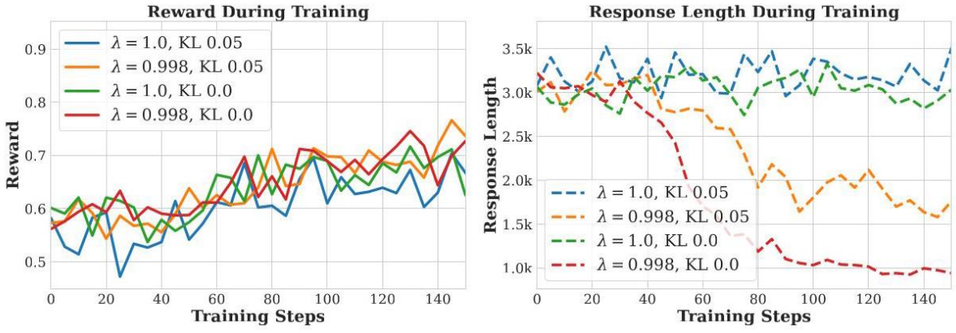

Figure 14: Comparison of different KL Loss, KL Penalty, and GAE λ values.

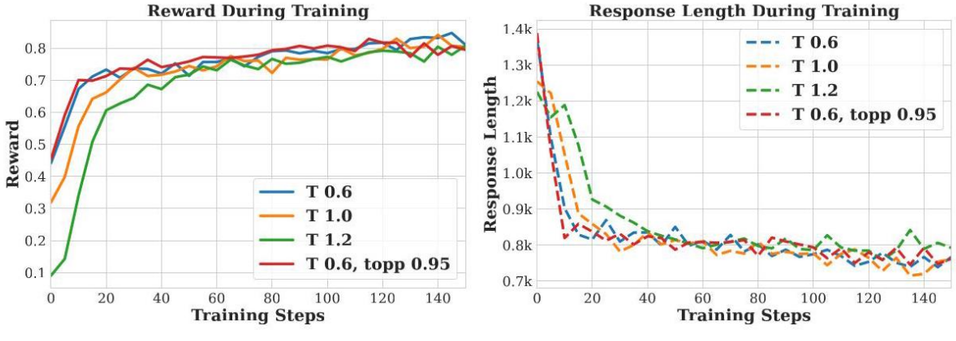

Figure 15: Comparison of different sampling strategies. T represents temperature and topp represents top-p sampling.

Sampling Strategy. We compare different sampling strategies, including temperature T = 0.6 , 1.0, 1.2 and T=0.6 topp=0.95, with different model initialization, training dataset and training hyperparameters compared to our main settings. Specifically, we use Qwen2.5-Math-7B[22] as initialization here and use MATH train set as training data. As for the training hyperparameters, here we adopt 1024 unique prompts and 8 responses for each prompt in each generation. We process the experiences into at most 8 mini-batches for both policy and critic training. From the experimental results, we find that most basic sampling strategy works well compared to changing temperature or topp. Considering the scalability of training recipe, we finally opt for the most basic sampling strategy that T and topp both equal to 1.0 .

More Ablations over KL Loss &KL Penalty &GAE λ Analysis. We analyze the GAE Lambda and KL Loss for the LLaMA3.1-Instruct-SFT model [23]. This model is trained on STILL-2 data [24]. From the experimental results, we find that the combination of GAE $\lambda = 1.0$ and no KL Loss performs best in terms of training stability and final performance. As shown in the figure, this configuration shows the most stable performance in both training reward and response length. Additionally, our early experiments also found that introducing KL Penalty (similar to reward shaping in RLHF) significantly affected the reasoning ability of the model. Based on these findings and for scalability, we finally chose the training strategy of not using KL constrains.