ON THE OPTIMIZATION DYNAMICS OF RLVR: GRADIENT GAP AND STEP SIZE THRESHOLDS

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement Learning with Verifiable Rewards (RLVR), which uses simple binary feedback to post-train large language models, has shown significant empirical success. However, a principled understanding of why it works has been lacking. This paper builds a theoretical foundation for RLVR by analyzing its training process at both the full-response (trajectory) and token levels. Central to our analysis is a quantity called the *Gradient Gap*, which formalizes the direction of improvement from low-reward to high-reward regions of the response space. We prove that convergence critically depends on aligning the update direction with this Gradient Gap. Moreover, we derive a sharp step-size threshold based on the magnitude of the Gradient Gap: below it, learning converges, whereas above it, performance collapses. Our theory further predicts how the critical step size must scale with *response length* and the *success rate*, thereby explaining why practical heuristics such as *length normalization* improve stability and showing that, with a fixed learning rate, the success rate can stagnate strictly below 100%. We validate these predictions through controlled bandit simulations and LLM experiments, including training Qwen2.5-7B with GRPO.

1 INTRODUCTION

Large language models (LLMs) have recently achieved significant advances through reinforcement learning post-training, which aligns them with complex tasks and preferences (Ziegler et al., 2019; Ouyang et al., 2022; Shao et al., 2024; Team et al., 2025). In particular, Reinforcement Learning with Verifiable Rewards (RLVR) has emerged as a powerful approach for post-training LLMs on tasks where success can be automatically checked (e.g. using a compiler or solver) (Guo et al., 2025). RLVR methods have shown impressive empirical gains by leveraging binary success/failure feedback instead of human judgments, thereby simplifying the RL pipeline. Techniques in this vein (e.g. variants of Proximal Policy Optimization, a.k.a. PPO (Schulman et al., 2017), like GRPO (Shao et al., 2024), DAPO (Yu et al., 2025b), Dr. GRPO (Liu et al., 2025), etc.) eliminate the need for learned reward or value models, relying purely on verifiable outcome signals. This has enabled LLMs to achieve state-of-the-art results on challenging reasoning and code-generation tasks, demonstrating the promise of RLVR-driven fine-tuning.

However, empirical progress in RLVR has far outpaced our theoretical understanding. The optimization process remains largely a black box: we do not fully understand why RL-based post-training works so well, under what conditions it might falter, or how to tune it for stable convergence. Recent PPO-based variants (e.g. GRPO, DAPO, Dr. GRPO) have sprung up to improve training stability, each introducing different heuristics like normalizing gradient updates by the output length or standardizing rewards by the group's variance. Yet it remains unclear which of these design choices truly matter; without a principled basis, their adoption is guided more by intuition than by theory. This gap is especially pronounced given RLVR's sparse binary rewards (each episode yields just a single success/failure bit), which make it difficult to analyze how gradient descent navigates the model's vast parameter space or how the policy's output distribution shifts toward higher-reward answers. In practice, practitioners often resort to trial-and-error for critical hyperparameters and algorithmic choices, where a mis-tuning can destabilize training or even cause catastrophic collapse (e.g., forgetting pre-trained knowledge or converging to trivial outputs). These challenges underscore the need for a rigorous theoretical framework to demystify RLVR's optimization dynamics and reduce reliance on guesswork.

This work establishes such a framework, providing a rigorous theoretical foundation for RLVR in LLM post-training. Our key contributions are:

- **Unified RLVR theory:** We develop a principled framework for RLVR under binary rewards, introducing the *Gradient Gap* to characterize the improvement direction from low- to high-reward responses.
- **Convergence guarantees:** We prove the existence of a sharp step-size threshold that separates stable convergence from divergence, providing clear guidance for safe hyperparameter tuning.
- **Length- and success-aware learning rates:** Our theory shows that the effective learning rate must shrink with output length and adapt to task difficulty, offering a theoretical explanation for the stabilizing effect of heuristics such as length normalization and clarifying why fixed step sizes can cause stagnation.
- **Empirical validation:** We validate our theory through bandit simulations and LLM experiments, including fine-tuning Qwen2.5-7B on GSM8K and DAPO17k datasets with GRPO, demonstrating close alignment between theory and practice.

1.1 Related Works

Recent efforts in RL-based language model post-training have introduced a family of GRPO-style algorithms that extend or modify Proximal Policy Optimization (PPO) for verifiable feedback settings. GRPO itself eliminates the value critic by estimating advantages from a group of sampled responses, using relative reward normalization instead of a learned baseline (Shao et al., 2024). Building on this idea, DAPO augmented GRPO by decoupling the PPO clipping range and dynamically filtering out cases where all responses in a batch are correct or all are incorrect (Yu et al., 2025b). Dr. GRPO revisits the advantage normalization procedure, arguing that removing length and variance normalizations (i.e. using only a mean baseline) can prevent bias in policy updates (Liu et al., 2025). Additional related papers are discussed in Appendix A.

In parallel, theoretical work has established convergence guarantees for policy gradient (PG) methods, including REINFORCE and actor-critic algorithms. In finite Markov decision processes with softmax policies, the PG objective often satisfies a PL condition, implying that any stationary point is globally optimal and that vanilla gradient ascent converges at a sublinear rate (Agarwal et al., 2021; Xiao, 2022). Actor-critic methods also achieve provable convergence by using two-timescale updates or pessimistic value estimation (Wu et al., 2020; Zanette et al., 2021). However, extending these guarantees to post-training large language models with verifiable binary rewards remains challenging, as sparse success/failure signals provide very limited gradient information.

2 Problem Set-Up

Language Model. We begin with a standard language model parameterized by $\theta \in \mathbb{R}^d$, which defines a conditional distribution $\pi_\theta(\vec{\boldsymbol{o}} \mid q)$ over sequences of tokens $\vec{\boldsymbol{o}} = (o_1, o_2, \ldots, o_{|\vec{\boldsymbol{o}}|})$ given an input prompt/question q. Output tokens $\{o_t\}_{t=1}^{|\vec{\boldsymbol{o}}|}$ are drawn from a finite vocabulary $\mathcal{T}$, and the generation process ends when the special end-of-sequence token $o_{|\vec{\boldsymbol{o}}|} = \texttt{EOS}$ is emitted.

The model generates tokens in an *autoregressive* fashion: at every step t, the next token o_t is sampled conditioned on the prompt q and all previously generated tokens $\vec{\boldsymbol{o}}_{<t} = (o_1, o_2, \ldots, o_{t-1})$. Formally, $\pi_\theta(\vec{\boldsymbol{o}} \mid q) = \prod_{t=1}^{|\vec{\boldsymbol{o}}|} \pi_\theta(o_t \mid q, \vec{\boldsymbol{o}}_{<t})$. Each conditional distribution is defined by a softmax over token logits $\boldsymbol{h}_\theta(\cdot)$:

$$\pi_\theta(o_t \mid q, \vec{\boldsymbol{o}}_{<t}) \;:=\; \frac{\exp\{\boldsymbol{h}_\theta(q, \vec{\boldsymbol{o}}_{\le t})\}}{\sum_{o' \in \mathcal{T}} \exp\{\boldsymbol{h}_\theta(q, \vec{\boldsymbol{o}}_{<t}, o')\}}\,. \tag{1}$$

Post-Training: Reinforcement Learning with Verifiable Rewards (RLVR). While a supervised language model can generate fluent text, it often struggles to align with task-specific goals such as math reasoning or code generation. Post-training addresses this limitation by adapting the model parameters θ to align more closely with an external reward signal that captures desirable behavior.

Formally, we assume access to an outcome reward model $r^\star(q, \vec{\boldsymbol{o}})$ that is directly verifiable and assessed at the end of a generated sequence: $r^\star = 1$ if the answer is correct (e.g., a valid proof step

or passing code execution) and $r^\star = 0$ otherwise. The aim of reinforcement learning in this context is to tune the model parameters θ so as to maximize the expected reward under the current policy:

$$\text{maximize}_{\theta \in \mathbb{R}^d} \quad J(\pi_\theta) \;:=\; \mathbb{E}_{q \sim \mathbb{P}(Q),\, \vec{\boldsymbol{o}} \sim \pi_\theta(\cdot|q)} \big[r^\star(q, \vec{\boldsymbol{o}}) \big] \,. \tag{2}$$

Policy Gradient. To optimize $J(\pi_\theta)$, we rely on policy gradient–based methods. At each iteration t, the parameters θ are updated according to

$$\theta_{k+1} \;=\; \theta_k + \eta_k \cdot \boldsymbol{w}_k \,, \tag{3}$$

where $\eta_k \geq 0$ is the learning rate and $\boldsymbol{w}_k \in \mathbb{R}^d$ is a normalized update direction with $\|\boldsymbol{w}_k\|_2 \leq 1$. For clarity, we denote the policy and logit function at step k as $\pi_k := \pi_{\theta_k}$ and $\boldsymbol{h}_k := \boldsymbol{h}_{\theta_k}$.

This generic formulation captures a broad family of post-training algorithms used in RLVR. Representative examples are:

REINFORCE: The classical policy gradient method updates parameters in the direction

$$\nabla_\theta \, J(\pi_k) \;=\; \mathbb{E}_{q \sim \mathbb{P}(Q),\, \vec{\boldsymbol{o}} \sim \pi_k(\cdot|q)} \big[A(q, \vec{\boldsymbol{o}}) \cdot \nabla_\theta \, \log \pi_k(\vec{\boldsymbol{o}} \mid q) \big] \,, \tag{4}$$

where the advantage function is given by $A(q, \vec{\boldsymbol{o}}) = r^\star(q, \vec{\boldsymbol{o}}) - \mathbb{E}_{\vec{\boldsymbol{o}}' \sim \pi_k(\cdot|q)}[r^\star(q, \vec{\boldsymbol{o}}')]$. In this case, the update rule $\theta_{t+1} = \theta_k + \alpha \cdot \nabla_\theta \, J(\pi_k)$ can be rewritten in our generic form by setting $\boldsymbol{w}_k = \nabla_\theta \, J(\pi_k) / \|\nabla_\theta \, J(\pi_k)\|_2$ and $\eta_k = \alpha \, \|\nabla_\theta \, J(\pi_k)\|_2$. Viewed in this way, Dr. GRPO (Liu et al., 2025) emerges as a variant that replaces the single-sample advantage with a group-wise demeaned version.

Group Relative Policy Optimization (GRPO): GRPO has recently become a standard choice for RLVR. The full algorithm incorporates clipping ratios and multi-step updates (see Appendix B.1). To connect it with the generic policy gradient form, we consider a simplified one-step approximation without clipping. In this case, the gradient direction is

$$\boldsymbol{g}_{\text{GRPO}}(\pi_k) \;=\; \mathbb{E}_{q \sim \mathbb{P}(Q),\, \vec{\boldsymbol{o}} \sim \pi_k(\cdot|q)} \Big[\frac{A(q, \vec{\boldsymbol{o}})}{\sigma(q)} \cdot \frac{1}{|\vec{\boldsymbol{o}}|} \nabla_\theta \, \log \pi_k(\vec{\boldsymbol{o}} \mid q) \Big] \,, \tag{5}$$

where the conditional standard deviation $\sigma(q)$ is given by $\sigma^2(q) = \text{Var}_{\vec{\boldsymbol{o}} \sim \pi_k(\cdot|q)}[r^\star(q, \vec{\boldsymbol{o}}) \mid q]$. In practice, GRPO is typically trained with a cosine learning rate schedule, which can be locally treated as a constant step size α. Within our generic update rule, this corresponds to setting $\boldsymbol{w}_k = \boldsymbol{g}_{\text{GRPO}} / \|\boldsymbol{g}_{\text{GRPO}}\|_2$ and $\eta_k = \alpha \, \|\boldsymbol{g}_{\text{GRPO}}\|_2$, so that both the response length $|\vec{\boldsymbol{o}}|$ and reward variability $\sigma(q)$ directly influence the effective step size η_k.

Objective. Our goal in this work is to understand how the choice of update direction $\boldsymbol{w}_k$ and step size η_k influences the convergence of RLVR. In particular, we ask: under what conditions can we guarantee convergence, and what design choices may lead to instability or failure modes?

3 Trajectory-Level Analysis

In this section, we study the optimization scheme (3) on a single prompt q. We take a *trajectory-level* view, where each response $\vec{\boldsymbol{o}}$ is treated as a single unit rather than a sequence of tokens. By abstracting away the internal structure, the analysis becomes simpler yet still revealing. We begin by outlining the key ingredients of this trajectory-level view, then examine both its success modes and failure cases. Although this setup is only a warm-up for the more detailed token-level analysis, it already highlights several nontrivial and illuminating properties of RLVR.

3.1 Key Ingredients: Gradient Gap and Gap Alignment

Recall that the optimization objective is the *correction rate* of the model π_θ on prompt q: $J_q(\pi_\theta) := \mathbb{E}_{\vec{\boldsymbol{o}} \sim \pi_\theta(\cdot|q)} \big[r^\star(q, \vec{\boldsymbol{o}}) \;\big|\; q \big]$. To analyze this, we partition the response space $\mathcal{O}$ into two sets based on the verifiable reward $r^\star(q, \cdot)$:

$$\mathcal{O}_q^+ \;:=\; \big\{ \vec{\boldsymbol{o}} \in \mathcal{O} \;\big|\; r^\star(q, \vec{\boldsymbol{o}}) = 1 \big\} \qquad \text{and} \qquad \mathcal{O}_q^- \;:=\; \big\{ \vec{\boldsymbol{o}} \in \mathcal{O} \;\big|\; r^\star(q, \vec{\boldsymbol{o}}) = 0 \big\} \,, \tag{6}$$

Here $\mathcal{O}_q^+$ represents desirable responses (correct solutions), while $\mathcal{O}_q^-$ contains undesirable ones. Accordingly, $J_q(\pi_\theta) = \mathbb{P}_{\vec{\boldsymbol{o}} \sim \pi_\theta(\cdot|q)} \big[\vec{\boldsymbol{o}} \in \mathcal{O}_q^+ \big]$ and $1 - J_q(\pi_\theta) = \mathbb{P}_{\vec{\boldsymbol{o}} \sim \pi_\theta(\cdot|q)} \big[\vec{\boldsymbol{o}} \in \mathcal{O}_q^- \big]$.

Conditional Policies. We further define conditional distributions over the positive and negative spaces:

$$\pi_\theta^+(\vec{\boldsymbol{o}} \mid q) = \pi_\theta\big(\vec{\boldsymbol{o}} \;\big|\; q, \mathcal{O}_q^+\big) \;:=\; \frac{\pi_\theta(\vec{\boldsymbol{o}} \mid q)}{J_q(\pi_\theta)} \cdot \mathbb{1}\{\vec{\boldsymbol{o}} \in \mathcal{O}_q^+\}\,, \tag{7a}$$

$$\pi_\theta^-(\vec{\boldsymbol{o}} \mid q) = \pi_\theta\big(\vec{\boldsymbol{o}} \;\big|\; q, \mathcal{O}_q^-\big) \;:=\; \frac{\pi_\theta(\vec{\boldsymbol{o}} \mid q)}{1 - J_q(\pi_\theta)} \cdot \mathbb{1}\{\vec{\boldsymbol{o}} \in \mathcal{O}_q^-\}\,. \tag{7b}$$

These describe how the model π_θ distributes probability mass within the "good" and "bad" regions, respectively.

Gradient Gap: A Direction for Improvement. Using the conditional policies, we measure the expected log-probability gradient / score function in each region:

$$\boldsymbol{g}_q^+(\pi_\theta) := \mathbb{E}_{\vec{\boldsymbol{o}} \sim \pi_\theta^+(\cdot|q)}\big[\nabla_\theta \log \pi_\theta(\vec{\boldsymbol{o}} \mid q)\big] \quad \text{and} \quad \boldsymbol{g}_q^-(\pi_\theta) := \mathbb{E}_{\vec{\boldsymbol{o}} \sim \pi_\theta^-(\cdot|q)}\big[\nabla_\theta \log \pi_\theta(\vec{\boldsymbol{o}} \mid q)\big]. \tag{8}$$

The difference between them,

$$\boldsymbol{g}_q^+(\pi_\theta) - \boldsymbol{g}_q^-(\pi_\theta), \tag{9}$$

is the *Gradient Gap*. Intuitively, it highlights how the model's parameters should be shifted to favor desirable responses over undesirable ones.

Crucially, the Gradient Gap is directly proportional[1] to the policy gradient given in equation (4):

$$\nabla_\theta J_q(\pi_\theta) \;=\; J_q(\pi_\theta)\{1 - J_q(\pi_\theta)\} \cdot \big(\boldsymbol{g}_q^+ - \boldsymbol{g}_q^-\big)\,. \tag{10}$$

This shows that the Gradient Gap captures the true direction of improvement. Unlike the full policy gradient $\nabla_\theta J_q(\pi_\theta)$, $\boldsymbol{g}_q^+(\pi_\theta) - \boldsymbol{g}_q^-(\pi_\theta)$ is not scaled down by the variability factor $J_q(1-J_q)$, making it a purer indicator of where to move.

Gap Alignment: Following the Right Direction. Consider now the optimization scheme (3). At iteration k, define $\boldsymbol{g}_q^+(k)$ and $\boldsymbol{g}_q^-(k)$ under the current policy π_k. The update vector $\boldsymbol{w}_k$ should ideally align with the improvement direction $\boldsymbol{g}_q^+(k) - \boldsymbol{g}_q^-(k)$.

We measure this alignment by the inner product

$$\Delta\mu_q(k) \;:=\; \boldsymbol{w}_k \cdot \big\{\boldsymbol{g}_q^+(k) - \boldsymbol{g}_q^-(k)\big\}\,. \tag{11}$$

If $\|\boldsymbol{w}_t\|_2 = 1$, this equals $\Delta\mu_q(k) \;=\; \|\boldsymbol{g}_q^+(k) - \boldsymbol{g}_q^-(k)\|_2 \cdot \cos\angle\big\{\boldsymbol{w}_t, \boldsymbol{g}_q^+(k) - \boldsymbol{g}_q^-(k)\big\}$, which depends both on the magnitude of the Gradient Gap and the angle of alignment.

In the convergence analysis, $\Delta\mu_q(k)$ will play a central role. For stable progress we require:

(i) $\Delta\mu_q(k)$ should be positive and preferably large, ensuring updates move in the right direction.

(ii) The step size η_k should adapt to its scale, preventing over- or under-shooting.

3.2 Main Findings

We now turn to the central findings of our analysis. Proofs will be deferred to Appendices C and D. Before presenting the results, let us impose a mild regularity condition on the policy score function.

Assumption 1 (Regularity of Trajectory Policy Score)**.** *The policy score function* $\nabla_\theta \log \pi_\theta(\vec{\boldsymbol{o}} \mid q)$ *behaves regularly with respect to the parameters* θ*:*

(a) (Boundedness) *There exists a constant* $G_{\mathrm{o}} < \infty$ *such that for all* θ *and* $(q, \vec{\boldsymbol{o}})$*,*

$$\big\|\nabla_\theta \log \pi_\theta(\vec{\boldsymbol{o}} \mid q)\big\|_2 \;\le\; G_{\mathrm{o}}. \tag{12}$$

(b) (Smoothness) *The policy score function is* L_{o}*-Lipschitz continuous with respect to* θ*:*

$$\big\|\nabla_\theta \log \pi_{\theta'}(\vec{\boldsymbol{o}} \mid q) - \nabla_\theta \log \pi_\theta(\vec{\boldsymbol{o}} \mid q)\big\|_2 \;\le\; L_{\mathrm{o}} \cdot \|\theta' - \theta\|_2. \tag{13}$$

Throughout this section, we use the shorthand $J_q(k) \;=\; J_q(\pi_k)$ to denote the performance at iteration k.

[1] A formal proof of this is found in Appendix B.2

3.2.1 Convergence and Stagnation

Armed with this set-up, we now state our main theorem, which distinguishes between two possible outcomes of learning: *successful convergence* to the optimum, or *stagnation* at a suboptimal performance plateau. The distinction hinges on how well the update directions align with the underlying objective. To formalize this, we introduce the notion of *Cumulative Gap Alignment*,

$$M(K) \;:=\; \textstyle\sum_{k=0}^{K-1} [\Delta\mu_q(k)]_+ \, \eta_k, \tag{14}$$

which accumulates the amount of "useful progress" made up to horizon K. Intuitively, $M(K)$ grows whenever the update direction is positively aligned with the true objective, and it stagnates when the updates fail to exploit the available signal.

Theorem 1 (Convergence and Stagnation)**.** *Assume that the step sizes satisfy* $\eta_k \le \frac{1}{2\sqrt{L_o}}$.

(a) ***(Stagnation)*** *Consider when* $J_q(0) < 1$. *If the alignment signal is too weak, in the sense that the cumulative alignment remains bounded* $M(K) \le C_0$ *and* $\sum_{k=0}^{\infty} \eta_k^2 \le C_0'/(L_o + 8G_o^2)$, *for some constants* $0 \le C_0, C_0' < \infty$, *then learning will stall. In this case, the performance remains strictly sub-optimal:* $J_q(k) \le J_q(0)\big(J_q(0) + \exp(C_0 + C_0')\,\{1 - J_q(0)\}\big)^{-1} < 1$.

(b) ***(Convergence)*** *Consider a case where* $J_q(0) > 0$. *Suppose the step size* η_k *is adapted to the strength of the alignment signal,*

$$\eta_k \;\le\; \frac{[\Delta\mu_q(k)]_+}{2\,(L_o + 8\,G_o^2)} \quad \textit{where } [\,\cdot\,]_+ = \max(0, \cdot). \tag{15a}$$

Then the performance is lower-bounded at any horizon K *by*

$$J_q(K) \;\ge\; \frac{J_q(0)}{J_q(0) + \{1 - J_q(0)\}\,\exp\big\{-\tfrac{1}{2}M(K)\big\}}. \tag{15b}$$

Moreover, if the alignment accumulates indefinitely, $\lim_{K\to\infty} M(K) = +\infty$, *then the policy is guaranteed to achieve perfect performance:* $\lim_{K\to\infty} J_q(K) = 1$.

The theorem establishes a clear dichotomy. Convergence is attainable only when update directions exhibit consistent alignment with the underlying objective and the step size is properly scaled to reflect this signal. In the absence of either alignment or adaptive scaling, progress stagnates and the policy remains confined to a suboptimal regime.

Sketch of Proof. The key step is the inequality

$$\left|\log\Big(\frac{J_q(k+1)}{1 - J_q(k+1)}\Big) - \log\Big(\frac{J_q(k)}{1 - J_q(k)}\Big) - \Delta\mu_q(k)\,\eta_k\right| \;\le\; (L_o + 8\,G_o^2)\,\eta_k^2\,, \tag{16}$$

which is stated formally in Lemma 1 of Appendix C.1.1. This inequality shows that $\Delta\mu_q(t)\,\eta_t$ captures the first-order Taylor approximation of the change in log-odds of J_q. Summing (16) over iterations and analyzing the resulting terms under different cases reveals that the Cumulative Gap Alignment $M(K)$ governs the value of J_q. This establishes the claims in Theorem 1.

3.2.2 The Importance of Properly Chosen Step Size η_k

According to condition (15a) in Theorem 1(b), the step size η_k must be carefully scaled to match the gap alignment $\Delta\mu_q(k)$. To illustrate this, we contrast two scenarios: a modest step size yields linear convergence, whereas an overly aggressive one causes failure.

Linear Convergence Under Proper Scaling. Suppose that every update direction provides a consistent signal, so that the Gap Alignment $\Delta\mu_q(k)$ is uniformly bounded below. In this case, a properly chosen fixed step size is sufficient to guarantee rapid improvement.

Corollary 1 (Linear Convergence with a Uniform Gap)**.** *If every update direction* $\boldsymbol{w}_t$ *provides a uniform gap,* $\Delta\mu_q(k) \ge \Delta\mu_q > 0$ *for all* $k \ge 0$, *then a simple fixed step size* η *satisfying*

$$\eta \;\le\; \min\left\{\frac{\Delta\mu_q}{2\,(L_o + 8\,G_o^2)}, \frac{1}{2\sqrt{L_o}}\right\},$$

drives the error to zero at a linear rate: $1 - J_q(K) \le \frac{1 - J_q(0)}{J_q(0)}\,\exp\big\{-\tfrac{1}{2}\,\Delta\mu_q\,\eta\cdot K\big\}$.

The Perils of Overshooting. The picture changes sharply when the step size is too large. If condition (15a) in Theorem 1(b) is violated, convergence may break down entirely. The next result shows that even with perfect update directions, learning can collapse under overly aggressive step sizes.

Theorem 2 (Catastrophic Failure from an Overly Large Step Size)**.** *There exists a problem instance under Assumption 1 with $G_{\mathrm{o}} \geq \sqrt{L_{\mathrm{o}}}$ where the Gap Alignment is uniformly positive, $\Delta\mu_q(k) \geq \Delta\mu_q > 0$ for all $k \geq 0$, yet using an overly large constant step size $\eta_k = \eta$ leads to failure. Specifically, if the step size satisfies*

$$\frac{60\,\Delta\mu_q}{L_{\mathrm{o}} + G_{\mathrm{o}}^2} \;\leq\; \eta \;\leq\; \frac{1}{2\sqrt{L_{\mathrm{o}} + G_{\mathrm{o}}^2}}\,,$$

where $0 < \Delta\mu_q \leq \frac{1}{120}\sqrt{L_{\mathrm{o}} + G_{\mathrm{o}}^2}$, the policy's performance will strictly **decrease** *at every step, ultimately converging to zero: $J_q(k) < J_q(k-1)$ and $\lim_{K\to\infty} J_q(k) = 0$.*

While the numerical constants (e.g., 60, 120) are not sharp, the phenomenon is robust: an oversized step size causes repeated overshooting, pushing the system toward collapse rather than improvement.

Intuition for the lower bound analysis. Our convergence analysis (Theorem 1) relies on equation (16), which uses a first-order approximation of the change in log-odds. For the lower bound, however, it is crucial to examine the second-order expansion. To this end, we define conditional variances over the positive (and negative) response space: $\mathrm{Var}^{+} := \mathrm{Var}_{\vec{\boldsymbol{o}}\sim\pi_k(\cdot|q,\mathcal{O}_q^{+})}\big[\boldsymbol{w}_k \cdot \nabla_\theta \log \pi_k(\vec{\boldsymbol{o}} \mid q)\big]$. The term Var^{-} is defined analogously. The second-order Taylor expansion gives

$$\log\Big(\frac{J_q(k+1)}{1-J_q(k+1)}\Big) - \log\Big(\frac{J_q(k)}{1-J_q(k)}\Big) \;=\; \Delta\mu_q(k)\,\eta_k + \{\mathrm{Var}^{+} - \mathrm{Var}^{-}\}\cdot\eta_k^2 + \mathcal{O}(\eta_k^3)\,. \quad (17)$$

In our construction, the linear term is always favorable:$\Delta\mu_q(k)\,\eta_k > 0$. The challenge comes from the quadratic term. If the variance over the negative space dominates, $\mathrm{Var}^{-} > \mathrm{Var}^{+}$, then for moderately large step sizes the second-order effect can overwhelm the first-order gain, pulling the log-odds downward and decreasing J_q.

This phenomenon is not just a theoretical artifact—it is highly plausible in practice. Real-world language models typically face an enormous negative space (many incorrect responses) with high variability, leading to large Var^{-}. In contrast, the positive space often contains only a few consistent modes, keeping Var^{+} relatively small. This imbalance highlights the danger of overshooting: unless the step size η_k is carefully calibrated, the variance contribution from the negative space can dominate and derail learning. To ensure both stability and progress, the step size must respect the scale $\eta_k \asymp \Delta\mu_q(k)/(L_{\mathrm{o}} + G_{\mathrm{o}}^2)$.

4 Token-Level Analysis

We now move towards a token-level analysis of RLVR, which sharpens the trajectory-level perspective developed earlier. While natural and general for abstract analysis, our analysis in Section 3 overlooks the autoregressive structure of LLMs: responses are generated token by token, with intermediate Chain-of-Thought (CoT) steps shaping the learning dynamics.

At the trajectory level, the regularity conditions in Assumption 1 are imposed on the policy score $\nabla_\theta \log \pi_\theta(\vec{\boldsymbol{o}} \mid q)$ of the entire response $\vec{\boldsymbol{o}}$. However, the score can be decomposed into token-wise contributions: $\nabla_\theta \log \pi_\theta(\vec{\boldsymbol{o}} \mid q) = \sum_{t=1}^{|\vec{\boldsymbol{o}}|} \nabla_\theta \log \pi_\theta(o_t \mid q, \vec{\boldsymbol{o}}_{<t})$, where every token o_t requires a forward pass from the language model and thus carries its own regularity properties. This makes it more natural—and ultimately more powerful—to impose assumptions at the token level. Doing so introduces response length as an explicit factor, which will be central to our analysis. Interestingly, as we will see, it also reveals how the training dynamics adapt to task difficulty under the current policy.

We refine Assumption 1 into the following token-level version.

Assumption 2 (Regularity of Token Policy Score)**.** *There exist $G_{\mathrm{p}}, L_{\mathrm{p}} \in (0, +\infty)$ such that*

$$\big\|\nabla_\theta \log \pi_\theta(o_t \mid q, \vec{\boldsymbol{o}}_{<t})\big\|_2 \;\leq\; G_{\mathrm{p}} < \infty \quad \textit{for all } \theta\textit{, question } q\textit{, response prefix } \vec{\boldsymbol{o}}_{<t} \textit{ and token } o_t,$$
$$\big\|\nabla_\theta \log \pi_{\theta'}(o_t \mid q, \vec{\boldsymbol{o}}_{<t}) - \nabla_\theta \log \pi_\theta(o_t \mid q, \vec{\boldsymbol{o}}_{<t})\big\|_2 \;\leq\; L_{\mathrm{p}} \cdot \|\theta' - \theta\|_2\,.$$

In addition, we propose a second key assumption concerning the distribution of response length.

Assumption 3 (Sub-Exponential Response Length)**.** *There exist constants $T_\infty, T_{\psi_1} \in (0, +\infty)$ such that for every question q and every policy π_θ, if $\vec{\boldsymbol{o}} \sim \pi_\theta(\cdot \mid q)$ and $\ell := |\vec{\boldsymbol{o}}|$ denotes the response length, then $1 \le \ell \le T_\infty$ almost surely, and $\|\ell\|_{\psi_1} \le T_{\psi_1}$.*[2]

Assumption 3 characterizes response length: T_∞ bounds the worst case, while T_{ψ_1} reflects the typical scale. It holds $\mathbb{E}_{\pi_\theta}[|\vec{\boldsymbol{o}}| \mid q] \le T_{\psi_1} \le T_\infty / \log 2$, so that T_{ψ_1} may be much smaller than T_∞.

With these two assumptions in place, we are ready to present our token-level convergence guarantee. The statement parallels the trajectory-level result, but now incorporates the finer granularity of token-wise dynamics. We retain the key quantities from Section 3.1, namely the Gap Alignment $\Delta\mu_q(k)$ from equation (11), and the Cumulative Gap Alignment $M(K)$ from equation (14).

Theorem 3 (Convergence at the Token-Level)**.** *Assume $J_q(0) > 0$. If the step size η_k is scaled to the strength of the alignment signal,*

$$\eta_k \;\le\; \min\left\{ \frac{[\Delta\mu_q(k)]_+ \,/\, 2}{L_{\mathrm{p}}\, T_\infty + G_{\mathrm{p}}^2 \,\min\left\{ \frac{T_{\psi_1}}{1 - J_q(k)}, 8\, T_\infty^2 \right\}},\; \frac{1}{2\sqrt{L_{\mathrm{p}}\, T_\infty + G_{\mathrm{p}}^2\, T_{\psi_1}}} \right\}, \tag{18a}$$

then the performance is guaranteed at any horizon K by

$$J_q(K) \;\ge\; \frac{J_q(0)}{J_q(0) + \{1 - J_q(0)\}\, \exp\left\{ -\tfrac{1}{2} M(K) \right\}}. \tag{18b}$$

This result closely mirrors the trajectory-level guarantee but introduces several new elements. The response length parameters T_∞ and T_{ψ_1} now play a direct role, reflecting the cost of token-level granularity. In addition, the factor $(1 - J_q)$ emerges in the step-size condition, linking stability to the current performance level of the policy. In a later discussion, we will examine the implications of condition (18a), with particular attention to how step-size choices manifest in practical algorithms such as GRPO and Dr. GRPO.

Complementing the positive result in Theorem 3, we now show that the step-size scalings with T_∞ and T_{ψ_1} are essentially tight, as established by the token-level analogue of Theorem 2 below.

Theorem 4 (Catastrophic Failure from an Overly Large Step Size at the Token Level)**.** *There exists a problem instance satisfying Assumption 2 with $G_{\mathrm{p}} \ge \sqrt{L_{\mathrm{p}}}$ where the Alignment Gap is always positive, $\Delta\mu_q(k) \ge \Delta\mu_q > 0$, yet choosing a constant step size $\eta_k = \eta$ that is too large leads to a complete failure of learning. Specifically, if the step size satisfies*

$$\frac{120\, \Delta\mu_q}{(L_{\mathrm{p}} + G_{\mathrm{p}}^2)\, T_\infty} \;\le\; \eta \;\le\; \frac{1}{2\sqrt{(L_{\mathrm{p}} + G_{\mathrm{p}}^2)\, T_\infty}}, \tag{19}$$

where $0 < \Delta\mu_q \le \frac{1}{240}\sqrt{(L_{\mathrm{p}} + G_{\mathrm{p}}^2)\, T_\infty}$, the policy's performance will strictly **decrease** *at every step, ultimately converging to zero: $J_q(k) < J_q(k-1)$ and $\lim_{K\to\infty} J_q(K) = 0$.*

This lower bound confirms that the step-size condition (18a) reflects an intrinsic barrier. Indeed, by treating $(1 - J_q(k))$ as constant and applying the crude bound $T_{\psi_1} \lesssim T_\infty$, the upper limit in (18a) reduces to $\eta_k \lesssim [\Delta\mu_q(k)]_+ / \{(L_{\mathrm{p}} + G_{\mathrm{p}}^2)\, T_\infty\}$, which matches the overshooting threshold in (19) up to constants. This alignment verifies the sharp dependence on response length in step-size selection.

Finally, note that the $(1 - J_q(k))$ factor only influences how fast convergence proceeds toward 1. In the lower bound construction of Theorem 4, $J_q(k)$ is strictly decreasing, so this term behaves like a constant and does not alter the failure guarantee. Hence, it affects the upper bound but not the lower bound.

Implications in GRPO and Dr. GRPO. We next examine how the update rules of GRPO and Dr. GRPO (or REINFORCE) fit into our token-level framework. For clarity, we restrict attention to the scaling behavior with respect to $\Delta\mu_q$, T_{ψ_1}, J_q, and $(1 - J_q)$ under a single prompt q.

In this regime, the GRPO gradient from equation (5) simplifies to

$$\boldsymbol{g}_{\mathrm{GRPO}}(\pi_k) \;\asymp\; \mathbb{E}_{\vec{\boldsymbol{o}} \sim \pi_k(\cdot|q)}\left[A(q, \vec{\boldsymbol{o}}) \cdot \nabla_\theta \log \pi_k(\vec{\boldsymbol{o}} \mid q) \right] / \left\{ T_{\psi_1} \sqrt{J_q(1 - J_q)} \right\}.$$

[2] For a random variable X, the ψ_1-Orlicz norm is $\|X\|_{\psi_1} := \inf\left\{ a > 0 \,:\, \mathbb{E}\left[\exp(|X|/a)\right] \le 2 \right\}$. Finiteness of $\|X\|_{\psi_1}$ is equivalent to X being sub-exponential.

The Dr. GRPO (or REINFORCE) gradient takes the form (4). Applying identity (10) gives

$$\boldsymbol{g}_{\mathrm{GRPO}} \asymp T_{\psi_1}^{-1}\sqrt{J_q(1-J_q)}\cdot\left(\boldsymbol{g}_q^+ - \boldsymbol{g}_q^-\right) \quad \text{and} \quad \boldsymbol{g}_{\mathrm{Dr.\,GRPO}} \asymp J_q(1-J_q)\cdot\left(\boldsymbol{g}_q^+ - \boldsymbol{g}_q^-\right).$$

An update step $\theta_{k+1} = \theta_k + \alpha \cdot \boldsymbol{g}(\pi_k)$ for $\boldsymbol{g} = \boldsymbol{g}_{\mathrm{GRPO}}$ or $\boldsymbol{g}_{\mathrm{Dr.\,GRPO}}$ can therefore be interpreted as moving in the direction $\boldsymbol{w}_k = (\boldsymbol{g}_q^+ - \boldsymbol{g}_q^-)/\|\boldsymbol{g}_q^+ - \boldsymbol{g}_q^-\|_2$, with alignment magnitude $\Delta\mu_q = \|\boldsymbol{g}_q^+ - \boldsymbol{g}_q^-\|_2$, and effective learning rates

$$\text{(GRPO)}\ \ \eta_k \asymp \Delta\mu_q \cdot T_{\psi_1}^{-1}\sqrt{J_q(1-J_q)} \quad \text{and} \quad \text{(Dr. GRPO)}\ \ \eta_k \asymp \Delta\mu_q \cdot J_q(1-J_q)\,. \tag{20}$$

On the other hand, condition (18a) in Theorem 3, under the simplification $L_{\mathrm{p}} \ll G_{\mathrm{p}}^2$ and retaining the $T_{\psi_1}/(1-J_q)$ term in the denominator, reduces to

$$\text{(Theorem 3, condition (18a))} \qquad \eta_k \lesssim \Delta\mu_q \cdot T_{\psi_1}^{-1}(1-J_q)\,. \tag{21}$$

Comparing equations (20) and (21) leads to several insights:

Gradient gap. Both GRPO and Dr. GRPO scale proportionally with the gap alignment $\Delta\mu_q$, consistent with the theoretical condition.

Sequence length. GRPO exhibits the correct $1/T_{\psi_1}$ scaling, aligning with the theory, offering an explanation for why length normalization empirically stabilizes training. In contrast, Dr. GRPO lacks this normalization.

Correction rate. After variance normalization, GRPO overshoots as $J_q \to 1$. We hypothesize that this may explain the observed stagnation of training at a correction rate strictly below 1.

Sketch of Proof for Theorem 3. The proof builds on the following refined token-level inequality:

$$\log\left(\frac{J_q(k+1)}{1-J_q(k+1)}\right) - \log\left(\frac{J_q(k)}{1-J_q(k)}\right) \;\geq\; \Delta\mu_q(k)\cdot\eta_k - \left(L_{\mathrm{p}}\,T_\infty + \frac{G_{\mathrm{p}}^2\,T_{\psi_1}}{1-J_q(k)}\right)\cdot\eta_k^2\,. \tag{22}$$

The formal statement of bound (22) is provided in Lemma 2 in Appendix C.2. In parallel, we adapt the trajectory-level result (16) to the token setting by taking $G_{\mathrm{o}} = G_{\mathrm{p}}T_\infty$ and $L_{\mathrm{o}} = L_{\mathrm{p}}T_\infty$. We then combine these two bounds, applying whichever is tighter in a given regime. The remaining steps follow the same structure as in Theorem 1(b).

The main technical challenge lies in proving inequality (22). The difficulty is that the Gradient Gap $\boldsymbol{g}_q^+ - \boldsymbol{g}_q^-$ is not a martingale, since it depends on the conditional distributions π_θ^+ and π_θ^-. To address this, we relate the log moment generating functions of the conditional score functions to those of the unconditional scores, which do form martingales. This step is crucial: it yields the sharp linear dependence on T_{ψ_1} in the η_k^2 term of (22). Without this refinement, a naive trajectory-level analysis would give only the weaker quadratic dependence $G_{\mathrm{p}}^2\,T_\infty^2$.

5 Numerical Experiments

5.1 REINFORCE on Contextual Bandits

We consider a contextual variant of the synthetic bandit experiment of Arnal et al. (2025, Section 5.1). contextual bandit with contexts $x \in [0,1]^d$ for $d = 10$. For a set of $N := 100$ arms, we generate linear scores for each context x, $\mathbf{s}(x) = \boldsymbol{\beta}^\top x \in \mathbb{R}^N$ for a matrix $\boldsymbol{\beta} \in \mathbb{R}^{d\times N}$, with standard normal entries. $r_y(x) := \arg\max_{y\in[N]} \mathbf{s}(x)$. We use linear logits initialized as $\ell_0(x) := \theta_0^\top x \in \mathbb{R}^N$, for $\theta_0 \sim \mathcal{N}(0, 0.01^2 \cdot \mathrm{Id}_{d\times d})$. The policy π_{θ_k} was then initialized as a softmax over ℓ_0 and the parameters θ_k were updated according to the REINFORCE exact gradient update at a training context x_k with stepsize $\eta := 0.1$:

$$\theta_{k+1} = \theta_k + \eta\ \mathbb{E}_{y\sim\pi_{\theta_k}(\cdot|x_k)}[(r_y(x_k) - J_{x_k}(\pi_{\theta_k}))\cdot\nabla_\theta\ \log\pi_{\theta_k}(y \mid x_k)]\,.$$

The training context x_k was selected at random among those (from an initial pool of 100 contexts drawn uniformly from $[0,1]^d$) with intermediate value function $J(x_k) \in [0.2, 0.8]$, following intuitions from curriculum learning for filtering out overly difficult or easy prompts (Zhang et al., 2025).

We construct three plots based on calculating the following for 500 randomly evaluated contexts x: the value function $J_x(\pi_{\theta_k})$, per-context cumulative gradient gap $\sum_{i=0}^k [\Delta\mu_x(i)]_+ \cdot \eta$, and the relative per-context cumulative gradient gap $\sum_{i=0}^k ([\Delta\mu_x(i)]_+ - [\Delta\mu_{x_i}(i)]_+)\cdot\eta$ which measures the discrepancy of the gradient gaps at the training contexts x_k.

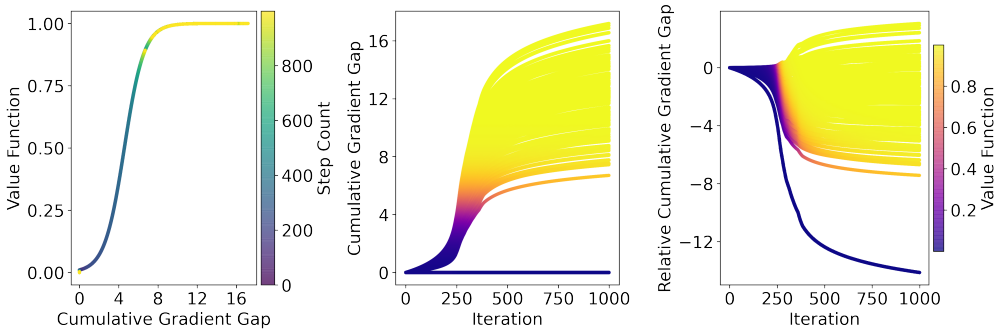

Figure 1: Contextual Bandit Experiments.

From the first subplot, we observe a distinctive logistic relationship between the cumulative gradient gap and the value function, reminiscent of our theory (Corollary 1).

From the second subplot, we see there are two regimes for each context's cumulative gradient gap curve: either fast exponential convergence (Corollary 1) or lack of improvement (Theorem 2).

In the third plot, we interestingly see that those contexts with close to 0 relative cumulative gradient gap (i.e., close to that of training contexts) experience faster convergence.

5.2 GRPO on Language Models

We validate our theory on three GRPO training runs for language model math reasoning: (1) Qwen2.5-7B on the GSM8k dataset (Cobbe et al., 2021) and (2) Qwen2.5-Math-7B on the DAPO-17k dataset (Yu et al., 2025a). For background, the GSM8k dataset consists of grade-school math word problems, while the more challenging DAPO-17k dataset consists of problems derived from past AIME and AMC competitions.

At each training step, we approximate the batch-average gradient gap magnitude $\mathbb{E}_q[\Delta\mu_q]$ using the relation $\boldsymbol{g}_{\text{GRPO}} \propto \sqrt{J_q(1-J_q)} \cdot \left(\boldsymbol{g}_q^+ - \boldsymbol{g}_q^-\right)$, as derived in Section 4. In Figure 2, we plot the cumulative gradient gap vs. the value function, colored by normalized step count. For all three datasets, we see a similar relationship between cumulative gradient gap and accuracy as in our theory and bandit experiment.

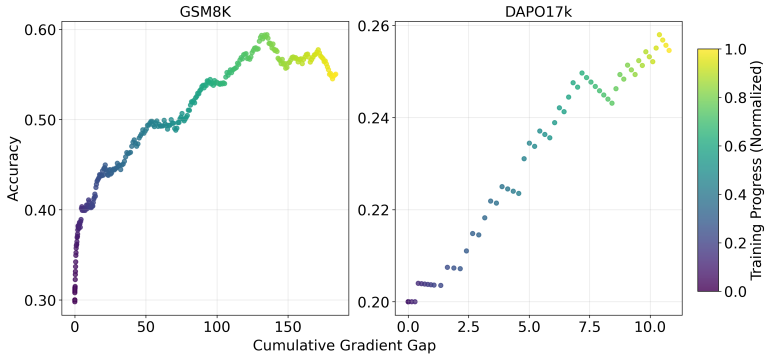

Figure 2: Cumulative Gradient Gap vs. Validated Accuracy for our experiments.

6 Discussion and Future Directions

Our analysis is restricted to the single-prompt setting, which enabled sharp characterizations of Gradient Gap alignment and step size scaling. In practice, however, training involves a diverse batch of prompts. In this regime, both the alignment signal $\Delta\mu_q(k)$ and the optimal step size η_k can vary substantially across prompts. A single update direction $\boldsymbol{w}_k$ may align well with some prompts but poorly with others, and a step size that is safe for one subset may be overly aggressive for another, leading to overshooting and limited overall gains.

These observations suggest several directions for future work: developing prompt-adaptive updates that adjust direction or scale based on batch heterogeneity, analyzing the statistical dynamics of RLVR under diverse prompt distributions, and extending the framework to sequential or curriculum-based training (Bengio et al., 2009; Chen et al., 2025; Zhang et al., 2025). Such extensions are essential for a full theory of RLVR in realistic multi-prompt settings.

REFERENCES

Alekh Agarwal, Sham M Kakade, Jason D Lee, and Gaurav Mahajan. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *Journal of Machine Learning Research*, 22(98):1–76, 2021.

Charles Arnal, Gaëtan Narozniak, Vivien Cabannes, Yunhao Tang, Julia Kempe, and Remi Munos. Asymmetric reinforce for off-policy reinforcement learning: Balancing positive and negative rewards. *arXiv preprint arXiv:2506.20520*, 2025.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th International Conference on Machine Learning (ICML '09)*, pages 41–48, New York, NY, USA, 2009. Association for Computing Machinery. ISBN 978-1-60558-516-1. doi: 10.1145/1553374.1553380. URL https://doi.org/10.1145/1553374.1553380.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford Series in Probability and Statistics. Oxford University Press, 2013. ISBN 9780199535255. doi: 10.1093/acprof:oso/9780199535255.001.0001.

Kianté Brantley, Mingyu Chen, Zhaolin Gao, Jason D. Lee, Wen Sun, Wenhao Zhan, and Xuezhou Zhang. Accelerating rl for llm reasoning with optimal advantage regression. *arXiv preprint arXiv:2505.20686*, 2025. URL https://arxiv.org/abs/2505.20686.

F. Chen. Outcome-based online reinforcement learning with general function approximation. *arXiv preprint arXiv:2505.20268*, 2025. URL https://arxiv.org/abs/2505.20268.

Xiaoyin Chen, Jiarui Lu, Minsu Kim, Dinghuai Zhang, Jian Tang, Alexandre Piché, Nicolas Gontier, Yoshua Bengio, and Ehsan Kamalloo. Self-evolving curriculum for llm reasoning. *CoRR*, abs/2505.14970, May 2025. doi: 10.48550/arXiv.2505.14970. URL https://arxiv.org/abs/2505.14970.

Xiaoyu Chen, Han Zhong, Zhuoran Yang, Zhaoran Wang, and Liwei Wang. Human-in-the-loop: Provably efficient preference-based reinforcement learning with general function approximation. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, volume 162, pages 3773–3793. PMLR, 2022. URL https://proceedings.mlr.press/v162/chen22ag.html.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Yihan Du, Anna Winnicki, Gal Dalal, Shie Mannor, and R. Srikant. Exploration-driven policy optimization in rlhf: Theoretical insights on efficient data utilization. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pages 11830–11887. PMLR, 2024. URL https://proceedings.mlr.press/v235/du24i.html.

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.

Zhiyuan He, Xufang Luo, Yike Zhang, Yuqing Yang, and Lili Qiu. ΔL normalization: Rethink loss aggregation in RLVR. *arXiv preprint arXiv:2509.07558*, 2025.

Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding r1-zero-like training: A critical perspective. *arXiv preprint arXiv:2503.20783*, 2025.

Long Ouyang, Jeff 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 Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. URL https://arxiv.org/abs/2203.02155. Also available as arXiv preprint arXiv:2203.02155.

Aldo Pacchiano, Aadirupa Saha, and Jonathan Lee. Dueling RL: Reinforcement learning with trajectory preferences. *arXiv preprint arXiv:2111.04850*, 2021. URL https://arxiv.org/abs/2111.04850.

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

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Yuda Song, Julia Kempe, and Remi Munos. Outcome-based exploration for llm reasoning. *arXiv preprint arXiv:2509.06941*, 2025.

Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, et al. Kimi k1. 5: Scaling reinforcement learning with llms. *arXiv preprint arXiv:2501.12599*, 2025.

Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, and Shengyi Huang. Trl: Transformer reinforcement learning. https://github.com/huggingface/trl, 2020.

Huaijie Wang, Shibo Hao, Hanze Dong, Shenao Zhang, Yilin Bao, Ziran Yang, and Yi Wu. Offline reinforcement learning for llm multi-step reasoning. *arXiv preprint arXiv:2412.16145*, 2024. URL https://arxiv.org/abs/2412.16145.

Yuanhao Wang, Qinghua Liu, and Chi Jin. Is RLHF more difficult than standard RL? a theoretical perspective. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2023. URL https://arxiv.org/abs/2306.14111. Also available as arXiv preprint arXiv:2306.14111.

Yue Frank Wu, Weitong Zhang, Pan Xu, and Quanquan Gu. A finite-time analysis of two time-scale actor-critic methods. *Advances in Neural Information Processing Systems*, 33:17617–17628, 2020.

Lin Xiao. On the convergence rates of policy gradient methods. *Journal of Machine Learning Research*, 23(282):1–36, 2022.

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

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, et al. DAPO: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025b.

Andrea Zanette, Martin J Wainwright, and Emma Brunskill. Provable benefits of actor-critic methods for offline reinforcement learning. *Advances in neural information processing systems*, 34:13626–13640, 2021.

Ruiqi Zhang, Daman Arora, Song Mei, and Andrea Zanette. Speed-rl: Faster training of reasoning models via online curriculum learning, 2025. URL https://arxiv.org/abs/2506.09016.

Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, et al. Group sequence policy optimization. *arXiv preprint arXiv:2507.18071*, 2025.

Banghua Zhu, Jiantao Jiao, and Michael I. Jordan. Principled reinforcement learning with human feedback from pairwise or k-wise comparisons. *arXiv preprint arXiv:2301.11270*, 2023. URL https://arxiv.org/abs/2301.11270.

Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul F. Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019. URL https://arxiv.org/abs/1909.08593.