

ARE COMPLICATED LOSS FUNCTIONS NECESSARY FOR TEACHING LLMs TO REASON?

Anonymous authors

Paper under double-blind review

ABSTRACT

Recent advances in large language models (LLMs) highlight the importance of post-training techniques for improving reasoning and mathematical ability. Group Relative Policy Optimization (GRPO) has shown promise in this domain by combining group-relative advantage estimation, PPO-style clipping, and KL regularization. However, its complexity raises the question of whether all components are necessary for fostering reasoning behaviors. We conduct a systematic analysis of GRPO and identify two key findings: (1) incorporating negative feedback is essential—training solely on actions above a baseline limits learning; and (2) PPO-style constraints, such as policy ratio clipping, are not required to improve mathematical reasoning or performance. Building on these insights, we propose REINFORCE with Group Relative Advantage (RGRA), a simplified variant that retains group-relative advantage estimation but removes PPO-style clipping and policy ratio terms. Experiments across standard mathematical benchmarks indicate that RGRA has the potential to achieve stronger performance than GRPO. Our results suggest that simpler REINFORCE-based approaches can effectively enhance reasoning in LLMs, offering a more transparent and efficient alternative to GRPO.

1 INTRODUCTION

Recent advancements in artificial intelligence have been largely driven by large language models (LLMs) (OpenAI et al., 2024; Team et al., 2025; Jiang et al., 2024), which demonstrate remarkable capabilities across a wide range of tasks, from text translation (Becker et al., 2024) to complex mathematical problem-solving (Wang et al., 2025). A key factor in their progress has been improvements in the post-training phase, which aligns model outputs more closely with human preferences and enhances task-specific performance.

To address this challenge, researchers introduced approaches such as Reinforcement Learning from Human Feedback (RLHF), which trains models using reward signals designed to approximate human preferences. Early RLHF methods, like Proximal Policy Optimization (PPO) (Schulman et al., 2017b), relied on two separate models: a policy model (the LLM to be aligned), a reward model (Ouyang et al., 2022a), and a value model which adds computational and implementation complexity.

Recently, Group Relative Policy Optimization (GRPO) (Shao et al., 2024a) has emerged as a promising alternative. Introduced in Shao et al. (2024a), GRPO eliminates the need for a value model by estimating advantages as normalized rewards across a sampled group of completions for each prompt. This innovation simplifies the training pipeline while achieving strong performance on mathematical reasoning benchmarks. Building on this, the DEEPSEEK-R1 model (DeepSeek-AI et al., 2025) demonstrated that combining GRPO with simple rule-based rewards for format and correctness can elicit extended reasoning traces and complex reasoning behaviors.

Despite its empirical success, GRPO’s loss function combines several components — group-relative advantage estimation, PPO-style clipping, and KL regularization — resulting in a level of complexity that may not be strictly necessary for effective learning. This observation is underscored by recent GRPO variants targeting scalability, stability, and efficiency: Prefix Grouper optimizes shared-prefix encoding for faster training (Liu et al., 2025b), CPPO prunes low-advantage completions to reduce sampling cost (Lin et al., 2025b), DAPO introduces an increased upper clipping range to increase

the exploration of the policy Yu et al. (2025), S-GRPO facilitates early exit, allowing the model to focus only on the essential reasoning steps (Dai et al., 2025), and GTPO addresses token-level conflicts and policy collapse by introducing trajectory-level protection mechanisms (Simoni et al., 2025). Together, these works highlight both the promise of GRPO-style optimization and the potential overengineering of its current formulation.

In this work, we take a complementary approach: rather than proposing yet another GRPO variant, we *systematically analyze and simplify* its loss function. By isolating and removing individual components, we aim to identify which elements are essential for enabling learning and which can be simplified or removed without substantially compromising performance. Furthermore, we evaluate whether simpler reinforcement learning methods — including variants of REINFORCE (Ahmadian et al., 2024) and Rejection Sampling Fine-Tuning (RAFT) (Liu et al., 2024) — can match or surpass the performance improvements achieved by GRPO-trained models on mathematical reasoning tasks. This analysis contributes both conceptual clarity and practical guidance for designing efficient and robust post-training strategies for reasoning-focused LLMs.

2 BACKGROUND

2.1 RELATED WORK

Reasoning in Large Language Models. Recent advancements in large language models (LLMs) have shown that, once models reach sufficient scale, they exhibit emergent behaviors, including the capacity for reasoning (Wei et al., 2022). Many studies have addressed how to improve reasoning. One approach is to provide explicit reasoning examples during inference or training. For example, Wei et al. (2023a) demonstrated that prompting models with structured reasoning instructions, such as chain-of-thought examples or simple cues like *Let’s think step by step*, can induce explicit multi-step reasoning traces. Other efforts focus on supervised fine-tuning with reasoning annotations. Early work by Rajani et al. (2019) improved reasoning in smaller models by training on human-written rationales, while later approaches such as STaR (Zelikman et al., 2022) employed self-generated rationales, fine-tuning models on their own successful reasoning trajectories. A complementary line of research addresses the decoding phase, leveraging search algorithms to encourage reasoning exploration. For example, Luo et al. (2024) employed tree-based search, such as Monte Carlo Tree Search, to dynamically explore reasoning paths during generation.

Reinforcement Learning in LLMs. Reinforcement learning (RL) has long been central in decision-making tasks (Mnih et al., 2016; 2015; Berner et al., 2019), and is now widely adopted for aligning and fine-tuning LLMs. RL from Human Feedback (RLHF) for LLMs was introduced in InstructGPT (Ouyang et al., 2022b) and subsequently refined by Anthropic (Bai et al., 2022). RLHF has become a cornerstone in training pipelines for models such as Claude 3 (Anthropic, 2024), Gemini (Anil et al., 2023), and GPT-4 (OpenAI, 2023). Typically, RLHF involves supervised fine-tuning, a reward model, and Proximal Policy Optimization (PPO) (Schulman et al., 2017a). PPO stabilizes training by constraining updates through a clipped surrogate objective, providing a practical alternative to Trust Region Policy Optimization (TRPO) (Schulman et al., 2015). Nonetheless, PPO remains sensitive to reward scaling and prone to instability (Wang et al., 2019; Garg et al., 2021; Moalla et al., 2024), which has motivated refinements such as TRGPPO (Wang et al., 2019), alphaPPO (Xu et al., 2023), and PPO-ALR (Jia et al., 2024). Recent analyses further question whether the standard RL challenges motivating PPO apply in the LLM setting: Ahmadian et al. (2024) argue that pre-trained LLMs represent strong policies whose variance properties differ substantially from typical RL agents, suggesting that simpler policy-gradient methods may suffice.

Advancements and Limitations in GRPO. DeepSeek introduced Group Relative Policy Optimization (GRPO) (Shao et al., 2024b; Guo et al., 2025) to eliminate the critic model by leveraging relative rewards across multiple responses. This technique achieves state-of-the-art performance on math benchmarks and demonstrates that reasoning can emerge as a by-product of reinforcement learning (Guo et al., 2025). Specifically, by training on problems with verifiable answers (e.g., mathematics) using simple correctness and format rewards, GRPO-based models autonomously learned to extend their reasoning length, effectively allocating more compute to reasoning — a phenomenon now referred to as inference-time scaling. As a result, GRPO has become a de facto standard for inducing reasoning in LLMs.

However, emerging studies identify several limitations. Bias effects can skew relative comparisons (He et al., 2025), gradient imbalance may undertrain rare yet informative tokens (Yang et al., 2025; Liu et al., 2025a), and, as in PPO, model performance can degrade or collapse (Dohare et al., 2023). To address these issues, multiple variants have been proposed, including efficiency-focused methods (e.g., CPPO (Lin et al., 2025a)), stability-oriented designs (e.g., S-GRPO (Dai et al., 2025)), and token-level conflict resolution (e.g., GTPO (Simoni et al., 2025)). Complementary analyses deepen our understanding of token-sharing conflicts across completions, as well as policy-collapse mechanisms, highlighting the limits of KL-based constraints and motivating entropy-based approaches (Cui et al., 2025).

Positioning of this Work. Building on these developments, we conduct a systematic analysis of GRPO with a focus on mathematical reasoning tasks. We identify which components of the algorithm are essential and which can be simplified without loss of performance. In particular, we investigate whether a REINFORCE-based approach with group-relative advantages — our proposed REINFORCE with Group Relative Advantage (RGRA) — can match or surpass GRPO, offering a more transparent and efficient alternative for reasoning-focused post-training.

2.2 PRELIMINARIES

Reinforcement Learning (RL) is a subfield of machine learning in which an agent learns by interacting with its environment in order to improve its performance over time. The goal is to find the policy π that maximizes the expected cumulative reward, which can be expressed as:

$$J(\theta) = \mathbb{E}_{\tau \sim \pi_\theta} [G(\tau)]$$

where $G(\tau)$ denotes the return of a trajectory τ sampled from policy π_θ .

REINFORCE and Policy Gradient Methods For large language models, the reinforcement learning algorithms most commonly used in the post-training phase belong to the class of policy gradient methods. These methods optimize the policy directly by adjusting its parameters through gradient ascent in the direction that maximizes the expected reward:

$$\theta \leftarrow \theta + \alpha \nabla_\theta J(\theta),$$

where α is the learning rate and $\nabla_\theta J(\theta)$ denotes the policy gradient. The policy gradient can be expressed as:

$$\nabla_\theta J(\theta) = \mathbb{E}_{\tau \sim \pi_\theta} [\nabla_\theta \log \pi_\theta(\tau) G(\tau)]$$

Here, $\pi_\theta(\tau)$ denotes the probability of generating the entire trajectory $\tau = (s_0, a_0, s_1, a_1, \dots, s_T, a_T)$ under the policy, and $G(\tau)$ represents the return, that is, the cumulative reward associated with the full trajectory. When $G(\tau)$ is estimated through Monte Carlo rollouts of complete episodes, the method corresponds to the REINFORCE algorithm, originally introduced in Williams (1992).

PPO A widely used algorithm in the post-training phase of large language models is Proximal Policy Optimization (PPO). PPO is an actor-critic method that stabilizes training and reduces variance in reinforcement learning by employing a clipped surrogate objective. In this setting, the actions a_t correspond to the output tokens o_t , while the state s_t is defined by the prompt q together with the previously generated tokens $o_{<t}$.

$$J_{\text{PPO}}(\theta) = \mathbb{E} [q \sim P(Q), o \sim \pi_{\theta_{\text{old}}}(O | q)]$$

$$\frac{1}{|o|} \sum_{t=1}^{|o|} \left\{ \min \left[\frac{\pi_\theta(o_t | q, o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_t | q, o_{i,<t})} A_t, \text{clip} \left(\frac{\pi_\theta(o_t | q, o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_t | q, o_{i,<t})}, 1 - \epsilon, 1 + \epsilon \right) A_t \right] \right\}$$

Here, π_θ and $\pi_{\theta_{\text{old}}}$ denote the current and previous policy models, respectively. The parameter ϵ is the clipping coefficient, which constrains the policy ratio to the interval $(1 - \epsilon, 1 + \epsilon)$. The advantage function A_t is typically estimated using Generalized Advantage Estimation (GAE) Schulman et al. (2018), which relies on the rewards r_t for each time step and a value function. The value function, usually parameterized by a model of comparable size to the policy model, acts as a baseline when computing the advantage.

GRPO Group Relative Policy Optimization improves RL for LLMs by observing that the value model can be omitted, and the baseline can instead be inferred directly from group statistics. In

particular, the advantage estimation is computed as:

$$\hat{A}_{i,t} = \frac{r_i - \text{mean}(r_1, \dots, r_G)}{\text{std}(r_1, \dots, r_G)}$$

where r_i denotes the reward assigned to output o_i within the group of G sampled outputs ($o_1, \dots, o_G$) generated from the same prompt. Moreover, the KL penalty with respect to a reference model, originally applied to the reward at each token Shao et al. (2024a), is instead incorporated directly into the loss function. This results in the following loss:

$$J_{\text{GRPO}}(\theta) = \mathbb{E} [q \sim P(Q), \{o_i\}_{i=1}^G \sim \pi_{\theta_{\text{old}}}(O | q)] \quad (1)$$

$$\frac{1}{G} \sum_{i=1}^G \frac{1}{|o_i|} \sum_{t=1}^{|o_i|} \left\{ \min \left[r_{i,t} \hat{A}_{i,t}, \text{clip}(r_{i,t}, 1 - \epsilon, 1 + \epsilon) \hat{A}_{i,t} \right] - \beta \text{D}_{\text{KL}} [\pi_{\theta} \| \pi_{\text{ref}}] \right\}$$

where

$$r_{i,t} = \frac{\pi_{\theta}(o_{i,t} | q, o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t} | q, o_{i,<t})}.$$

and β is the parameter controlling the KL regularization.

RAFT RAFT aims to provide an alternative to traditional RL methods for LLMs. The proposed algorithm operates by sampling multiple responses for each prompt in a batch, ranking these generated completions using the rewards, and then selecting the highest-ranked response for each prompt. These top responses are then used to construct a new dataset, which serves as the training data for supervised fine-tuning using cross-entropy loss.

3 EXPERIMENTS

3.1 EXPERIMENTAL SETUP

Training Datasets We constructed our training set using problems drawn from GSM8K (Cobbe et al., 2021), a widely used benchmark for grade-school mathematics reasoning. From the training split of the dataset we randomly sampled 1,800 instances. This dataset was selected as it has been explicitly decontaminated from the training corpora of the models employed in our study (Qwen et al., 2025), ensuring unbiased evaluation. It serves as the basis for a quantitative assessment of both performance improvements and training stability.

Benchmarks To comprehensively assess the emerging capabilities of the different models in reasoning tasks, we select nine different benchmarks that reflect a diverse level of complexity. We consider five Math-English benchmarks: the testing split of **GSM8K** Cobbe et al. (2021), **MATH** (Hendrycks et al., 2021b), **Gaokao2023-Math-En** Liao et al. (2024), **OlympiadBench** (He et al., 2024), **AMC23** Yang et al. (2024). Then, we consider two Chinese Math benchmarks **CMATH** Wei et al. (2023b) and the **CN-Middle-School** Yang et al. (2024). Finally, we consider two STEM benchmarks: **MMLU-STEM** (English) Hendrycks et al. (2021a) and **Gaokao2024** (Chinese) Zhong et al. (2023).

Models We evaluate the proposed training schemes on two instruction-tuned variants of the Qwen2.5 family (Qwen et al., 2025), specifically the 0.5B and 1.5B parameter models and the instruction-tuned 1B parameters model of the Llama3.2 family Grattafiori et al. (2024). Those models are trained on the GSM8K benchmark, enabling a comparative analysis across small scales.

Evaluation To evaluate our model, we calculate the accuracy on a variety of standard benchmark datasets, including English and Chinese. We evaluate the benchmark accuracy of the different models, to understand the capabilities of the models to generalize over different tasks. We also consider training metrics such as average response length, which allows to monitor the ability to learn reasoning, and the average reward obtained by the models during training.

Experimental Details To fine-tune the models, we employ LoRA with a rank of 128, effectively reducing the number of trainable parameters to approximately 10% of the original model size. In addition, we incorporate other efficiency techniques such as gradient accumulation and gradient checkpointing. For inference, we adopt VLLM as the underlying engine. For each prompt in the

dataset, we generate a group of 8 completions. The maximum number of tokens generated per completion is set to 512. For the reward system, we implemented two distinct reward signals. The first is a format reward, granting 0.1 points to outputs that follow the specified format. The second is a correctness reward, awarding 1 point for answers that correctly solve the given task. A complete list of experimental parameters can be found in Appendix A.

3.2 EXPERIMENTS

We present a series of experiments designed to simplify and decompose the GRPO loss function. Our motivation stems from the observation that, although GRPO has proven effective in improving model performance on mathematical tasks, its structure, which combines group-relative advantage estimation, PPO-style clipping, and KL regularization, introduces a level of complexity that may not be strictly necessary for effective learning.

By systematically isolating and removing individual components, we aim to identify which elements are essential for enabling learning and which can be simplified or omitted without significantly degrading performance.

To this end, we examine three distinct variants of GRPO, evaluating how each simplification affects mathematical performance and the emergence of reasoning abilities in LLMs. Specifically, we analyze:

- **Positive-only Advantages:** We examine the effect of training exclusively on actions that outperform the current baseline, focusing learning on high-reward behaviors and ignoring negative feedback.

$$\mathcal{J}_{\text{GRPO-pos}}(\theta) = \mathbb{E} \left[q \sim P(Q), \{o_i\}_{i=1}^G \sim \pi_{\theta_{\text{old}}}(O | q) \right] \\ \frac{1}{G} \sum_{i=1}^G \frac{1}{|o_i|} \sum_{t=1}^{|o_i|} \left\{ \min \left[r_{i,t} \tilde{A}_{i,t}, \text{clip} \left(r_{i,t}, 1 - \epsilon, 1 + \epsilon \right) \tilde{A}_{i,t} \right] - \beta \text{D}_{\text{KL}} [\pi_{\theta} \parallel \pi_{\text{ref}}] \right\}$$

where the modified advantage term is given by:

$$\tilde{A}_{i,t} = \begin{cases} \hat{A}_{i,t} & \text{if } \hat{A}_{i,t} > 0 \\ 0 & \text{otherwise} \end{cases}$$

- **RGRA - Removing PPO-style Constraints:** Inspired by Ahmadian et al. (2024), we investigate the necessity of PPO-style clipping. In this variant, we remove policy ratios and clipping, proposing a REINFORCE variant that preserves GRPO’s group-relative advantage estimation. We refer to this simplified approach as RGRA, characterized by the following gradient:

$$\nabla_{\theta} \mathcal{J}_{\text{RGRA}}(\theta) = \mathbb{E} \left[q \sim P(Q), \{o_i\}_{i=1}^G \sim \pi_{\theta}(O | q) \right] \quad (2)$$

$$\frac{1}{G} \sum_{i=1}^G \frac{1}{|o_i|} \sum_{t=1}^{|o_i|} \left\{ \nabla_{\theta} \log \pi_{\theta}(o_{i,t} | q, o_{i,<t}) \cdot \hat{A}_{i,t} - \beta \nabla_{\theta} \text{D}_{\text{KL}} [\pi_{\theta} \parallel \pi_{\text{ref}}] \right\}$$

- **REINFORCE with Direct Rewards:** In this variant, we start from RGRA, remove the group-relative advantage estimation, and train directly on the raw reward signal.

Following these investigations, we also evaluated the impact of adopting a simpler rejection sampling strategy that uses a simple cross-entropy, RAFT Dong et al. (2023). The results of the proposed fine-tuning techniques are further compared with those achieved by the fine-tuned version of the models considered.

4 RESULTS AND DISCUSSION

Figure 1 reports average reward and response length during training on GSM8K for Qwen2.5 0.5B, 1.5B and Llama3.2 1B models across different objectives. Training with positive-only advantages



Figure 1: Training metrics (10-step running average).

Math-English Benchmarks							
Model	Method	GSM8K	MATH	Gaokao2023 Math-En	Olympiad Bench	AMC23	Avg
Qwen2.5-0.5-it	-	41.5	22.6	21.0	6.2	7.5	19.8
Qwen2.5-1.5-it	-	61.1	38.9	35.1	11.4	17.5	32.8
Llama3.2-1.0-it	-	37.9	18.9	14.5	4.6	10.0	13.4
Qwen2.5-0.5-it	GRPO	50.9	30.3	30.4	8.9	7.5	25.6
Qwen2.5-1.5-it	GRPO	71.0	44.2	38.7	12.6	20.0	37.3
Llama3.2-1.0-it	GRPO	43.0	22.9	17.4	4.6	12.5	20.1
Qwen2.5-0.5-it	GRPO-pos	35.6	21.1	19.7	6.1	2.5	17.0
Qwen2.5-1.5-it	GRPO-pos	70.6	41.0	38.7	10.7	17.5	35.7
Llama3.2-1.0-it	GRPO-pos	41.3	21.8	18.2	5.0	12.5	19.8
Qwen2.5-0.5-it	RGRA	53.1	32.1	29.1	8.3	10.0	26.5
Qwen2.5-1.5-it	RGRA	72.7	46.7	42.6	12.0	17.5	38.3
Llama3.2-1.0-it	RGRA	43.3	21.4	19.0	5.0	12.5	20.2
Qwen2.5-0.5-it	RAFT	14.1	12.0	10.9	4.0	2.5	8.7
Qwen2.5-1.5-it	RAFT	67.0	40.0	36.6	11.3	25.0	36.0
Llama3.2-1.0-it	RAFT	41.8	21.0	17.4	4.7	10.0	19.0
Qwen2.5-0.5-it	REINFORCE	44.7	26.1	24.7	4.9	12.5	22.6
Qwen2.5-1.5-it	REINFORCE	63.6	37.6	31.9	8.7	12.5	30.9
Llama3.2-1.0-it	REINFORCE	41.1	22.0	17.7	4.0	10.0	19.0
Qwen2.5-0.5-it	ft	39.5	20.7	20.8	5.0	7.5	18.7
Qwen2.5-1.5-it	ft	63.8	33.2	28.1	10.7	2.5	27.7
Llama3.2-1.0-it	ft	33.8	17.7	13.0	4.7	5.0	14.8

Table 1: Performance of trained models on English Math benchmarks. All models are trained using gsm8k Dataset.

Chinese Math Benchmarks				
Model	Method	CMATH	CN-Middle-School	Avg
Qwen2.5-0.5-it	-	34.3	42.6	38.5
Qwen2.5-1.5-it	-	52.3	51.5	51.9
Llama3.2-1.0-it	-	29.5	24.8	27.2
Qwen2.5-0.5-it	GRPO	51.2	51.5	51.4
Qwen2.5-1.5-it	GRPO	75.0	56.4	65.7
Llama3.2-1.0-it	GRPO	33.5	26.7	30.1
Qwen2.5-0.5-it	GRPO-pos	46.3	36.6	41.4
Qwen2.5-1.5-it	GRPO-pos	71.2	59.4	65.3
Llama3.2-1.0-it	GRPO-pos	35.7	24.8	30.3
Qwen2.5-0.5-it	RGRA	54.8	55.4	55.1
Qwen2.5-1.5-it	RGRA	72.3	66.3	69.3
Llama3.2-1.0-it	RGRA	27.5	25.7	26.6
Qwen2.5-0.5-it	RAFT	35.2	32.7	34.0
Qwen2.5-1.5-it	RAFT	66.2	55.4	60.8
Llama3.2-1.0-it	RAFT	34.8	21.8	28.3
Qwen2.5-0.5-it	REINFORCE	42.7	43.6	43.2
Qwen2.5-1.5-it	REINFORCE	71.0	55.4	63.2
Llama3.2-1.0-it	REINFORCE	30.0	25.7	27.9
Qwen2.5-0.5-it	ft	29.2	42.6	35.9
Qwen2.5-1.5-it	ft	65.3	53.5	59.4
Llama3.2-1.0-it	ft	26.2	15.8	21.0

Table 2: Performance of trained models on Chinese Math benchmarks. All models are trained using gsm8k Dataset.

STEM Benchmarks				
Model	Method	English (MMLU-STEM)	Chinese (Gaokao2024)	Avg
Qwen2.5-0.5-it	-	40.6	24.4	32.5
Qwen2.5-1.5-it	-	59.2	31.5	45.4
Llama3.2-1.0-it	-	31.7	13.7	22.7
Qwen2.5-0.5-it	GRPO	41.3	21.2	31.3
Qwen2.5-1.5-it	GRPO	58.7	32.6	45.7
Llama3.2-1.0-it	GRPO	32.6	17.2	24.9
Qwen2.5-0.5-it	GRPO-pos	39.7	19.6	29.7
Qwen2.5-1.5-it	GRPO-pos	59.5	33.9	46.7
Llama3.2-1.0-it	GRPO-pos	32.4	12.8	22.6
Qwen2.5-0.5-it	RGRA	42.0	26.5	34.3
Qwen2.5-1.5-it	RGRA	60.1	41.2	50.7
Llama3.2-1.0-it	RGRA	33.5	11.4	22.5
Qwen2.5-0.5-it	RAFT	39.7	20.2	30.0
Qwen2.5-1.5-it	RAFT	58.2	33.9	46.1
Llama3.2-1.0-it	RAFT	31.6	14.0	22.8
Qwen2.5-0.5-it	REINFORCE	41.1	25.7	33.4
Qwen2.5-1.5-it	REINFORCE	57.9	31.1	44.5
Llama3.2-1.0-it	REINFORCE	32.8	11.1	22.0
Qwen2.5-0.5-it	ft	39.4	17.1	28.3
Qwen2.5-1.5-it	ft	55.5	24.4	40.0
Llama3.2-1.0-it	ft	31.7	12.9	22.3

Table 3: Performance of trained models on STEM benchmarks (English: MMLU, Chinese: Gaokao2024). All models are trained using gsm8k Dataset.

and RAFT exhibits severe instability, particularly in the 0.5B model, where both reward and response length collapse within the first 20 steps. This collapse manifests as degenerate outputs of minimal length, indicating a reward-hacking phenomenon where the model exploits the absence of negative feedback by converging toward trivial responses. Although the 1.5B and 1B models trained under these regimes avoid immediate collapse, they still demonstrate reward stagnation and gradual shortening of responses, suggesting that discarding negative feedback systematically biases models toward under-exploration and degraded reasoning. By contrast, GRPO and RGRA maintain stable training dynamics in both model sizes, both achieving comparable reward trajectories. These findings reinforce that advantage estimation is essential for stabilizing reinforcement learning in LLMs, while PPO-style clipping is not strictly required when initializing from strong policies. However, training with direct REINFORCE on raw rewards collapses even in the larger 1.5B model, underscoring the indispensable role of advantage estimation for stability.

Math-English Benchmarks The performance of trained models on English mathematical reasoning tasks is reported in Table 1. First, we can observe that models trained with RAFT or positive-only GRPO show worse performance compared to both GRPO and RGRA. Especially for the Qwen2.5 0.5B model. This poor result aligns with the observed training collapse and response truncation. Second, GRPO achieves consistent improvements over instruction-tuned baselines, particularly on GSM8K and MATH, but RGRA outperforms GRPO in most settings. Specifically, RGRA achieves the highest average performance across the Math-English benchmarks for the three models used, surpassing GRPO in 17 out of 27 individual comparisons. This further supports the claim that PPO-style constraints are not necessary for effective learning from strong initialization.

Chinese Math Benchmarks The results on CMATH and CN-Middle-School (Table 2) reveal a similar pattern. RAFT and positive-only GRPO exhibit weaker performance with respect to GRPO. In contrast, RGRA delivers the strongest results in the Qwen2.5 models, achieving average accuracies of 55.1 (0.5B) and 65.3 (1.5B), exceeding both standard GRPO and fine-tuned baselines.

STEM Benchmarks Performance in STEM-related evaluations (MMLU-STEM and Gaokao2024-STEM, see Table 3) further illustrates the limitations of training regimes that ignore negative feedback. RAFT and positive-only GRPO again fail to surpass fine-tuned baselines. GRPO

shows modest improvements and outperforms the other methods on Llama3.2. By contrast, RGRA achieves the best improvements on the Qwen2.5 models.

Emergence of Reasoning Behaviors Beyond raw benchmark scores, training methods also differ in their ability to induce reasoning behaviors. On the Countdown dataset, we observe that RAFT and positive-only GRPO models fail to generate explicit reasoning steps, instead outputting direct final answers (Figure 2(a)). Conversely, GRPO and RGRA models exhibit emergent reasoning, including explicit re-evaluation of intermediate answers (Figure 2(b)). This highlights that robust training regimes not only stabilize learning and improve benchmark scores but also foster the development of interpretable reasoning strategies in LLMs.

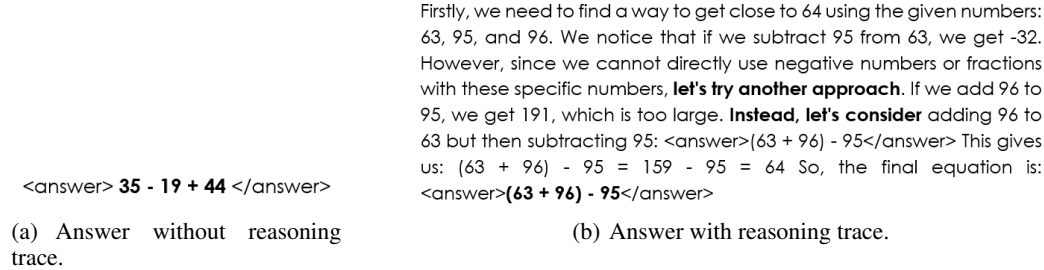


Figure 2: Comparison of answers with and without reasoning traces.

5 CONCLUSION

In this work, we investigated the GRPO loss function with the goal of disentangling its components and identifying which are essential for effective post-training of large language models. Through a series of controlled ablations and benchmark evaluations, we analyzed the impact of retaining or removing specific elements of the objective on learning stability, mathematical reasoning, and generalization across tasks and languages.

Our results demonstrate three key findings. First, negative feedback is indispensable: methods that ignore it—such as RAFT or positive-only GRPO—exhibit instability, collapse, and consistently degraded performance. Second, advantage estimation is crucial: removing it, as in REINFORCE with direct rewards, destabilizes learning even in larger models with strong initial policies. Third, PPO-style clipping is unnecessary: eliminating clipping and policy ratios does not harm stability; instead, it simplifies training and can lead to improved performance.

These insights motivated the introduction of RGRA, a simplified variant of GRPO that discards PPO-style constraints while preserving group-relative advantage estimation. Across training dynamics, multilingual mathematical benchmarks, STEM evaluations, and reasoning behavior analyses, RGRA not only achieves stable learning but also surpasses GRPO on 17 over 27 tasks, establishing it as a competitive reinforcement learning objective for reasoning tasks.

Overall, this study advances our understanding of how reinforcement learning objectives shape the post-training of large language models. By showing that GRPO can be simplified without sacrificing performance, we lay the groundwork for future research on reinforcement learning strategies that further enhance reasoning capabilities and generalization in LLMs.

Future works will consider exploring additional tasks outside the mathematical domain. An additional research work could address larger models, which was not possible here due to hardware constraints.

6 REPRODUCIBILITY STATEMENT

We provide detailed descriptions of each algorithm in Section 3 and Appendix A, including the techniques, fine-tuned hyperparameters, and infrastructures used in our experiments. The link to our code is https://anonymous.4open.science/r/math_llms-FE4E/README.md.

REFERENCES

- Arash Ahmadian, Chris Cremer, Matthias Gallé, Marzieh Fadaee, Julia Kreutzer, Olivier Pietquin, Ahmet Üstün, and Sara Hooker. Back to Basics: Revisiting REINFORCE Style Optimization for Learning from Human Feedback in LLMs, February 2024. URL <http://arxiv.org/abs/2402.14740>. arXiv:2402.14740 [cs].
- Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Slav Petrov, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy P. Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul Ronald Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anaïs White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, and et al. Gemini: A family of highly capable multimodal models. *CoRR*, abs/2312.11805, 2023. doi: 10.48550/ARXIV.2312.11805. URL <https://doi.org/10.48550/arXiv.2312.11805>.
- Anthropic. The claude 3 model family: Opus, sonnet, haiku — model card. Model card, Anthropic, March 2024. URL <https://assets.anthropic.com/m/61e7d27f8c8f5919/original/Claude-3-Model-Card.pdf>. Accessed: 2025-07-22.
- Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom B. Brown, Jack Clark, Sam McCandlish, Chris Olah, Benjamin Mann, and Jared Kaplan. Training a helpful and harmless assistant with reinforcement learning from human feedback. *CoRR*, abs/2204.05862, 2022. doi: 10.48550/ARXIV.2204.05862. URL <https://doi.org/10.48550/arXiv.2204.05862>.
- Jonas Becker, Jan Philip Wahle, Bela Gipp, and Terry Ruas. Text generation: A systematic literature review of tasks, evaluation, and challenges, 2024. URL <https://arxiv.org/abs/2405.15604>.
- Christopher Berner, Greg Brockman, Brooke Chan, Vicki Cheung, Przemyslaw Debiak, Christy Dennison, David Farhi, Quirin Fischer, Shariq Hashme, Christopher Hesse, Rafal Józefowicz, Scott Gray, Catherine Olsson, Jakub Pachocki, Michael Petrov, Henrique Pondé de Oliveira Pinto, Jonathan Raiman, Tim Salimans, Jeremy Schlatter, Jonas Schneider, Szymon Sidor, Ilya Sutskever, Jie Tang, Filip Wolski, and Susan Zhang. Dota 2 with large scale deep reinforcement learning. *CoRR*, abs/1912.06680, 2019. URL <http://arxiv.org/abs/1912.06680>.
- Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training Verifiers to Solve Math Word Problems, November 2021. URL <http://arxiv.org/abs/2110.14168>. arXiv:2110.14168 [cs].
- Ganqu Cui, Yuchen Zhang, Jiacheng Chen, Lifan Yuan, Zhi Wang, Yuxin Zuo, Haozhan Li, Yuchen Fan, Huayu Chen, Weize Chen, Zhiyuan Liu, Hao Peng, Lei Bai, Wanli Ouyang, Yu Cheng, Bowen Zhou, and Ning Ding. The entropy mechanism of reinforcement learning for reasoning language models, 2025. URL <https://arxiv.org/abs/2505.22617>.
- Muzhi Dai, Chenxu Yang, and Qingyi Si. S-grpo: Early exit via reinforcement learning in reasoning models. *arXiv preprint arXiv:2505.07686*, 2025.
- DeepSeek-AI, Daya Guo, et al. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning, January 2025. URL <http://arxiv.org/abs/2501.12948>. arXiv:2501.12948 [cs].
- Shibhansh Dohare, Qingfeng Lan, and A Rupam Mahmood. Overcoming policy collapse in deep reinforcement learning. In *Sixteenth European Workshop on Reinforcement Learning*, 2023.

- Hanze Dong, Wei Xiong, Deepanshu Goyal, Yihan Zhang, Winnie Chow, Rui Pan, Shizhe Diao, Jipeng Zhang, Kashun Shum, and Tong Zhang. RAFT: Reward rAnked FineTuning for Generative Foundation Model Alignment, December 2023. URL <http://arxiv.org/abs/2304.06767>. arXiv:2304.06767 [cs].
- Saurabh Garg, Joshua Zhanson, Emilio Parisotto, Adarsh Prasad, J. Zico Kolter, Zachary C. Lipton, Sivaraman Balakrishnan, Ruslan Salakhutdinov, and Pradeep Ravikumar. On proximal policy optimization’s heavy-tailed gradients. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 3610–3619. PMLR, 2021. URL <http://proceedings.mlr.press/v139/garg21b.html>.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhatta, Kushal Lakhotia, Lauren Rantala-Yearay, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Rapparthi, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collet, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vitor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuwei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu

- Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkan Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Rutu Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaoqian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. Llama 3 Herd of Models, November 2024. URL <http://arxiv.org/abs/2407.21783>. arXiv:2407.21783 [cs].
- 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.
- Andre He, Daniel Fried, and Sean Welleck. Rewarding the unlikely: Lifting GRPO beyond distribution sharpening. *CoRR*, abs/2506.02355, 2025. doi: 10.48550/ARXIV.2506.02355. URL <https://doi.org/10.48550/arXiv.2506.02355>.
- Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Leng Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, Jie Liu, Lei Qi, Zhiyuan Liu, and Maosong Sun. Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems, 2024. URL <https://arxiv.org/abs/2402.14008>.

- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring Massive Multitask Language Understanding, January 2021a. URL <http://arxiv.org/abs/2009.03300>. arXiv:2009.03300 [cs].
- 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. In Joaquin Vanschoren and Sai-Kit Yeung (eds.), *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*, 2021b.
- Lu Jia, Binglin Su, Du Xu, Yewei Wang, Jing Fang, and Jun Wang. Policy optimization algorithm with activation likelihood-ratio for multi-agent reinforcement learning. *Neural Process. Lett.*, 56(6):247, 2024. doi: 10.1007/S11063-024-11705-X. URL <https://doi.org/10.1007/s11063-024-11705-x>.
- Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, L  lio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Th  ophile Gerv  t, Thibaut Lavril, Thomas Wang, Timoth  e Lacroix, and William El Sayed. Mixtral of experts, 2024. URL <https://arxiv.org/abs/2401.04088>.
- Minpeng Liao, Wei Luo, Chengxi Li, Jing Wu, and Kai Fan. MARIO: MATH Reasoning with code Interpreter Output – A Reproducible Pipeline, February 2024. URL <http://arxiv.org/abs/2401.08190>. arXiv:2401.08190 [cs].
- Zhihang Lin, Mingbao Lin, Yuan Xie, and Rongrong Ji. Cppo: Accelerating the training of group relative policy optimization-based reasoning models. *arXiv preprint arXiv:2503.22342*, 2025a.
- Zhihang Lin, Mingbao Lin, Yuan Xie, and Rongrong Ji. Cppo: Accelerating the training of group relative policy optimization-based reasoning models, 2025b. URL <https://arxiv.org/abs/2503.22342>.
- Tianqi Liu, Yao Zhao, Rishabh Joshi, Misha Khalman, Mohammad Saleh, Peter J. Liu, and Jialu Liu. Statistical rejection sampling improves preference optimization, 2024. URL <https://arxiv.org/abs/2309.06657>.
- Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding rl-zero-like training: A critical perspective. *arXiv preprint arXiv:2503.20783*, 2025a.
- Zikang Liu, Tongtian Yue, Yepeng Tang, Longteng Guo, Junxian Cai, Qingbin Liu, Xi Chen, and Jing Liu. Prefix grouper: Efficient grpo training through shared-prefix forward, 2025b. URL <https://arxiv.org/abs/2506.05433>.
- Liangchen Luo, Yinxiao Liu, Rosanne Liu, Samrat Phatale, Meiqi Guo, Harsh Lara, Yunxuan Li, Lei Shu, Yun Zhu, Lei Meng, Jiao Sun, and Abhinav Rastogi. Improve Mathematical Reasoning in Language Models by Automated Process Supervision, December 2024. URL <http://arxiv.org/abs/2406.06592>. arXiv:2406.06592 [cs].
- Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin A. Riedmiller, Andreas Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharmashan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nat.*, 518(7540):529–533, 2015. doi: 10.1038/NATURE14236. URL <https://doi.org/10.1038/nature14236>.
- Volodymyr Mnih, Adri   Puigdom  nech Badia, Mehdi Mirza, Alex Graves, Timothy P. Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In Maria-Florina Balcan and Kilian Q. Weinberger (eds.), *Proceedings of the 33rd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, volume 48 of *JMLR Workshop and Conference Proceedings*, pp. 1928–1937. JMLR.org, 2016. URL <http://proceedings.mlr.press/v48/mnih16.html>.

Skander Moalla, Andrea Miele, Daniil Pyatko, Razvan Pascanu, and Caglar Gulcehre. No representation, no trust: Connecting representation, collapse, and trust issues in PPO. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/81166fbd9cc5ad14031c5db69d3fd6a8-Abstract-Conference.html.

OpenAI. GPT-4 technical report. *CoRR*, abs/2303.08774, 2023. doi: 10.48550/ARXIV.2303.08774. URL <https://doi.org/10.48550/arXiv.2303.08774>.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kopic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rameesh Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O’Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selman, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report, 2024. URL <https://arxiv.org/abs/2303.08774>.

- 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, 2022a.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022b. URL http://papers.nips.cc/paper_files/paper/2022/hash/blefde53be364a73914f58805a001731-Abstract-Conference.html.
- Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 Technical Report, January 2025. URL <http://arxiv.org/abs/2412.15115>. arXiv:2412.15115 [cs].
- Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. Explain Yourself! Leveraging Language Models for Commonsense Reasoning, June 2019. URL <https://arxiv.org/abs/1906.02361v1>.
- John Schulman, Sergey Levine, Pieter Abbeel, Michael I. Jordan, and Philipp Moritz. Trust region policy optimization. In Francis R. Bach and David M. Blei (eds.), *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pp. 1889–1897. JMLR.org, 2015. URL <http://proceedings.mlr.press/v37/schulman15.html>.
- John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017a.
- John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal Policy Optimization Algorithms, August 2017b. URL <http://arxiv.org/abs/1707.06347>. arXiv:1707.06347 [cs].
- John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. High-Dimensional Continuous Control Using Generalized Advantage Estimation, October 2018. URL <http://arxiv.org/abs/1506.02438>. arXiv:1506.02438 [cs].
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models, April 2024a. URL <http://arxiv.org/abs/2402.03300>. arXiv:2402.03300 [cs].
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024b.
- Marco Simoni, Aleksandar Fontana, Giulio Rossolini, and Andrea Saracino. Gtpo: Trajectory-based policy optimization in large language models. *arXiv preprint arXiv:2508.03772*, 2025.
- Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillcrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul R. Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, Jack Krawczyk,

Cosmo Du, Ed Chi, Heng-Tze Cheng, Eric Ni, Purvi Shah, Patrick Kane, Betty Chan, Manaal Faruqui, Aliaksei Severyn, Hanzhao Lin, YaGuang Li, Yong Cheng, Abe Ittycheriah, Mahdis Mahdieh, Mia Chen, Pei Sun, Dustin Tran, Sumit Bagri, Balaji Lakshminarayanan, Jeremiah Liu, Andras Orban, Fabian Gra, Hao Zhou, Xinying Song, Aurelien Boffy, Harish Ganapathy, Steven Zheng, HyunJeong Choe, goston Weisz, Tao Zhu, Yifeng Lu, Siddharth Gopal, Jarrod Kahn, Maciej Kula, Jeff Pitman, Rushin Shah, Emanuel Taropa, Majd Al Merey, Martin Baeuml, Zhifeng Chen, Laurent El Shafey, Yujing Zhang, Olcan Sercinoglu, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anas White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, Alexandre Frechette, Charlotte Smith, Laura Culp, Lev Proleev, Yi Luan, Xi Chen, James Lottes, Nathan Schucher, Federico Lebron, Alban Rustemi, Natalie Clay, Phil Crone, Tomas Kocisky, Jeffrey Zhao, Bartek Perz, Dian Yu, Heidi Howard, Adam Bloniarz, Jack W. Rae, Han Lu, Laurent Sifre, Marcello Maggioni, Fred Alcober, Dan Garrette, Megan Barnes, Shantanu Thakoor, Jacob Austin, Gabriel Barth-Maron, William Wong, Rishabh Joshi, Rahma Chaabouni, Deeni Fatiha, Arun Ahuja, Gaurav Singh Tomar, Evan Senter, Martin Chadwick, Ilya Kornakov, Nithya Attaluri, Iaki Iturrate, Ruibo Liu, Yunxuan Li, Sarah Cogan, Jeremy Chen, Chao Jia, Chenjie Gu, Qiao Zhang, Jordan Grimstad, Ale Jakse Hartman, Xavier Garcia, Thanumalayan Sankaranarayana Pillai, Jacob Devlin, Michael Laskin, Diego de Las Casas, Dasha Valter, Connie Tao, Lorenzo Blanco, Adri Puigdomnech Badia, David Reitter, Mianna Chen, Jenny Brennan, Clara Rivera, Sergey Brin, Shariq Iqbal, Gabriela Surita, Jane Labanowski, Abhi Rao, Stephanie Winkler, Emilio Parisotto, Yiming Gu, Kate Olszewska, Ravi Addanki, Antoine Miech, Annie Louis, Denis Teplyashin, Geoff Brown, Elliot Catt, Jan Bala-guer, Jackie Xiang, Pidong Wang, Zoe Ashwood, Anton Briukhov, Albert Webson, Sanjay Ganapathy, Smit Sanghavi, Ajay Kannan, Ming-Wei Chang, Axel Stjerngren, Josip Djolonga, Yuting Sun, Ankur Bapna, Matthew Aitchison, Pedram Pejman, Henryk Michalewski, Tianhe Yu, Cindy Wang, Juliette Love, Junwhan Ahn, Dawn Bloxwich, Kehang Han, Peter Humphreys, Thibault Sellam, James Bradbury, Varun Godbole, Sina Samangooei, Bogdan Damoc, Alex Kaskasoli, Sbastien M. R. Arnold, Vijay Vasudevan, Shubham Agrawal, Jason Riesa, Dmitry Lepikhin, Richard Tanburn, Srivatsan Srinivasan, Hyeontaek Lim, Sarah Hodgkinson, Pranav Shyam, Johan Ferret, Steven Hand, Ankush Garg, Tom Le Paine, Jian Li, Yujia Li, Minh Giang, Alexander Neitz, Zaheer Abbas, Sarah York, Machel Reid, Elizabeth Cole, Aakanksha Chowdhery, Dipanjan Das, Dominika Rogoziska, Vitaliy Nikolaev, Pablo Sprechmann, Zachary Nado, Lukas Zilka, Flavien Prost, Luheng He, Marianne Monteiro, Gaurav Mishra, Chris Welty, Josh Newlan, Dawei Jia, Miltiadis Allamanis, Clara Huiyi Hu, Raoul de Liedekerke, Justin Gilmer, Carl Saroufim, Shruti Rijhwani, Shaobo Hou, Disha Shrivastava, Anirudh Baddepudi, Alex Goldin, Adnan Ozturel, Albin Cassirer, Yunhan Xu, Daniel Sohn, Devendra Sachan, Reinald Kim Amplayo, Craig Swanson, Dessie Petrova, Shashi Narayan, Arthur Guez, Siddhartha Brahma, Jessica Landon, Miteyan Patel, Ruizhe Zhao, Kevin Vilella, Luyu Wang, Wenhao Jia, Matthew Rahtz, Mai Gimnez, Legg Yeung, James Keeling, Petko Georgiev, Diana Mincu, Boxi Wu, Salem Haykal, Rachel Saputro, Kiran Vodrahalli, James Qin, Zeynep Cankara, Abhanshu Sharma, Nick Fernando, Will Hawkins, Behnam Neyshabur, Solomon Kim, Adrian Hutter, Priyanka Agrawal, Alex Castro-Ros, George van den Driessche, Tao Wang, Fan Yang, Shuo yiin Chang, Paul Komarek, Ross McIlroy, Mario Lui, Guodong Zhang, Wael Farhan, Michael Sharman, Paul Natsev, Paul Michel, Yamini Bansal, Siyuan Qiao, Kris Cao, Siamak Shakeri, Christina Butterfield, Justin Chung, Paul Kishan Rubenstein, Shivani Agrawal, Arthur Mensch, Kedar Soparkar, Karel Lenc, Timothy Chung, Aedan Pope, Loren Maggiore, Jackie Kay, Priya Jhakra, Shibo Wang, Joshua Maynez, Mary Phuong, Taylor Tobin, Andrea Tacchetti, Maja Trebacz, Kevin Robinson, Yash Katariya, Sebastian Riedel, Paige Bailey, Kefan Xiao, Nimesh Ghelani, Lora Aroyo, Ambrose Slone, Neil Houlsby, Xuehan Xiong, Zhen Yang, Elena Gribovskaya, Jonas Adler, Mateo Wirth, Lisa Lee, Music Li, Thais Kagohara, Jay Pavagadhi, Sophie Bridgers, Anna Bortsova, Sanjay Ghemawat, Zafarali Ahmed, Tianqi Liu, Richard Powell, Vijay Bolina, Mariko Iinuma, Polina Zablotskaia, James Besley, Da-Woon Chung, Timothy Dozat, Ramona Comanescu, Xiance Si, Jeremy Greer, Guolong Su, Martin Polacek, Raphal Lopez Kaufman, Simon Tokumine, Hexiang Hu, Elena Buchatskaya, Yingjie Miao, Mohamed Elhawaty, Aditya Siddhant, Nenad Toma-sev, Jinwei Xing, Christina Greer, Helen Miller, Shereen Ashraf, Aurko Roy, Zizhao Zhang, Ada Ma, Angelos Filos, Milos Besta, Rory Blevins, Ted Klimentko, Chih-Kuan Yeh, Soravit Chang-pinyo, Jiaqi Mu, Oscar Chang, Mantas Pajarskas, Carrie Muir, Vered Cohen, Charline Le Lan, Krishna Haridasan, Amit Marathe, Steven Hansen, Sholto Douglas, Rajkumar Samuel, Mingqiu Wang, Sophia Austin, Chang Lan, Jiepu Jiang, Justin Chiu, Jaime Alonso Lorenzo, Lars Lowe

Sjösund, Sébastien Cevey, Zach Gleicher, Thi Avrahami, Anudhyan Boral, Hansa Srinivasan, Vittorio Selo, Rhys May, Konstantinos Aisopos, Léonard Hussenot, Livio Baldini Soares, Kate Baumli, Michael B. Chang, Adrià Recasens, Ben Caine, Alexander Pritzel, Filip Pavetic, Fabio Pardo, Anita Gergely, Justin Frye, Vinay Ramasesh, Dan Horgan, Kartikeya Badola, Nora Kassner, Subhrajit Roy, Ethan Dyer, Víctor Campos Campos, Alex Tomala, Yunhao Tang, Dalia El Badawy, Elspeth White, Basil Mustafa, Oran Lang, Abhishek Jindal, Sharad Vikram, Zhitao Gong, Sergi Caelles, Ross Hemsley, Gregory Thornton, Fangxiaoyu Feng, Wojciech Stokowiec, Ce Zheng, Phoebe Thacker, Çağlar Ünlü, Zhishuai Zhang, Mohammad Saleh, James Svensson, Max Bileschi, Piyush Patil, Ankesh Anand, Roman Ring, Katerina Tsihlias, Arpi Vezer, Marco Selvi, Toby Shevlane, Mikel Rodriguez, Tom Kwiatkowski, Samira Daruki, Keran Rong, Allan Dafoe, Nicholas FitzGerald, Keren Gu-Lemberg, Mina Khan, Lisa Anne Hendricks, Marie Pellet, Vladimir Feinberg, James Cobon-Kerr, Tara Sainath, Maribeth Rauh, Sayed Hadi Hashemi, Richard Ives, Yana Hasson, Eric Noland, Yuan Cao, Nathan Byrd, Le Hou, Qingze Wang, Thibault Sottiaux, Michela Paganini, Jean-Baptiste Lespiau, Alexandre Moufarek, Samer Hassan, Kaushik Shivakumar, Joost van Amersfoort, Amol Mandhane, Pratik Joshi, Anirudh Goyal, Matthew Tung, Andrew Brock, Hannah Sheahan, Vedant Misra, Cheng Li, Nemanja Rakićević, Mostafa Dehghani, Fangyu Liu, Sid Mittal, Junhyuk Oh, Seb Noury, Eren Sezener, Fantine Huot, Matthew Lamm, Nicola De Cao, Charlie Chen, Sidharth Mudgal, Romina Stella, Kevin Brooks, Gautam Vasudevan, Chenxi Liu, Mainak Chain, Nivedita Melinker, Aaron Cohen, Venus Wang, Kristie Seymore, Sergey Zubkov, Rahul Goel, Summer Yue, Sai Krishnakumaran, Brian Albert, Nate Hurley, Motoki Sano, Anhad Mohanane, Jonah Joughin, Egor Filonov, Tomasz Kepa, Yomna Eldawy, Jiawern Lim, Rahul Rishi, Shirin Badiehzadegan, Taylor Bos, Jerry Chang, Sanil Jain, Sri Gayatri Sundara Padmanabhan, Subha Puttagunta, Kalpesh Krishna, Leslie Baker, Norbert Kalb, Vamsi Bedapudi, Adam Kurzrok, Shuntong Lei, Anthony Yu, Oren Litvin, Xiang Zhou, Zhichun Wu, Sam Sobell, Andrea Siciliano, Alan Papir, Robby Neale, Jonas Braggagnolo, Tej Toor, Tina Chen, Valentin Anklin, Feiran Wang, Richie Feng, Milad Gholami, Kevin Ling, Lijuan Liu, Jules Walter, Hamid Moghaddam, Arun Kishore, Jakub Adamek, Tyler Mercado, Jonathan Mallinson, Siddhinita Wandekar, Stephen Cagle, Eran Ofek, Guillermo Garrido, Clemens Lombriser, Maksim Mukha, Botu Sun, Hafeezul Rahman Mohammad, Josip Matak, Yadi Qian, Vikas Peswani, Pawel Janus, Quan Yuan, Leif Schelin, Oana David, Ankur Garg, Yifan He, Oleksii Duzhyi, Anton Älgmyr, Timothée Lottaz, Qi Li, Vikas Yadav, Luyao Xu, Alex Chinien, Rakesh Shivanna, Aleksandr Chuklin, Josie Li, Carrie Spadine, Travis Wolfe, Kareem Mohamed, Subhabrata Das, Zihang Dai, Kyle He, Daniel von Dincklage, Shyam Upadhyay, Akanksha Maurya, Luyan Chi, Sebastian Krause, Khalid Salama, Pam G Rabinovitch, Pavan Kumar Reddy M, Aarush Selvan, Mikhail Dektiarev, Golnaz Ghiasi, Erdem Guven, Himanshu Gupta, Boyi Liu, Deepak Sharma, Idan Heimlich Shtacher, Shachi Paul, Oscar Akerlund, François-Xavier Aubet, Terry Huang, Chen Zhu, Eric Zhu, Elico Teixeira, Matthew Fritze, Francesco Bertolini, Liana-Eleonora Marinescu, Martin Bölle, Dominik Paulus, Khyatti Gupta, Tejasi Latkar, Max Chang, Jason Sanders, Roopa Wilson, Xuewei Wu, Yi-Xuan Tan, Lam Nguyen Thiet, Tulsee Doshi, Sid Lall, Swaroop Mishra, Wanming Chen, Thang Luong, Seth Benjamin, Jasmine Lee, Ewa Andrejczuk, Dominik Rabiej, Vipul Ranjan, Krzysztof Styrz, Pengcheng Yin, Jon Simon, Malcolm Rose Harriott, Mudit Bansal, Alexei Robsky, Geoff Bacon, David Greene, Daniil Mirylenka, Chen Zhou, Obaid Sarvana, Abhimanyu Goyal, Samuel Andermatt, Patrick Siegler, Ben Horn, Assaf Israel, Francesco Pongetti, Chih-Wei "Louis" Chen, Marco Selvatici, Pedro Silva, Kathie Wang, Jackson Tolins, Kelvin Guu, Roey Yoge, Xiaochen Cai, Alessandro Agostini, Maulik Shah, Hung Nguyen, Noah Ó Donnaile, Sébastien Pereira, Linda Friso, Adam Stambler, Adam Kurzrok, Chenkai Kuang, Yan Romanikhin, Mark Geller, ZJ Yan, Kane Jang, Cheng-Chun Lee, Wojciech Fica, Eric Malmi, Qijun Tan, Dan Banica, Daniel Balle, Ryan Pham, Yanping Huang, Diana Avram, Hongzhi Shi, Jasjot Singh, Chris Hidey, Niharika Ahuja, Pranab Saxena, Dan Dooley, Srividya Pranavi Potharaju, Eileen O'Neill, Anand Gokulchandran, Ryan Foley, Kai Zhao, Mike Dusenberry, Yuan Liu, Pulkit Mehta, Ragha Kotikalapudi, Chalence Safranek-Shrader, Andrew Goodman, Joshua Kessinger, Eran Globen, Prateek Kolhar, Chris Golewski, Ali Ibrahim, Yang Song, Ali Eichenbaum, Thomas Brovelli, Sahitya Potluri, Preethi Lahoti, Cip Baetu, Ali Ghorbani, Charles Chen, Andy Crawford, Shalini Pal, Mukund Sridhar, Petru Gurita, Asier Mujika, Igor Petrovski, Pierre-Louis Cedoz, Chenmei Li, Shiyuan Chen, Nicolò Dal Santo, Siddharth Goyal, Jitesh Punjabi, Karthik Kappaganthu, Chester Kwak, Pallavi LV, Sarmishta Velury, Himadri Choudhury, Jamie Hall, Premal Shah, Ricardo Figueira, Matt Thomas, Minjie Lu, Ting Zhou, Chintu Kumar, Thomas Jurdi, Sharat Chikkerur, Yenai Ma, Adams Yu, Soo Kwak, Victor Ahdel, Sujeevan Rajayogam, Travis Choma, Fei Liu, Aditya Barua, Colin Ji, Ji Ho

Park, Vincent Hellendoorn, Alex Bailey, Taylan Bilal, Huanjie Zhou, Mehrdad Khatir, Charles Sutton, Wojciech Rzadkowski, Fiona Macintosh, Roopali Vij, Konstantin Shagin, Paul Medina, Chen Liang, Jinjing Zhou, Pararth Shah, Yingying Bi, Attila Dankovics, Shipra Banga, Sabine Lehmann, Marissa Bredesen, Zifan Lin, John Eric Hoffmann, Jonathan Lai, Raynald Chung, Kai Yang, Nihal Balani, Arthur Bražinskas, Andrei Sozanschi, Matthew Hayes, Héctor Fernández Alcalde, Peter Makarov, Will Chen, Antonio Stella, Liselotte Snijders, Michael Mandl, Ante Kärman, Paweł Nowak, Xinyi Wu, Alex Dyck, Krishnan Vaidyanathan, Raghavender R, Jessica Mallet, Mitch Rudominer, Eric Johnston, Sushil Mittal, Akhil Udathu, Janara Christensen, Vishal Verma, Zach Irving, Andreas Santucci, Gamaleldin Elsayed, Elnaz Davoodi, Marin Georgiev, Ian Tenney, Nan Hua, Geoffrey Cideron, Edouard Leurent, Mahmoud Alnahlawi, Ionut Georgescu, Nan Wei, Ivy Zheng, Dylan Scandinaro, Heinrich Jiang, Jasper Snoek, Mukund Sundararajan, Xuezhi Wang, Zack Ontiveros, Itay Karo, Jeremy Cole, Vinu Rajashekhar, Lara Tume, Eyal Ben-David, Rishub Jain, Jonathan Uesato, Romina Datta, Oskar Bunyan, Shimu Wu, John Zhang, Piotr Stanczyk, Ye Zhang, David Steiner, Subhajit Naskar, Michael Azzam, Matthew Johnson, Adam Paszke, Chung-Cheng Chiu, Jaume Sanchez Elias, Afroz Mohiuddin, Faizan Muhammad, Jin Miao, Andrew Lee, Nino Vieillard, Jane Park, Jiageng Zhang, Jeff Stanway, Drew Garmon, Abhijit Karmarkar, Zhe Dong, Jong Lee, Aviral Kumar, Luowei Zhou, Jonathan Evens, William Isaac, Geoffrey Irving, Edward Loper, Michael Fink, Isha Arkatkar, Nanxin Chen, Izhak Shafran, Ivan Petrychenko, Zhe Chen, Johnson Jia, Anselm Levskaya, Zhenkai Zhu, Peter Grabowski, Yu Mao, Alberto Magni, Kaisheng Yao, Javier Snider, Norman Casagrande, Evan Palmer, Paul Suganthan, Alfonso Castaño, Irene Giannoumis, Wooyeol Kim, Mikołaj Rybiński, Ashwin Sreevatsa, Jennifer Prendki, David Soergel, Adrian Goedeckemeyer, Willi Gierke, Mohsen Jafari, Meenu Gaba, Jeremy Wiesner, Diana Gage Wright, Yawen Wei, Harsha Vashisht, Yana Kulizhskaya, Jay Hoover, Maigo Le, Lu Li, Chimezie Iwuanyanwu, Lu Liu, Kevin Ramirez, Andrey Khorlin, Albert Cui, Tian LIN, Marcus Wu, Ricardo Aguilar, Keith Pallo, Abhishek Chakladar, Ginger Perng, Elena Allica Abellan, Mingyang Zhang, Ishita Dasgupta, Nate Kushman, Ivo Penchev, Alena Repina, Xihui Wu, Tom van der Weide, Priya Ponnappalli, Caroline Kaplan, Jiri Simsa, Shuangfeng Li, Olivier Dousse, Fan Yang, Jeff Piper, Nathan Ie, Rama Pasmurthy, Nathan Lintz, Anitha Vijayakumar, Daniel Andor, Pedro Valenzuela, Minnie Lui, Cosmin Paduraru, Daiyi Peng, Katherine Lee, Shuyuan Zhang, Somer Greene, Duc Dung Nguyen, Paula Kurylowicz, Cassidy Hardin, Lucas Dixon, Lili Janzer, Kiam Choo, Ziqiang Feng, Biao Zhang, Achintya Singhal, Dayou Du, Dan McKinnon, Natasha Antropova, Tolga Bolukbasi, Orgad Keller, David Reid, Daniel Finchelstein, Maria Abi Raad, Remi Crocker, Peter Hawkins, Robert Dadashi, Colin Gaffney, Ken Franko, Anna Bulanova, Rémi Leblond, Shirley Chung, Harry Askham, Luis C. Cobo, Kelvin Xu, Felix Fischer, Jun Xu, Christina Sorokin, Chris Alberti, Chu-Cheng Lin, Colin Evans, Alek Dimitriev, Hannah Forbes, Dylan Banarse, Zora Tung, Mark Omernick, Colton Bishop, Rachel Sterneck, Rohan Jain, Jiawei Xia, Ehsan Amid, Francesco Piccinno, Xingyu Wang, Praseem Banzal, Daniel J. Mankowitz, Alex Polozov, Victoria Krakovna, Sasha Brown, MohammadHossein Bateni, Dennis Duan, Vlad Firoiu, Meghana Thotakuri, Tom Natan, Matthieu Geist, Ser tan Girgin, Hui Li, Jiayu Ye, Ofir Roval, Reiko Tojo, Michael Kwong, James Lee-Thorp, Christopher Yew, Danila Sinopalnikov, Sabela Ramos, John Mellor, Abhishek Sharma, Kathy Wu, David Miller, Nicolas Sonnerat, Denis Vnukov, Rory Greig, Jennifer Beattie, Emily Caveness, Libin Bai, Julian Eisenschlos, Alex Korchemniy, Tomy Tsai, Mimi Jasarevic, Weize Kong, Phuong Dao, Zeyu Zheng, Frederick Liu, Fan Yang, Rui Zhu, Tian Huey Teh, Jason Sanmiya, Evgeny Gladchenko, Nejc Trdin, Daniel Toyama, Evan Rosen, Sasan Tavakkol, Linting Xue, Chen Elkind, Oliver Woodman, John Carpenter, George Papamakarios, Rupert Kemp, Sushant Kafle, Tanya Grunina, Rishika Sinha, Alice Talbert, Diane Wu, Denese Owusu-Afriyie, Cosmo Du, Chloe Thornton, Jordi Pont-Tuset, Pradyumna Narayana, Jing Li, Saaber Fatehi, John Wieting, Omar Ajmeri, Benigno Uribe, Yeongil Ko, Laura Knight, Amélie Hélie, Ning Niu, Shane Gu, Chenxi Pang, Yeqing Li, Nir Levine, Ariel Stolovich, Rebeca Santamaria-Fernandez, Sonam Goenka, Wenny Yustalim, Robin Strudel, Ali Elqursh, Charlie Deck, Hyo Lee, Zonglin Li, Kyle Levin, Raphael Hoffmann, Dan Holtmann-Rice, Olivier Bachem, Sho Arora, Christy Koh, Soheil Hassas Yeganeh, Siim Pöder, Mukarram Tariq, Yanhua Sun, Lucian Ionita, Mojtaba Seyedhosseini, Pouya Tafti, Zhiyu Liu, Anmol Gulati, Jasmine Liu, Xinyu Ye, Bart Chrzasczcz, Lily Wang, Nikhil Sethi, Tianrun Li, Ben Brown, Shreya Singh, Wei Fan, Aaron Parisi, Joe Stanton, Vinod Koverkathu, Christopher A. Choquette-Choo, Yunjie Li, TJ Lu, Abe Ittycheriah, Prakash Shroff, Mani Varadarajan, Sanaz Bahargam, Rob Willoughby, David Gaddy, Guillaume Desjardins, Marco Cornero, Brona Robenek, Bhavishya Mittal, Ben Albrecht, Ashish Shenoy, Fedor Moiseev, Henrik Jacobsson, Alireza Ghaffarkhah, Morgane Riviére, Alanna Wal-

- ton, Clément Crepy, Alicia Parrish, Zongwei Zhou, Clement Farabet, Carey Radebaugh, Praveen Srinivasan, Claudia van der Salm, Andreas Fidjeland, Salvatore Scellato, Eri Latorre-Chimoto, Hanna Klimczak-Plucińska, David Bridson, Dario de Cesare, Tom Hudson, Piermaria Mendolichio, Lexi Walker, Alex Morris, Matthew Mauger, Alexey Guseynov, Alison Reid, Seth Odoom, Lucia Loher, Victor Cotruta, Madhavi Yenugula, Dominik Grewe, Anastasia Petrushkina, Tom Duerig, Antonio Sanchez, Steve Yadlowsky, Amy Shen, Amir Globerson, Lynette Webb, Sahil Dua, Dong Li, Surya Bhupatiraju, Dan Hurt, Haroon Qureshi, Ananth Agarwal, Tomer Shani, Matan Eyal, Anuj Khare, Shreyas Rammohan Belle, Lei Wang, Chetan Tekur, Mihir Sanjay Kale, Jinliang Wei, Ruoxin Sang, Brennan Saeta, Tyler Liechty, Yi Sun, Yao Zhao, Stephan Lee, Pandu Nayak, Doug Fritz, Manish Reddy Vuyyuru, John Aslanides, Nidhi Vyas, Martin Wicke, Xiao Ma, Evgenii Eltyshov, Nina Martin, Hardie Cate, James Manyika, Keyvan Amiri, Yelin Kim, Xi Xiong, Kai Kang, Florian Luisier, Nilesh Tripuraneni, David Madras, Mandy Guo, Austin Waters, Oliver Wang, Joshua Ainslie, Jason Baldridge, Han Zhang, Garima Pruthi, Jakob Bauer, Feng Yang, Riham Mansour, Jason Gelman, Yang Xu, George Polovets, Ji Liu, Honglong Cai, Warren Chen, XiangHai Sheng, Emily Xue, Sherjil Ozair, Christof Angermueller, Xiaowei Li, Anoop Sinha, Weiren Wang, Julia Wiesinger, Emmanouil Koukoumidis, Yuan Tian, Anand Iyer, Madhu Gurusurthy, Mark Goldenson, Parashar Shah, MK Blake, Hongkun Yu, Anthony Urbanowicz, Jennimaria Palomaki, Chrisantha Fernando, Ken Durden, Harsh Mehta, Nikola Momchev, Elahé Rahimtoroghi, Maria Georgaki, Amit Raul, Sebastian Ruder, Morgan Redshaw, Jinhyuk Lee, Denny Zhou, Komal Jalan, Dinghua Li, Blake Hechtman, Parker Schuh, Milad Nasr, Kieran Milan, Vladimir Mikulik, Juliana Franco, Tim Green, Nam Nguyen, Joe Kelley, Aroma Mahendru, Andrea Hu, Joshua Howland, Ben Vargas, Jeffrey Hui, Kshitij Bansal, Vikram Rao, Rakesh Ghiya, Emma Wang, Ke Ye, Jean Michel Sarr, Melanie Moranski Preston, Madeleine Elish, Steve Li, Aakash Kaku, Jigar Gupta, Ice Pasupat, Da-Cheng Juan, Milan Someswar, Tejvi M., Xinyun Chen, Aida Amini, Alex Fabrikant, Eric Chu, Xuanyi Dong, Amruta Muthal, Senaka Butthipitiya, Sarthak Jauhari, Nan Hua, Urvashi Khandelwal, Ayal Hitron, Jie Ren, Larissa Rinaldi, Shahar Drath, Avigail Dabush, Nan-Jiang Jiang, Harshal Godhia, Uli Sachs, Anthony Chen, Yicheng Fan, Hagai Taitelbaum, Hila Noga, Zhuyun Dai, James Wang, Chen Liang, Jenny Hamer, Chun-Sung Ferng, Chenel Elkind, Aviel Atias, Paulina Lee, Vít Lístík, Mathias Carlen, Jan van de Kerkhof, Marcin Pikuś, Krunoslav Zaher, Paul Müller, Sasha Zykova, Richard Stefanec, Vitaly Gatsko, Christoph Hirsichall, Ashwin Sethi, Xingyu Federico Xu, Chetan Ahuja, Beth Tsai, Anca Stefanoiu, Bo Feng, Keshav Dhandhanian, Manish Katyal, Akshay Gupta, Atharva Parulekar, Divya Pitta, Jing Zhao, Vivaan Bhatia, Yashodha Bhavnani, Omar Alhadlaq, Xiaolin Li, Peter Danenberg, Dennis Tu, Alex Pine, Vera Filippova, Abhipso Ghosh, Ben Limonchik, Bhargava Urala, Chaitanya Krishna Lanka, Derik Clive, Yi Sun, Edward Li, Hao Wu, Kevin Hongtongsak, Ianna Li, Kalind Thakkar, Kuanysh Omarov, Kushal Majmundar, Michael Alverson, Michael Kucharski, Mohak Patel, Mudit Jain, Maksim Zabelin, Paolo Pelagatti, Rohan Kohli, Saurabh Kumar, Joseph Kim, Swetha Sankar, Vineet Shah, Lakshmi Ramachandruni, Xiangkai Zeng, Ben Bariach, Laura Weidinger, Tu Vu, Alek Andreev, Antoine He, Kevin Hui, Sheleem Kashem, Amar Subramanya, Sissie Hsiao, Demis Hassabis, Koray Kavukcuoglu, Adam Sadovsky, Quoc Le, Trevor Strohman, Yonghui Wu, Slav Petrov, Jeffrey Dean, and Oriol Vinyals. Gemini: A family of highly capable multimodal models, 2025. URL <https://arxiv.org/abs/2312.11805>.
- Ruonan Wang, Runxi Wang, Yunwen Shen, Chengfeng Wu, Qinglin Zhou, and Rohitash Chandra. Evaluation of llms for mathematical problem solving, 2025. URL <https://arxiv.org/abs/2506.00309>.
- Yuhui Wang, Hao He, Xiaoyang Tan, and Yaozhong Gan. Trust region-guided proximal policy optimization. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d’Alché-Buc, Emily B. Fox, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 624–634, 2019. URL <https://proceedings.neurips.cc/paper/2019/hash/a666587afda6e89aec274a3657558a27-Abstract.html>.
- Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent Abilities of Large Language Models, October 2022. URL <http://arxiv.org/abs/2206.07682>. arXiv:2206.07682 [cs].

- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models, January 2023a. URL <http://arxiv.org/abs/2201.11903>. arXiv:2201.11903 [cs].
- Tianwen Wei, Jian Luan, Wei Liu, Shuang Dong, and Bin Wang. CMATH: Can Your Language Model Pass Chinese Elementary School Math Test?, June 2023b. URL <http://arxiv.org/abs/2306.16636>. arXiv:2306.16636 [cs].
- Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8(3):229–256, May 1992. ISSN 1573-0565. doi: 10.1007/BF00992696. URL <https://doi.org/10.1007/BF00992696>.
- Haotian Xu, Zheng Yan, Junyu Xuan, Guangquan Zhang, and Jie Lu. Improving proximal policy optimization with alpha divergence. *Neurocomputing*, 534:94–105, 2023. doi: 10.1016/J.NEUCOM.2023.02.008. URL <https://doi.org/10.1016/j.neucom.2023.02.008>.
- 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, September 2024. URL <http://arxiv.org/abs/2409.12122>. arXiv:2409.12122 [cs].
- Zhihe Yang, Xufang Luo, Zilong Wang, Dongqi Han, Zhiyuan He, Dongsheng Li, and Yunjian Xu. Do not let low-probability tokens over-dominate in RL for llms. *CoRR*, abs/2505.12929, 2025. doi: 10.48550/ARXIV.2505.12929. URL <https://doi.org/10.48550/arXiv.2505.12929>.
- Qiyang Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Mu Qiao, Yonghui Wu, and Mingxuan Wang. DAPO: An Open-Source LLM Reinforcement Learning System at Scale, May 2025. URL <http://arxiv.org/abs/2503.14476>. arXiv:2503.14476 [cs].
- Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah D. Goodman. STaR: Bootstrapping Reasoning With Reasoning, May 2022. URL <http://arxiv.org/abs/2203.14465>. arXiv:2203.14465 [cs].
- Rui Zheng, Shihan Dou, Songyang Gao, Yuan Hua, Wei Shen, Binghai Wang, Yan Liu, Senjie Jin, Qin Liu, Yuhao Zhou, Limao Xiong, Lu Chen, Zhiheng Xi, Nuo Xu, Wenbin Lai, Minghao Zhu, Cheng Chang, Zhangyue Yin, Rongxiang Weng, Wensen Cheng, Haoran Huang, Tianxiang Sun, Hang Yan, Tao Gui, Qi Zhang, Xipeng Qiu, and Xuanjing Huang. Secrets of RLHF in Large Language Models Part I: PPO, July 2023. URL <http://arxiv.org/abs/2307.04964>. arXiv:2307.04964 [cs].
- Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. AGIEval: A Human-Centric Benchmark for Evaluating Foundation Models, September 2023. URL <http://arxiv.org/abs/2304.06364>. arXiv:2304.06364 [cs].

A HYPERPARAMETERS USED

Table 4 summarizes the training hyperparameters.

Table 4: Hyper-parameters used for training on GSM8K.

Experiment	Hyper-parameter	Value
GRPO	Batch size	24
	Learning rate	1×10^{-5}
	Group size G	8
	Temperature	1
	KL coefficient	0.005
	Max new tokens	512
	Warmup steps	5
GRPO_pos	Batch size	24
	Learning rate	1×10^{-5}
	Group size G	8
	Temperature	1
	KL coefficient	0.005
	Max new tokens	512
	Warmup steps	5
RGRA	Batch size	24
	Learning rate	1×10^{-5}
	Group size G	8
	Temperature	1
	KL coefficient	0.005
	Max new tokens	512
	Warmup steps	5
REINFORCE	Batch size	24
	Learning rate	1×10^{-5}
	Group size G	8
	Temperature	1
	KL coefficient	0.005
	Max new tokens	512
	Warmup steps	5
RAFT	Batch size	24
	Update epochs per stage	1
	Learning rate	1×10^{-5}
	Group size G	8
	Temperature	1
	Max new tokens	512
	Warmup steps	5
SFT	Learning rate	1×10^{-4}
	Decay mode	Linear
	Epochs	1
	Batch size	8
	Warmup steps	5