BEYOND OUTCOME REWARDS: VERIFIABLE INFORMATION GAIN FOR REASONING REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement Learning with Verifiable Rewards (RLVR) improves large language model reasoning, but its outcome-level rewards provide limited supervision for intermediate reasoning steps. Recent self-distillation methods introduce teacher guidance, yet densely supervising entire teacher trajectories may inject redundant signals and constrain policy exploration. We propose **VIRAL**, a **V**erifiable **I**nformation-guided **R**easoning-level **A**dvantage **L**earning framework that selectively supervises critical reasoning patterns. VIRAL constructs a privileged teacher conditioned on the ground-truth solution and the student's response, identifies informative teacher patterns via ground-truth information gain and influential student patterns via uncertainty reduction, and establishes semantic correspondence between their reasoning trajectories. For the resulting critical patterns, VIRAL measures local teacher–student discrepancy and incorporates it into the RLVR advantage, while leaving the remaining trajectory governed by the original outcome reward. Experiments across three LLMs and one VLM on eleven reasoning benchmarks demonstrate consistent improvements over strong RLVR and self-distillation baselines.

1 INTRODUCTION

Reinforcement Learning with Verifiable Rewards (RLVR) (Shao et al., 2024; Lambert et al., 2024; Guo et al., 2025) has emerged as a powerful post-training paradigm for improving the complex reasoning capabilities of large language models (LLMs) across diverse domains (Xue et al., 2025; Liu et al., 2025; Dou et al., 2025; Dong et al., 2026). By optimizing model outputs with automatically verifiable correctness signals, RLVR effectively enhances reasoning performance. However, existing methods typically rely on coarse outcome-level rewards that assign largely uniform supervision across tokens, resulting in sparse and poorly localized credit signals that fail to identify the intermediate reasoning steps most responsible for the final outcome (Wang et al., 2025; 2026; He et al., 2026a). In a reasoning trajectory, many tokens serve merely explanatory or transitional roles, while only a small subset captures critical decisions, intermediate conclusions, or error-prone branching points (Wang et al., 2025; 2026). Assigning the same outcome-level reward to all tokens therefore dilutes the learning signal for these decisive steps, causing RLVR to underemphasize the most consequential reasoning components and limiting further performance gains (He et al., 2026b).

Recent studies use entropy-based signals (Chen et al., 2025; Wang et al., 2025) or special functional tokens (Wang et al., 2026) to reweight advantages or restrict policy-gradient updates. However, these signals are *heuristic* rather than *verifiable*: they may capture uncertainty or reasoning structure, but do not directly quantify how an intermediate reasoning component contributes to final correctness. Another line of work explores on-policy distillation (OPD) (Agarwal et al., 2024), where a stronger teacher supervises the student on its own trajectories. To remove the need for an external teacher, on-policy self-distillation (OPSD) (Zhao et al., 2026a) lets the same model act as both student and teacher under different contexts, with the teacher accessing privileged information such as the ground-truth solution. Yet directly distilling privileged teacher signals may over-regularize the student toward teacher-preferred trajectories, limiting policy exploration and potentially harming generalization (Zhao et al., 2026b; Pan et al., 2026). Recent methods therefore combine RLVR with self-distillation (Pan et al., 2026; Yang et al., 2026), using teacher–student discrepancies to

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

modulate token-level advantages. However, privileged information may improve the teacher trajectory globally without making every intermediate component equally informative. Teacher-derived signals remain model-dependent, and densely applying them throughout the trajectory can introduce substantial redundancy and noise. As shown in Fig. 1 and Fig. 3a, even when the teacher accesses the ground-truth solution and ultimately produces a correct response, many intermediate tokens contribute little to final correctness. From the perspective of token-level information gain, most tokens yield near-zero gains, while a considerable fraction even exhibit negative gains. This reveals an important distinction: privileged information can improve the teacher trajectory globally, yet does not make every local reasoning component informative for student supervision. This raises a key question: *how can we extract the teacher reasoning components that are most informative for final correctness and selectively translate them into fine-grained supervision without constraining policy exploration?*

Therefore, effectively supervising critical reasoning components with fine-grained learning signals throughout intermediate reasoning has become a key challenge in RLVR (Wang et al., 2025; 2026; He et al., 2026a; Chen et al., 2025). Based on the above analysis, a new RLVR-based optimization paradigm is needed to provide fine-grained supervision for critical intermediate reasoning processes while preserving the model's capacity for exploration. Ground-truth solutions and final answers contain rich offline verifiable information that can reveal which intermediate reasoning components are most relevant to final correctness (Hendrycks et al., 2021; Lightman et al., 2024). This motivates a central requirement: *how can we extract the truly informative components from such offline trajectories and transform them into verifiable, fine-grained learning signals for RLVR optimization?*

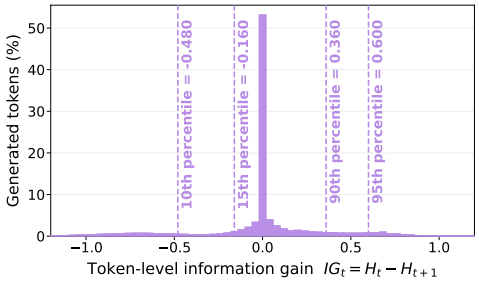


Figure 1: We sample 100 DeepMath (He et al., 2026c) problems and use Qwen3-8B as the teacher model conditioned on the ground-truth solution s, with all generated responses reaching the correct final answer. For each token o_i, we compute $\mathrm{IG}(o_i) = H(q, o_{<i}, s) - H(q, o_{\leq i}, s)$.

Motivated by this need, we propose **V**erifiable **I**nformation-guided **R**easoning-level **A**dvantage **L**earning (VIRAL), an RLVR framework for selective intermediate supervision. VIRAL instantiates the same model as a student and a privileged teacher: the student independently explores from the problem, while the teacher additionally observes the ground-truth solution and student response to produce a corrected, reasoning-aligned trajectory. Rather than supervising the entire teacher trajectory, VIRAL identifies correctness-relevant teacher patterns via ground-truth information gain and influential student patterns via uncertainty reduction. It further introduces **R**easoning **A**lignment-guided **A**dvantage **O**ptimization (RAAO), which aligns teacher and student patterns, measures their local distributional discrepancy, and selectively adjusts the RLVR advantage for critical patterns while leaving the remaining trajectory governed by the outcome reward. In this way, VIRAL transfers ground-truth-informed supervision to critical reasoning decisions without enforcing full-trajectory imitation, thereby preserving RLVR exploration.

Our contributions are threefold. First, we reveal that even ground-truth-informed teacher trajectories contain many locally uninformative reasoning components, motivating selective rather than dense intermediate supervision. Second, we propose VIRAL, which identifies critical teacher and student reasoning patterns and selectively optimizes them through reasoning alignment-guided advantage refinement. Third, experiments across multiple LLMs, a VLM, and eleven reasoning benchmarks consistently demonstrate the effectiveness and generality of VIRAL.

2 Preliminaries and Related Work

In this section, we review related work on RLVR and introduce the background of OPD and GRPO. Due to space constraints, the details of OPD and GRPO are deferred to Appendix C.

Reinforcement learning with verifiable rewards. Reinforcement learning with verifiable rewards (RLVR) has emerged as an effective paradigm for improving the reasoning capabilities of large language models, particularly in domains such as mathematics and code where solution correctness

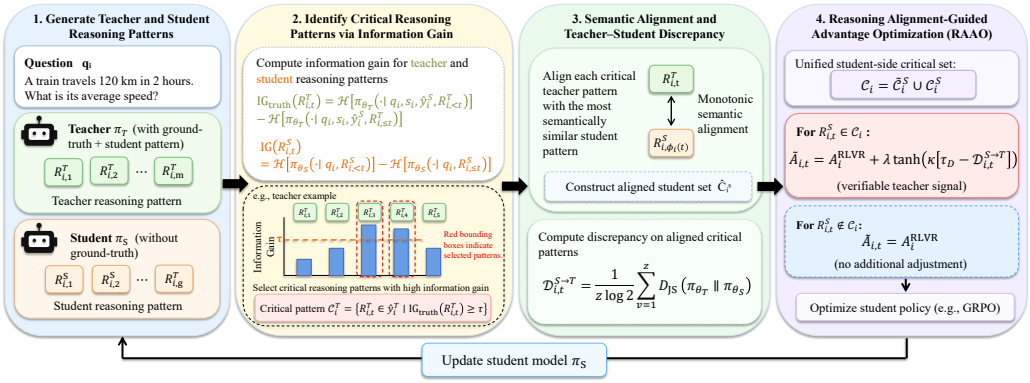


Figure 2: Overview of VIRAL, including critical reasoning pattern identification, teacher–student alignment, and reasoning alignment-guided advantage optimization.

can be automatically verified (Shao et al., 2024; Guo et al., 2025; Yu et al., 2025). However, conventional RLVR typically propagates a sequence-level reward or advantage uniformly across the entire reasoning trajectory, providing limited supervision for the intermediate reasoning steps that determine final correctness. Recent studies have therefore explored finer-grained credit assignment by learning process rewards from outcome supervision (Cui et al., 2025), emphasizing high-entropy tokens that correspond to critical reasoning decisions (Wang et al., 2025), or reweighting token-level updates according to their estimated contribution to the final outcome (He et al., 2026a;b; Wang et al., 2025; 2026). Some studies (Pan et al., 2026; Yang et al., 2026)combine online-policy self-distillation with RLVR to provide verifiable supervision for intermediate reasoning. However, indiscriminately supervising teacher-generated tokens ignores their varying contributions to correctness and may introduce substantial noise and redundancy. Therefore, identifying teacher reasoning components worth supervising is the key motivation of VIRAL.

3 Approach of VIRAL

In this section, we introduce **V**erifiable **I**nformation-guided **R**easoning-level **A**dvantage **L**earning (VIRAL), which consists of two key components: identifying critical reasoning patterns from teacher and student trajectories, and **R**easoning **A**lignment-guided **A**dvantage **O**ptimization (RAAO), which selectively refines their advantages based on teacher–student discrepancies.

3.1 Problem Setup

We consider a dataset of problem–solution pairs, $\mathcal{S} = \{(q_i, s_i)\}_{i=1}^{n}$, where q_i denotes the i-th problem and s_i its ground-truth solution. Under VIRAL, we instantiate a teacher model $\pi_{\theta_\mathrm{T}}(\cdot)$ and a student model $\pi_{\theta_\mathrm{S}}(\cdot)$ that share the same underlying model parameters but differ in their accessible information and roles. The student independently solves q_i using only the problem, $\hat{y}_i^{\mathrm{S}} = \pi_{\theta_\mathrm{S}}(q_i)$, whereas the teacher additionally observes the ground-truth solution s_i and the student trajectory $\hat{y}_i^{\mathrm{S}}$, generating $\hat{y}_i^{\mathrm{T}} = \pi_{\theta_\mathrm{T}}\left(q_i, s_i, \hat{y}_i^{\mathrm{S}}\right)$. With this ground-truth information, the teacher evaluates the student's reasoning against the solution and produces a teacher-guided reasoning trajectory.

The goal of VIRAL is to transform the rich offline supervision in s_i into an on policy signal for supervising the model's own reasoning process. Rather than directly imitating the ground-truth trajectory, VIRAL uses s_i as privileged information to evaluate and guide new reasoning, providing fine-grained intermediate supervision that complements the sparse outcome-level rewards of RLVR.

3.2 Teacher-student instantiation in VIRAL

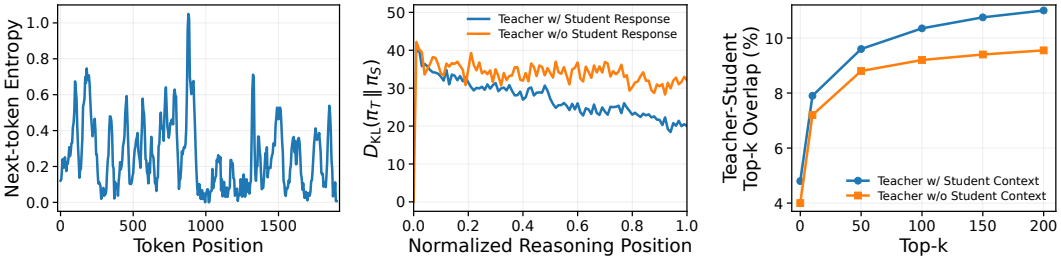
For each query q_i, multiple valid reasoning trajectories may lead to the correct answer. Directly aligning the student response with a single online reasoning trajectory, as in OPSD (Zhao et al., 2026a), can therefore bias the student toward a particular reasoning path and reduce reasoning diversity (Zhao et al., 2026b; Pan et al., 2026). To mitigate this issue, VIRAL reconstructs the teacher

response conditioned jointly on the ground-truth solution and the student's own trajectory. This design has two benefits. First, conditioning on $\hat{y}_i^{\mathrm{S}}$ naturally aligns the teacher trajectory with the student's original reasoning process, yielding more relevant supervision. Second, the teacher can refine the student's reasoning while preserving its underlying reasoning pattern, correcting erroneous steps and improving valid ones rather than forcing the student onto a different solution trajectory. More example can be founded in Appendix F.1

To verify whether this *teacher–student interaction paradigm* achieves these objectives, we conduct empirical analyses from two perspectives: reasoning-distribution distance, which measures teacher–student alignment, and reasoning-pattern consistency, which evaluates whether the teacher refines the student response while preserving its underlying reasoning structure.

Reasoning Distribution Discrepancy Analysis: We conduct this analysis on DeepMath (He et al., 2026c) using Qwen3-8B as the base model. We randomly sample 100 problems and compare the student distribution $\pi_{\theta_{\mathrm{S}}}(q_i)$ with two teacher distributions: one conditioned on $(q_i, s_i, \hat{y}_i^{\mathrm{S}})$, and the other conditioned only on (q_i, s_i). We measure the teacher–student discrepancy using the KL divergence, $D = D_{\mathrm{KL}}\left(\pi_{\theta_{\mathrm{T}}} \,\|\, \pi_{\theta_{\mathrm{S}}}\right)$, where a smaller divergence indicates closer alignment between the teacher and student distributions. As shown in Fig. 3b, conditioning the teacher on the student response $\hat{y}_i^{\mathrm{S}}$ consistently reduces the teacher–student KL divergence, indicating that the resulting teacher distribution remains more closely aligned with the student's reasoning state and provides more relevant supervision.

Reasoning-Pattern Consistency Analysis: We further examine whether conditioning the teacher on the student response improves consistency with the student's reasoning pattern. Following the top-k overlap metric (Li et al., 2026), let $\mathcal{S}_{i,t}^{\mathrm{S}} = \mathrm{TopK}(p_{i,t}^{\mathrm{S}}, k)$ and $\mathcal{S}_{i,t}^{\mathrm{T}} = \mathrm{TopK}(p_{i,t}^{\mathrm{T}}, k)$, where $p_{i,t}^{\mathrm{S}}$ and $p_{i,t}^{\mathrm{T}}$ are the student's and teacher's next-token distributions at reasoning step t for query q_i. We measure consistency as $M_{\mathrm{overlap}} = \mathbb{E}_{i,t}\left[\left|\mathcal{S}_{i,t}^{\mathrm{S}} \cap \mathcal{S}_{i,t}^{\mathrm{T}}\right| / k\right]$, where a larger value indicates stronger teacher–student agreement over high-probability candidate tokens. As shown in Fig. 3c, the teacher conditioned on the student response achieves a consistently higher overlap across all values of k, so the improvement is not limited to the highest-probability tokens but extends to a broader candidate set. This suggests that such conditioning encourages the teacher to preserve the student's token preferences and follow similar reasoning branches, yielding stronger reasoning-pattern consistency.



(a) Distribution of entropy.

(b) Teacher–student reasoning discrepancy.

(c) Teacher–student overlap distribution

Figure 3: We conduct reasoning analysis on 100 DeepMath problems using Qwen3-8B. (a) Teacher entropy with ground-truth answers still fluctuates substantially, producing many high-uncertainty tokens. (b) Teacher–student reasoning discrepancy with and without conditioning the teacher on the student response. (c) Teacher–student reasoning-pattern consistency with and without conditioning the teacher on the student response.

3.2.1 Reasoning-Level Semantic Correspondence between Teacher and Student Responses.

Since the teacher model may refine, insert, or rewrite intermediate reasoning steps of student response, the teacher and student trajectories are not necessarily position-wise aligned. To address this issue, we introduce a *monotonic semantic alignment* mechanism to establish reasoning pattern-level correspondence between the teacher and student responses. Specifically, we decompose the teacher and student responses into a sequence of reasoning patterns, denoted as $\hat{y}_i^{\mathrm{T}} = [R_{i,1}^{\mathrm{T}}, R_{i,2}^{\mathrm{T}}, \cdots, R_{i,m}^{\mathrm{T}}]$ and $\hat{y}_i^{\mathrm{S}} = [R_{i,1}^{\mathrm{S}}, R_{i,2}^{\mathrm{S}}, \cdots, R_{i,g}^{\mathrm{S}}]$, respectively. Each reasoning pattern R_v consists of a sequence of z tokens, i.e., $R_v = [o_{v,1}, o_{v,2}, \cdots, o_{i,z}]$. For each reasoning pattern $R_{i,k}^{\mathrm{T}}$ in the teacher response, we

employ monotonic semantic alignment to identify the most semantically corresponding reasoning pattern $R^{\mathrm{S}}_{i,j}$ in the student response. The detailed procedure is provided in the AppendixC.2

3.3 Information-Gain-Guided Reasoning-Aligned RLVR

3.3.1 Quantifying Reasoning-Pattern Informativeness via Information Gain

For the teacher response $\hat{y}_i^{\mathrm{T}} = [R^{\mathrm{T}}_{i,1}, R^{\mathrm{T}}_{i,2}, \cdots, R^{\mathrm{T}}_{i,v}]$, different reasoning patterns may contribute unequally to the final answer and therefore provide varying degrees of useful supervision to the student model. Directly using the entire teacher response may thus introduce uninformative tokens as noise and hinder the learning of effective reasoning patterns (Pan et al., 2026), as teacher-generated trajectories often contain explanatory or transitional content that contributes little to the core reasoning process (He et al., 2026a; Wang et al., 2025; 2026). Therefore, it is crucial to identify the most informative components of $\hat{y}_i^{\mathrm{T}}$ and selectively use them to supervise the student's reasoning.

To identify the most informative components of the teacher model response, motivated by prior studies (He et al., 2026a; Qian et al., 2025; Ton et al., 2024; Wen et al., 2025), we introduce a ground-truth Information Gain (IG) measure over the teacher reasoning trajectory, denoted as $\mathrm{IG}_{\text{truth}}(\cdot)$, defined by

$$\mathrm{IG}_{\mathrm{truth}}\left(R^{\mathrm{T}}_{i,t}\right) = \mathcal{H}\big[\pi_{\theta_T}\left(\cdot \mid q_i, s_i, \hat{y}_i^{\mathrm{S}}, R^{\mathrm{T}}_{i,<t}\right)\big] - \mathcal{H}\big[\pi_{\theta_T}\left(\cdot \mid q_i, s_i, \hat{y}_i^{\mathrm{S}}, R^{\mathrm{T}}_{i,\leq t}\right)\big]. \tag{1}$$

where $R^{\mathrm{T}}_{i,<t}$ and $R^{\mathrm{T}}_{i,\leq t}$ denote the teacher reasoning prefixes ending at the $(t-1)$-th and t-th reasoning patterns, respectively, and $\mathcal{H}(\cdot)$ denotes the entropy of the model's predictive distribution. Since the teacher has access to the ground-truth solution s_i, its reasoning trajectory can consistently lead to the correct answer. Accordingly, $\mathrm{IG}_{\text{truth}}(R^T_{i,t})$ serves as a ground-truth-informed reasoning-pattern-level signal that measures how much $R^T_{i,t}$ reduces the teacher's predictive uncertainty under privileged conditioning: larger values indicate patterns that are more informative for correctness-oriented reasoning, while smaller values suggest less informative or potentially noisy supervision. Unlike prior information-theoretic measures of reasoning-pattern significance (Wang et al., 2025; Qian et al., 2025; Ton et al., 2024; Chen et al., 2025; He et al., 2026a), $\mathrm{IG}_{\text{truth}}(R^T_{i,t})$ directly grounds pattern importance in verifiable correctness, enabling principled identification of the patterns most relevant to the correct solution. It is note that $\mathrm{IG}_{\text{truth}}$ leverages the off-policy verifiable information in the ground-truth solution s_i to construct on-policy verifiable supervision signals for the current reasoning trajectory, thereby guiding intermediate reasoning during RLVR optimization.

Following the same principle, for the student model, we define $\mathrm{IG}(R^S_{i,t})$ to quantify the contribution of reasoning pattern $R^S_{i,t}$ to its predicted answer $\hat{e}_i$:

$$\mathrm{IG}(R^S_{i,t}) = \mathcal{H}\big[\pi_{\theta_S}\big(\cdot \mid q_i, R^S_{i,<t}\big)\big] - \mathcal{H}\big[\pi_{\theta_S}\big(\cdot \mid q_i, R^S_{i,\leq t}\big)\big]. \tag{2}$$

Different from the teacher-side $\mathrm{IG}_{\text{truth}}(R^T_{i,t})$, $\mathrm{IG}(R^S_{i,t})$ is computed without access to the ground-truth solution s_i. Therefore, it is not directly tied to answer correctness, but instead measures the influence of each reasoning pattern on the student's own reasoning trajectory and predictive distribution. Patterns with higher influence are thus particularly important for RLVR, as they play a larger role in determining how the student's reasoning evolves.

3.3.2 Reasoning Alignment-Guided Advantage Optimization

Having quantified the correctness-relevant informativeness of each teacher reasoning pattern $R^{\mathrm{T}}_{i,t} \in \hat{y}_i^{\mathrm{T}}$ via $\mathrm{IG}_{\text{truth}}(R^{\mathrm{T}}_{i,t})$, a natural question arises: *how can such correctness-relevant information be used to refine the student's credit assignment without forcing it to imitate the teacher trajectory?* To this end, within VIRAL, we introduce a **R**easoning **A**lignment-guided **A**dvantage **O**ptimization (RAAO) mechanism, which establishes correspondence between critical teacher and student reasoning patterns and leverages their local discrepancies to selectively modulate the corresponding advantages, while retaining the original outcome-level reward for the remaining reasoning trajectory. RAAO provides targeted supervision for critical intermediate reasoning patterns while preserving exploration elsewhere.

Critical Reasoning Pattern Identification and Alignment: First, we define the set of *critical reasoning patterns* as $\mathcal{C}^T_i = \left\{R^T_{i,t} \in \hat{y}_i^{\mathrm{T}} \mid \mathrm{IG}_{\text{truth}}(R^T_{i,t}) \geq \tau_T\right\}$, where τ_T is the selection threshold.

Intuitively, $\mathcal{C}_i^T$ captures the reasoning steps that are decisive in transforming the student response $\hat{y}_i^{\mathrm{S}}$ into the teacher response $\hat{y}_i^{\mathrm{T}}$. For each $R_{i,t}^T \in \mathcal{C}_i^T$, we identify the corresponding student-side pattern $R_{i,t}^S$ using the monotonic semantic alignment in Sec. 3.2.1, yielding the aligned student set $\hat{\mathcal{C}}_i^S$. We then quantify the discrepancy between each critical teacher pattern $R_{i,t}^{\mathrm{T}}$ and its aligned student counterpart $\hat{R}_{i,t}^{\mathrm{S}}$ through the *critical reasoning pattern divergence* $\mathcal{D}_{i,t}^{\mathrm{S}\to\mathrm{T}}$, defined as

$$\mathcal{D}_{i,t}^{\mathrm{S}\to\mathrm{T}} = \frac{1}{z\log 2}\sum_{v=1}^{z} D_{\mathrm{JS}}\big(\pi_{\theta_T}\big(\cdot \mid q_i, \hat{y}_i^{\mathrm{T}}, \hat{y}_{i,<t}^{\mathrm{S}}, R_{i,t,<v}^{\mathrm{S}}\big) \,\big\|\, \pi_{\theta_S}\big(\cdot \mid q_i, \hat{y}_{i,<t}^{\mathrm{S}}, R_{i,t,<v}^{\mathrm{S}}\big)\big) \tag{3}$$

where $\hat{y}_{i,<t}^{\mathrm{S}}$ denote the student reasoning prefixes consisting of the first $t-1$ reasoning patterns, while $\hat{R}_{i,t,v<}^{\mathrm{S}}$ denote the prefixes of the first $v-1$ tokens within their t-th reasoning patterns. Accordingly, $\pi_{\theta_T}\big(\cdot \mid q_i, \hat{y}_i^{\mathrm{T}}, \hat{y}_{i,<t}^{\mathrm{S}}, R_{i,t,<v}^{\mathrm{S}}\big)$ and $\pi_{\theta_S}\Big(\cdot \mid q_i, \hat{y}_{i,<t}^{\mathrm{S}}, \hat{R}_{i,t,v<}^{\mathrm{S}}\Big)$ are the corresponding teacher and student next-token predictive distributions, and $D_{\mathrm{JS}}(\cdot)$ denotes the Jensen–Shannon divergence. In Eq. 3, we replace s_i with the student-anchored refinement $\hat{y}_i^{\mathrm{T}}$ as privileged context, and re-forward the student trajectory under this context via teacher forcing, so that teacher and student distributions are compared on the same student prefix. Thus, $\mathcal{D}_{i,t}^{\mathrm{S}\to\mathrm{T}}$ quantifies the local discrepancy between the student's original decision distribution and the teacher's ground-truth-guided corrected distribution: a smaller value indicates stronger consistency with the corrected reasoning direction, whereas a larger value reflects greater probability redistribution and hence a larger reasoning deviation.

Following the same principle, we apply Eq. 2 to the student model to construct a corresponding reasoning pattern set $\mathcal{C}_i^S$, consisting of reasoning patterns that induce the largest decrease in the student's uncertainty. These reasoning patterns are regarded as reasoning-critical for the student, as they contribute most substantially to resolving uncertainty toward the final answer, regardless of whether the resulting reasoning is ultimately correct.

Model optimization: We construct a unified set of critical reasoning patterns $\mathcal{C}_i$ by combining the student-side patterns $\hat{\mathcal{C}}_i^S$ aligned with teacher critical patterns and the patterns $\mathcal{C}_i^S$ independently identified by the student: $\mathcal{C}_i = \hat{\mathcal{C}}_i^S \cup \mathcal{C}_i^S$. The two sets are complementary: $\mathcal{C}_i^S$ captures patterns critical to the student's own reasoning, whereas $\hat{\mathcal{C}}_i^S$ contains student patterns aligned with teacher patterns carrying correctness-relevant information; thus, both warrant targeted supervision. Critical reasoning patterns that cannot be reliably matched to a corresponding pattern in $\hat{y}_i^{\mathrm{T}}$ through Sec. 3.2.1 are excluded from $\mathcal{C}_i$, since the lack of teacher–student correspondence prevents reliable transfer of correctness-relevant information from the ground-truth-informed teacher supervision. For any reasoning pattern $R_{i,t}^S \in \mathcal{C}_i$, its final advantage $\widetilde{A}_{i,t}$ is defined as follows:

$$\widetilde{A}_{i,t} = A_i^{\mathrm{RLVR}} + \lambda \tanh\big(\kappa\left[\tau_D - \mathcal{D}_{i,t}^{\mathrm{S}\to\mathrm{T}}\right]\big), \tag{4}$$

where λ and κ control the strength and sensitivity of the supervision adjustment, respectively, and A_i^{RLVR} denotes the response-level advantage derived from the outcome reward in Eq. 8. The term $\tanh\big(\kappa\left[\tau_D - \mathcal{D}_{i,t}^{\mathrm{S}\to\mathrm{T}}\right]\big)$ provides pattern-specific supervision for the tokens associated with the t-th reasoning pattern. When $\mathcal{D}_{i,t}^{\mathrm{S}\to\mathrm{T}} < \tau_D$, the student pattern is close to its teacher counterpart and receives a positive gain that decreases with the discrepancy; when $\mathcal{D}_{i,t}^{\mathrm{S}\to\mathrm{T}} > \tau_D$, the gain becomes negative, penalizing substantial deviation from the teacher. For each reasoning pattern assigned a final advantage, we apply the corresponding advantage to all tokens within that pattern and optimize the policy using the objective in Eq. 7.

In VIRAL, the privileged teacher influences optimization only through the pattern-level advantage adjustment; the importance ratio $\pi_\theta/\pi_{\theta_{\text{old}}}$ retains its original definition in Eq. 6 and is computed solely from the current and old student policies. This design is necessary because the teacher–student discrepancy only measures local reasoning alignment and does not by itself provide a reliable signal of final-answer correctness. We include the analysis before and after optimization in Appendix B.6.

3.4 Theoretical Analysis

In this section, we analyze how supervising the identified critical reasoning patterns affects the prediction error of the student model. Specifically, we establish an upper bound on the student error probability in terms of the expected teacher–student discrepancy over critical reasoning patterns.

Table 1: Performance comparison of different post-training methods across various base models. The best result in each block is highlighted in **bold**.

Base Model	Method	Benchmark						Avg.
		MATH500	AIME24	AIME25	AIME26	AMC23	Minerva	
Qwen2.5-Bespoke-7B	Base	78.4	18.2	16.9	14.5	59.2	26.9	35.7
	GRPO	81.5	20.7	19.2	19.7	62.3	29.4	38.8
	STEER	82.2	22.4	20.5	15.2	57.4	28.2	37.7
	OPSD	70.4	17.5	14.4	16.7	54.9	24.4	33.1
	RLSD	80.9	21.4	20.4	21.3	60.2	**29.6**	39.0
	VIRAL	**83.7**	**24.1**	**21.7**	**22.4**	**63.5**	28.7	**40.7**
Qwen3-8B	Base	88.2	62.7	57.2	54.9	93.4	36.4	65.6
	GRPO	89.5	66.5	63.6	61.8	94.2	37.2	68.8
	STEER	88.7	67.4	65.5	60.5	94.4	37.0	68.9
	OPSD	90.5	64.7	62.4	62.4	91.5	36.4	68.0
	RLSD	91.3	66.9	64.9	62.0	93.5	37.7	69.4
	RLCSD	91.7	66.4	64.6	61.7	93.9	37.5	69.3
	VIRAL	**92.4**	**68.2**	**66.7**	**62.9**	**95.1**	**38.7**	**70.7**
LLaMA3.1-Bespoke-8B	Base	67.4	10.9	8.4	6.5	48.6	17.2	26.5
	GRPO	75.7	12.4	11.9	10.4	55.7	23.9	31.7
	STEER	77.9	12.7	11.4	11.2	**59.7**	22.4	32.6
	OPSD	66.2	10.4	9.7	7.4	52.5	14.7	26.8
	RLSD	75.8	13.3	13.5	10.7	59.2	22.3	32.5
	VIRAL	**79.4**	**15.3**	**14.2**	**12.9**	59.5	**24.8**	**34.4**

Table 2: Comparison on multimodal reasoning tasks using Qwen2.5-VL-7B as the base model. **Bold** indicates the best result in each block.

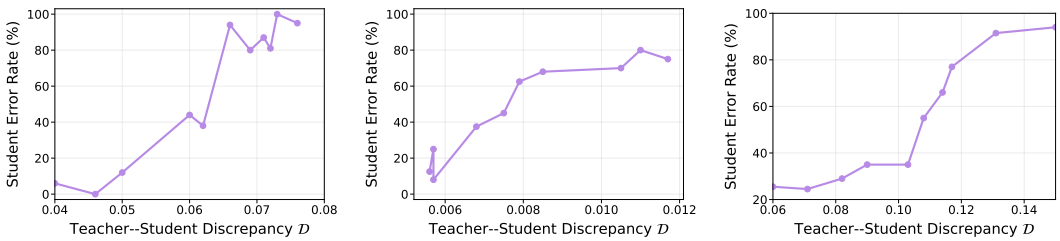
Method	Benchmark					Avg.
	MMMU	MMStar	MathVision	VisuLogic	MathKangaroo	
Base	53.2	61.9	25.4	26.1	27.3	38.8
GRPO	55.3	62.4	27.2	27.2	28.5	40.1
STEER	55.7	63.7	28.5	27.9	29.4	41.0
OPSD	53.9	61.7	25.8	26.5	27.3	39.0
RLSD	55.2	63.4	26.7	27.4	29.4	40.4
MM-Eureka	55.7	62.6	26.9	26.6	29.1	40.2
VIRAL	**57.9**	**63.9**	**30.7**	**28.2**	**29.7**	**42.1**

Theorem 3.1. *Let q_i denote the problem, $\mathcal{D}_{i,t}^{\mathrm{S}\to\mathrm{T}}$ the normalized teacher–student discrepancy at reasoning pattern t, and $\overline{\mathcal{D}}_{i,t}$ its expected discrepancy conditioned on q_i, where $\overline{\mathcal{D}}_{i,t} = \mathbb{E}\left[\mathcal{D}_{i,t}^{\mathrm{S}\to\mathrm{T}} \mid q_i\right]$. With residual error ϵ, it holds*

$$\Pr\left(\hat{e}_i^{\mathrm{S}} \neq e_i \mid q_i\right) \leq \epsilon + \min\left\{\sqrt{z\log 2 \sum_{t\in\mathcal{C}_i} \overline{\mathcal{D}}_{i,t}},\ \sqrt{1 - 4^{-z\sum_{t\in\mathcal{C}_i} \overline{\mathcal{D}}_{i,t}}}\right\}, \tag{5}$$

where $\hat{e}_i^{\mathrm{S}}$ denotes the answer predicted by the student model for question q_i, e_i denotes its corresponding ground-truth answer, z denotes the number of tokens contained in each reasoning pattern, and ϵ denotes the residual prediction error that remains after incorporating ground-truth-guided supervision on the selected critical reasoning patterns.

The proof can be found in Appendix D. According to Theorem 3.1, reducing the teacher–student discrepancy on critical reasoning patterns tightens the upper bound on student prediction error. Thus, the optimization should adapt its training signal to pattern-level alignment: well-aligned patterns should receive stronger positive signals to reinforce reliable reasoning, whereas highly discrepant patterns should receive weaker or negative adjustments to avoid unreliable teacher supervision. Motivated by this, VIRAL augments the original RLVR advantage with a pattern-level adjustment: $\lambda \tanh\left(\kappa[\tau_D - \mathcal{D}_{i,t}^{\mathrm{S}\to\mathrm{T}}]\right)$. This adjustment decreases monotonically with the teacher–student discrepancy $\mathcal{D}_{i,t}^{\mathrm{S}\to\mathrm{T}}$. When the discrepancy is below τ_D, it is positive, increasing the advantage of the corresponding critical reasoning pattern and reinforcing its learning; when the discrepancy exceeds τ_D, it becomes negative, reducing the pattern-level advantage and suppressing patterns that substantially deviate from the teacher. In this way, VIRAL translates the theoretical objective of reducing critical-pattern discrepancy into a practical pattern-level advantage adjustment for policy optimization. We provide optimization results that further support our theoretical analysis in Appendix B.6.



(a) Qwen2.5-7B-Instruct (b) Qwen3-8B (c) LLaMA3.1-8B-Instruct

Figure 4: Relationship between the normalized critical-pattern discrepancy $\frac{z}{|\mathcal{C}_i|}\sum_{t\in\mathcal{C}_i}\overline{\mathcal{D}}_{i,t}$ and the student error rate across three base models.

Empirical Analysis. To empirically examine the relationship between critical-pattern discrepancy and reasoning correctness, we randomly sample 1,000 problems from DeepMath and analyze three base models: Qwen2.5-7B-Instruct, Qwen3-8B, and LLaMA3.1-8B-Instruct. For each problem q_i, we define the normalized critical-pattern discrepancy as $M_i = \frac{z}{|\mathcal{C}_i|}\sum_{t\in\mathcal{C}_i}\overline{\mathcal{D}}_{i,t}$, where z is the reasoning-pattern length and $|\mathcal{C}_i|$ is the number of selected critical patterns, thereby normalizing variations in $|\mathcal{C}_i|$ across examples. This quantity is directly related to the discrepancy term in Theorem 3.1: $z\sum_{t\in\mathcal{C}_i}\overline{\mathcal{D}}_{i,t} = |\mathcal{C}_i|M_i$. Thus, for fixed $|\mathcal{C}_i|$, the discrepancy-dependent component of the prediction-error upper bound is monotonically increasing in M_i, so a smaller M_i yields a tighter bound. As shown in Fig. 4, this relationship is also reflected empirically. Across all three model families, examples with smaller M_i tend to have lower student error rates, while larger M_i is generally associated with higher error rates. Although local fluctuations exist, the overall positive trend remains consistent across models. These results provide empirical evidence that stronger teacher–student alignment on critical reasoning patterns is associated with more reliable reasoning outcomes, consistent with Theorem 3.1.

4 Experiments

4.1 Experiment setup

Models and Benchmark. We first conduct reasoning experiments on three large language models: Qwen2.5-7B-Instruct (Qwen et al., 2025), Qwen3-8B (Yang et al., 2025), and LLaMA3.1-8B-Instruct (Grattafiori et al., 2024). To strengthen the reasoning initialization of Qwen2.5-7B-Instruct and LLaMA3.1-8B-Instruct, we adopt cold-start checkpoints obtained through supervised fine-tuning on Bespoke-Stratos-17k (Labs, 2025), denoted as **Qwen2.5-Bespoke-7B** and **LLaMA3.1-Bespoke-8B**, respectively. In addition, we extend our method to VLMs, using Qwen2.5-VL-7B-Instruct (Wang et al., 2024b) as the base model to further evaluate its applicability in multimodal reasoning scenarios. For LLM reasoning evaluation, we assess reasoning performance on six mathematical reasoning benchmarks: MATH500 (Lightman et al., 2024), AIME 2024 (Zhang & Math-AI, 2024), AIME 2025 (Zhang & Math-AI, 2025), AIME 2026 (Zhang & Math-AI, 2026), AMC 2023 (Team Math-AI, 2023), and Minerva (Lewkowycz et al., 2022). For multimodal evaluation, we use five benchmarks spanning general multimodal understanding and visual reasoning: MMMU (Yue et al., 2024), MMStar (Chen et al., 2024), MathVision (Wang et al., 2024a), VisuLogic (Xu et al., 2026), and MathKangaroo (Sáez et al., 2025).

Baseline. We compare VIRAL with GRPO (Shao et al., 2024), OPSD (Zhao et al., 2026a), STEER (Hao et al., 2026), RLSD (Yang et al., 2026), RLCSD (Pan et al., 2026), covering representative baselines from different methodological categories, including standard RLVR, self-distillation, entropy-aware optimization, and distillation-enhanced RLVR. For the multimodal setting, we further include MM-Eureka (Meng et al., 2025), a method specifically developed for multimodal reasoning, as an additional baseline to enable a more comprehensive evaluation. Due to space constraints, the training and evaluation settings are provided in Appendix B.1.

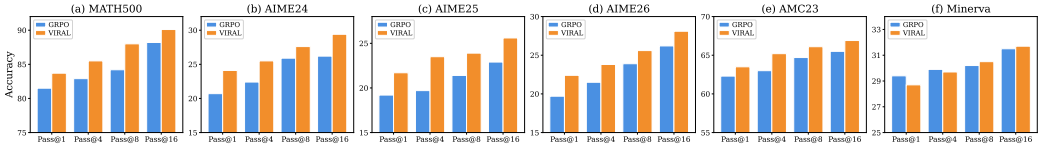


Figure 5: Effect of scaling test-time compute across different benchmarks (Pass@K), where Qwen2.5-Bespoke-7B is considered as the base model.

Table 3: Ablation study of VIRAL, where Qwen2.5-Bespoke-7B is considered as base model. **Bold** indicates the best result in each block.

Method	Benchmark						Avg.
	MATH500	AIME24	AIME25	AIME26	AMC23	Minerva	
VIRAL $_{\text{w/o GT-IG Selection}}$	80.2	19.5	20.5	19.2	57.2	28.2	37.5
VIRAL $_{\text{w/o Student-conditioned Teacher}}$	82.3	23.9	20.6	20.4	60.3	**29.2**	39.5
VIRAL $_{\text{w/Student Critical Only}}$	81.4	22.6	19.7	18.5	58.7	27.4	38.1
VIRAL $_{\text{w/ Teacher Critical Only}}$	82.2	**25.4**	19.2	21.2	61.5	28.2	39.6
VIRAL	**83.7**	24.1	**21.7**	**22.4**	**63.5**	28.7	**40.7**

4.2 Performance of VIRAL

Main results. As shown in Table 1, VIRAL consistently improves reasoning performance across all three base models, achieving the best average accuracy in every model block. For Qwen3-8B, we further evaluate performance under the thinking mode with the maximum token budget, with the results reported in Appendix B.3. In addition, a statistical analysis of our experimental results is provided in Appendix B.8.

Performance on VLM Tasks. As shown in Table 2, VIRAL consistently outperforms all baselines across the five multimodal reasoning benchmarks, achieving the best average accuracy of 42.1. Compared with the strongest baseline, STEER, VIRAL improves the average performance by 1.1 points, with particularly clear gains on MMMU and MathVision. These results suggest that VIRAL generalizes effectively from language reasoning to multimodal reasoning tasks.

Ablation analysis of VIRAL. As shown in Table 3, the full VIRAL achieves the best average performance of 40.7. Removing GT-IG selection causes the largest degradation, reducing the average accuracy by 3.2 points, highlighting the importance of identifying correctness-relevant reasoning patterns. Using only student- or teacher-side critical patterns also leads to noticeable drops, indicating that the two provide complementary supervision. Moreover, removing the student-conditioned teacher decreases the average performance to 39.5, validating the benefit of conditioning teacher reasoning on the student's trajectory.

Effect of Test-Time Scaling: As shown in Fig. 5, increasing test-time compute generally improves performance for both GRPO and VIRAL across all benchmarks. Notably, VIRAL consistently achieves larger gains than GRPO on MATH500, AIME24, AIME25, AIME26, and AMC23, and the advantage is largely maintained as Pass@K increases. On Minerva, the two methods become comparable at larger K. These results suggest that VIRAL can effectively leverage additional test-time computation, demonstrating favorable scaling behavior across diverse reasoning tasks.

5 Conclusion

In this work, we propose VIRAL, a verifiable information-guided RLVR framework that selectively supervises critical intermediate reasoning patterns rather than the entire reasoning trajectory. VIRAL identifies informative teacher and student reasoning patterns and introduces RAAO to refine their advantages according to local teacher–student discrepancies, while preserving exploration on the remaining trajectory. Extensive experiments across multiple LLMs and a VLM demonstrate consistent improvements over strong RLVR and self-distillation baselines, highlighting the effectiveness and generality of selective reasoning-level supervision.

AI USE STATEMENT

Generative AI tools were used for literature search and summarization, text editing and refinement, and assistance with analysis and plotting code. All AI-assisted text and code were reviewed and verified by the authors. The research questions, experimental design, and interpretation of results are the authors' own, and the authors take full responsibility for the final content of this work.

ETHICS STATEMENT

This work studies reinforcement learning for mathematical and multimodal reasoning using existing models and public datasets. It does not involve human-subject research or the collection of new personal data. Improving reasoning capabilities may benefit downstream applications but may also facilitate misuse. Moreover, ground-truth-conditioned teacher supervision and verifiable outcome rewards do not guarantee that intermediate reasoning is correct, unbiased, or safe. Our evaluation focuses on reasoning performance and does not establish suitability for safety-critical deployment.

REPRODUCIBILITY STATEMENT

Section 3 describes the privileged teacher construction, critical reasoning pattern selection, teacher–student discrepancy, and advantage adjustment. Appendix C.2 details the semantic alignment procedure, and Figure 9 provides the teacher prompt. Appendix B.1 reports the training data, hyperparameters, and evaluation configuration. The assumptions and proofs supporting the theoretical analysis are provided in Appendix D. We will release the training and evaluation code, configuration files, and data-processing scripts to facilitate reproduction of our results.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations*, volume 2024, pp. 21246–21263, 2024.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *Advances in Neural Information Processing Systems*, 37:27056–27087, 2024.

Minghan Chen, Guikun Chen, Wenguan Wang, and Yi Yang. Seed-grpo: Semantic entropy enhanced grpo for uncertainty-aware policy optimization. *arXiv preprint arXiv:2505.12346*, 2025.

Ganqu Cui, Lifan Yuan, Zefan Wang, Hanbin Wang, Yuchen Zhang, Jiacheng Chen, Wendi Li, Bingxiang He, Yuchen Fan, Tianyu Yu, et al. Process reinforcement through implicit rewards. *arXiv preprint arXiv:2502.01456*, 2025.

Guanting Dong, Hangyu Mao, Kai Ma, Licheng Bao, Yifei Chen, Zhongyuan Wang, Zhongxia Chen, Jiazhen Du, Huiyang Wang, Fuzheng Zhang, et al. Agentic reinforced policy optimization. *ICLR*, 2026.

Zhihao Dou, Qinjian Zhao, Zhongwei Wan, Dinggen Zhang, Weida Wang, Towsif Raiyan, Benteng Chen, Qingtao Pan, Yang Ouyang, Zhiqiang Gao, et al. Plan then action: High-level planning guidance reinforcement learning for llm reasoning. *arXiv preprint arXiv:2510.01833*, 2025.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Etash Guha, Ryan Marten, Sedrick Keh, Negin Raoof, Georgios Smyrnis, Hritik Bansal, Marianna Nezhurina, Jean Mercat, Trung Vu, Zayne Sprague, et al. Openthoughts: Data recipes for reasoning models. In *International Conference on Learning Representations*, volume 2026, pp. 108059–108130, 2026.

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

Zhezheng Hao, Hong Wang, Haoyang Liu, Jian Luo, Jiarui Yu, Hande Dong, Qiang Lin, Can Wang, and Jiawei Chen. Rethinking entropy interventions in rlvr: An entropy change perspective. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 31105–31133, 2026.

Yinhan He, Yaochen Zhu, Mingjia Shi, Wendy Zheng, Lin Su, Xiaoqing Wang, Qi Guo, and Jundong Li. Iapo: Information-aware policy optimization for token-efficient reasoning. *arXiv preprint arXiv:2602.19049*, 2026a.

Yuhang He, Haodong Wu, Siyi Liu, Hongyu Ge, Hange Zhou, Keyi Wu, Zhuo Zheng, Qihong Lin, Zixin Zhong, and Yongqi Zhang. Rethinking token-level credit assignment in rlvr: A polarity-entropy analysis. *arXiv preprint arXiv:2604.11056*, 2026b.

Zhiwei He, Tian Liang, Jiahao Xu, Qiuzhi Liu, Xingyu Chen, Yue Wang, Linfeng Song, Dian Yu, Zhenwen Liang, Wenxuan Wang, et al. Deepmath-103k: A large-scale, challenging, decontaminated, and verifiable mathematical dataset for advancing reasoning. In *International Conference on Learning Representations*, volume 2026, pp. 138306–138322, 2026c.

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. *NeurIPS*, 2021.

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

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, et al. Tulu 3: Pushing frontiers in open language model post-training. *arXiv preprint arXiv:2411.15124*, 2024.

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, et al. Solving quantitative reasoning problems with language models. *Advances in neural information processing systems*, 35:3843–3857, 2022.

Yaxuan Li, Yuxin Zuo, Bingxiang He, Jinqian Zhang, Chaojun Xiao, Cheng Qian, Tianyu Yu, Huanang Gao, Wenkai Yang, Zhiyuan Liu, et al. Rethinking on-policy distillation of large language models: Phenomenology, mechanism, and recipe. *arXiv preprint arXiv:2604.13016*, 2026.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations*, volume 2024, pp. 39578–39601, 2024.

Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Wanli Ouyang. Flow-grpo: Training flow matching models via online rl. *arXiv preprint arXiv:2505.05470*, 2025.

Fanqing Meng, Lingxiao Du, Zongkai Liu, Zhixiang Zhou, Quanfeng Lu, Daocheng Fu, Tiancheng Han, Botian Shi, Wenhai Wang, Junjun He, et al. Mm-eureka: Exploring the frontiers of multimodal reasoning with rule-based reinforcement learning. *arXiv preprint arXiv:2503.07365*, 2025.

Leyi Pan, Shuchang Tao, Yunpeng Zhai, Lingzhe Zhang, Zhaoyang Liu, Bolin Ding, Aiwei Liu, and Lijie Wen. Rlcsd: Reinforcement learning with contrastive on-policy self-distillation. *arXiv preprint arXiv:2606.11709*, 2026.

Chen Qian, Dongrui Liu, Haochen Wen, Zhen Bai, Yong Liu, and Jing Shao. Demystifying reasoning dynamics with mutual information: Thinking tokens are information peaks in llm reasoning. *arXiv preprint arXiv:2506.02867*, 2025.

A Yang Qwen, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengpeng Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, pp. 2412–15115, 2025.

Arnau Igualde Sáez, Lamyae Rhomrasi, Yusef Ahsini, Ricardo Vinuesa, Sergio Hoyas, Jose P García Sabater, Marius J Alfonso, and J Alberto Conejero. Evaluating visual mathematics in multimodal llms: A multilingual benchmark based on the kangaroo tests. *arXiv preprint arXiv:2506.07418*, 2025.

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

Team Math-AI. AMC 2023. `https://huggingface.co/datasets/math-ai/amc23`, 2023. 40 problems compiled from the 2023 AMC 12A and 12B.

Jean-Francois Ton, Muhammad Faaiz Taufiq, and Yang Liu. Understanding chain-of-thought in llms through information theory. *arXiv preprint arXiv:2411.11984*, 2024.

Haozhe Wang, Qixin Xu, Che Liu, Junhong Wu, Fangzhen Lin, and Wenhu Chen. Emergent hierarchical reasoning in llms through reinforcement learning. *ICLR*, 2026.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with math-vision dataset. *Advances in Neural Information Processing Systems*, 37:95095–95169, 2024a.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024b.

Shenzhi Wang, Le Yu, Chang Gao, Chujie Zheng, Shixuan Liu, Rui Lu, Kai Dang, Xionghui Chen, Jianxin Yang, Zhenru Zhang, et al. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for llm reasoning. *NeurIPS*, 2025.

Zichen Wen, Yifeng Gao, Weijia Li, Conghui He, and Linfeng Zhang. Token pruning in multimodal large language models: Are we solving the right problem? In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 15537–15549, 2025.

Weiye Xu, Jiahao Wang, Weiyun Wang, Zhe Chen, Wengang Zhou, Aijun Yang, Lewei Lu, Houqiang Li, Xiaohua Wang, Xizhou Zhu, et al. Visulogic: A benchmark for evaluating visual reasoning in multi-modal large language models. In *International Conference on Learning Representations*, volume 2026, pp. 25966–26003, 2026.

Zeyue Xue, Jie Wu, Yu Gao, Fangyuan Kong, Lingting Zhu, Mengzhao Chen, Zhiheng Liu, Wei Liu, Qiushan Guo, Weilin Huang, et al. Dancegrpo: Unleashing grpo on visual generation. *arXiv preprint arXiv:2505.07818*, 2025.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Chenxu Yang, Chuanyu Qin, Qingyi Si, Minghui Chen, Naibin Gu, Dingyu Yao, Zheng Lin, Weiping Wang, Jiaqi Wang, and Nan Duan. Self-distilled rlvr. *arXiv preprint arXiv:2604.03128*, 2026.

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

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9556–9567, 2024.

Yifan Zhang and Team Math-AI. American invitational mathematics examination (aime) 2024, 2024.

Yifan Zhang and Team Math-AI. American invitational mathematics examination (aime). *Mathematical Association of America (MAA)*, 2025.

Yifan Zhang and Team Math-AI. American invitational mathematics examination (aime) 2026, 2026.

Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. Self-distilled reasoner: On-policy self-distillation for large language models. *arXiv preprint arXiv:2601.18734*, 2026a.

Ziqi Zhao, Xinyu Ma, Liu Yang, Yujie Feng, Daiting Shi, Jingzhou He, Xin Xin, Zhaochun Ren, and Xiao-Ming Wu. Rosd: Reflective on-policy self-distillation for language model reasoning across domains. *arXiv preprint arXiv:2605.28014*, 2026b.

APPENDIX

A CODE IMPLEMENTATION

Our code is available here: link.

B EXPERIMENTAL SETTINGS AND FURTHER EXPERIMENTAL RESULTS

B.1 EXPERIMENT SETTING

Training Data. For LLM training, we randomly sample 40K examples from OpenThoughts-Math (Guha et al., 2026) for reinforcement learning, where the problem is used as the student prompt and the reference solution provides privileged information to the teacher. For VLM reasoning tasks, we follow (Meng et al., 2025) and use the same training data for training.

Training details. Unless otherwise specified, we train all models for 15 epochs using AdamW with a learning rate of 1×10^{-6}, a weight decay of 1×10^{-2}, a global batch size of 128, and a KL coefficient of 1×10^{-2}. During rollout sampling, we generate 5 responses per query with a temperature of 1.0 and top-p of 1.0. The maximum prompt length is set to 8,192 tokens for all models. The maximum response length is set to 8,192 tokens for Qwen2.5-Bespoke-7B and LLaMA3.1-Bespoke-8B, and to 12,000 tokens for Qwen3-8B. For each base model, we use the same maximum response length across all compared methods to ensure a fair comparison. All other method-specific hyperparameters follow the default settings reported in the corresponding papers. We train Qwen2.5-Bespoke-7B and LLaMA3.1-Bespoke-8B for 10 epochs, while Qwen3-8B-Instruct is trained for 6 epochs, using a global batch size of 128 for all models. For each base model, all RL-based baselines are trained with the same number of epochs and global batch size to ensure a fair comparison.

We set the reasoning-pattern length to $z = 6$ for all experiments. The remaining method-specific hyperparameters are selected using a 1,000-sample subset of DeepMath for small-scale pilot training. We perform grid search over candidate configurations and select the setting that achieves the highest average accuracy reward on the tuning set after training. Specifically, the information-gain threshold τ_T is set to the 90th percentile of the information-gain distribution, such that only the top 10% of reasoning patterns with the highest information gain are retained. The advantage-adjustment strength is set to $\lambda = 0.05$. We set $\kappa = 0.4$ for Qwen2.5-Bespoke-7B and LLaMA3.1-Bespoke-8B, and $\kappa = 1.0$ for Qwen3-8B. The discrepancy threshold τ_D is set to the 40th percentile of the teacher–student discrepancy distribution over all reasoning patterns.

Evaluation Setup. For each base model, all compared methods are evaluated under the same configuration to ensure a fair comparison. Each benchmark is evaluated independently for 64 runs, and

the reported result is the average performance over these runs. During evaluation, we use a temperature of 0.6 and top-p of 0.95. The maximum output length is set to 8,192 tokens for Qwen2.5-Bespoke-7B and LLaMA3.1-Bespoke-8B, and to 18,000 tokens for Qwen3-8B. All other evaluation settings are kept identical across methods for the same base model.

B.2 Further Ablation Studies of VIRAL

Variant I. To investigate whether the improvement of VIRAL simply comes from applying sparse supervision to a subset of reasoning patterns, we replace the information-gain-based critical-pattern selection with random selection. Specifically, for each response, we randomly sample the same number of reasoning patterns as selected by VIRAL, while keeping the subsequent teacher–student discrepancy estimation and optimization procedure unchanged. This variant isolates whether the effectiveness of VIRAL arises from the informativeness of the selected reasoning patterns rather than sparsity alone.

Variant II. To study whether the teacher–student discrepancy needs to be incorporated through advantage adjustment, we retain the original GRPO advantage and instead introduce the discrepancy as an auxiliary loss over the selected critical reasoning patterns. Specifically, rather than modifying the token-level advantage, we optimize $\mathcal{L}_{\text{Variant-II}} = \mathcal{L}_{\text{GRPO}} + \beta \frac{1}{|\mathcal{C}_i|} \sum_{t \in \mathcal{C}_i} \mathcal{D}_{i,t}^{\text{S}\to\text{T}}$, where β controls the strength of the additional alignment objective. This variant directly regularizes the student toward the ground-truth-guided teacher distribution while leaving the original GRPO credit assignment unchanged.

Variant III. To evaluate the necessity of reasoning-level semantic correspondence, we replace the monotonic semantic alignment in VIRAL with position-wise matching. Specifically, the t-th teacher reasoning pattern is directly paired with the t-th student reasoning pattern, $R_{i,t}^{\text{T}} \leftrightarrow R_{i,t}^{\text{S}}$, without considering their semantic similarity. All other components of VIRAL remain unchanged. This variant examines whether explicitly resolving the structural mismatch between teacher and student reasoning trajectories is important for obtaining reliable pattern-level supervision.

Table 4: Further ablation studies of VIRAL, where Qwen2.5-Bespoke-7B is considered as the base model. **Bold** indicates the best result in each column.

Method	Benchmark						Avg.
	MATH500	AIME24	AIME25	AIME26	AMC23	Minerva	
Variant I	79.4	20.9	19.2	19.4	58.9	25.5	37.2
Variant II	80.2	22.9	19.5	20.7	61.2	26.8	38.6
Variant III	81.4	21.7	20.9	20.4	61.8	27.2	38.9
VIRAL	**83.7**	**24.1**	**21.7**	**22.4**	**63.5**	**28.7**	**40.7**

As shown in Table 4, the full VIRAL consistently outperforms all three variants, achieving the highest average accuracy of 40.7. Variant I exhibits the largest degradation, with the average performance dropping to 37.2, indicating that the improvement of VIRAL cannot be attributed merely to sparse supervision; rather, selecting informative critical reasoning patterns is essential. Variant II reaches an average accuracy of 38.6, suggesting that directly imposing the teacher–student discrepancy as an auxiliary regularization loss is less effective than incorporating it into the advantage for fine-grained credit assignment. Variant III performs slightly better with an average accuracy of 38.9, but still falls behind the full VIRAL, demonstrating that position-wise matching cannot reliably capture the correspondence between structurally different teacher and student reasoning trajectories. Overall, these results validate the importance of information-guided critical-pattern selection, advantage-level supervision, and reasoning-level semantic alignment in VIRAL.

B.3 Effect of Maximum Generation Length

We further evaluate all methods on Qwen3-8B with the maximum generation length increased to 38,912 tokens. To reduce evaluation cost, each configuration is evaluated over 24 runs. As shown in Table 5, VIRAL achieves the highest average accuracy of 73.4, outperforming GRPO, STEER, OPSD, and RLSD by 1.5, 0.8, 2.1, and 0.6 points, respectively. VIRAL also achieves the best performance on five of the six benchmarks, with STEER, RLCSD and RLSD slightly outperforming it

Table 5: Performance comparison of different post-training methods under an increased maximum generation length, where Qwen3-8B is considered as the base model. **Bold** indicates the best result in each column.

Method	Benchmark						Avg.
	MATH500	AIME24	AIME25	AIME26	AMC23	Minerva	
Base	88.9	74.2	66.4	60.2	93.8	38.6	70.4
GRPO	90.5	76.5	68.2	62.3	94.6	39.0	71.9
STEER	92.2	76.2	**69.2**	63.9	94.9	39.4	72.6
OPSD	91.4	75.4	66.7	61.7	93.8	38.7	71.3
RLSD	92.9	77.2	68.9	63.6	94.2	39.7	72.8
RLCSD	93.1	76.4	68.9	63.2	94.6	39.2	72.6
VIRAL	**93.5**	**77.9**	68.7	**64.1**	**95.8**	**40.6**	**73.4**

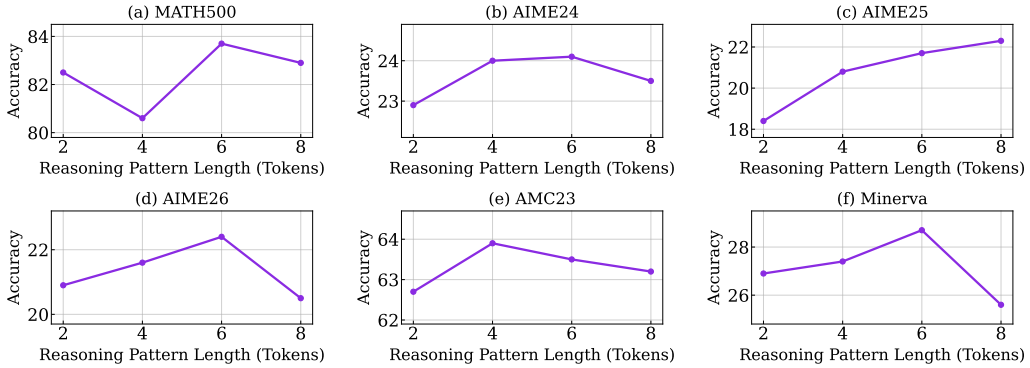


Figure 6: Sensitivity analysis of reasoning pattern length, where Qwen2.5-Bespoke-7B is considered as the base model.

only on AIME25. These results show that the advantage of VIRAL is maintained under a substantially larger generation budget, suggesting that its improvements are not dependent on a restrictive output-length setting and remain robust when the model is allowed to perform longer reasoning.

B.4 Sensitivity Analysis

We investigate the sensitivity of VIRAL to the reasoning pattern length by varying the number of tokens in each pattern from 2 to 8. As shown in Fig. 6, the performance remains relatively stable across different settings, indicating that VIRAL is not highly sensitive to this hyperparameter. In general, a moderate pattern length of 4–6 tokens performs favorably across most benchmarks, with a length of 6 achieving the best results on MATH500, AIME24, AIME26, and Minerva. When the pattern becomes overly long, performance tends to decrease on several benchmarks, particularly AIME26 and Minerva, suggesting that longer patterns may introduce less relevant information and weaken fine-grained supervision. An exception is AIME25, where performance continues to improve with increasing pattern length. Overall, these results suggest that a moderate reasoning pattern length provides a good balance between capturing sufficient semantic information and preserving fine-grained supervision.

B.5 Performance Comparison of VIRAL Built Upon Different RLVR Methods

Beyond GRPO, we further integrate VIRAL with two alternative RLVR methods, DAPO and Dr.GRPO, using Qwen2.5-Bespoke as the base model. As shown in Table 6, VIRAL consistently improves performance across all six benchmarks under both RLVR backbones. Specifically, VIRAL improves the average accuracy of DAPO from 38.5 to 40.6 (+2.1 points) and Dr.GRPO from 37.5 to 39.2 (+1.7 points). These results indicate that the effectiveness of VIRAL is not tied to a specific RLVR algorithm and can provide complementary gains across different optimization frameworks.

Table 6: Performance comparison of VIRAL built upon different RLVR methods, where Qwen2.5-Bespoke is considered as the base model.

Method	Benchmark						Avg.
	MATH500	AIME24	AIME25	AIME26	AMC23	Minerva	
DAPO	80.2	21.9	20.5	20.2	60.7	27.5	38.5
VIRAL $_{\text{w/DAPO}}$	**83.3**	**23.7**	**22.4**	**21.9**	**62.6**	**29.4**	**40.6**
Dr.GRPO	81.4	19.3	18.9	17.7	61.2	26.2	37.5
VIRAL $_{\text{w/Dr.GRPO}}$	**83.2**	**22.1**	**20.5**	**19.4**	**62.9**	**27.1**	**39.2**

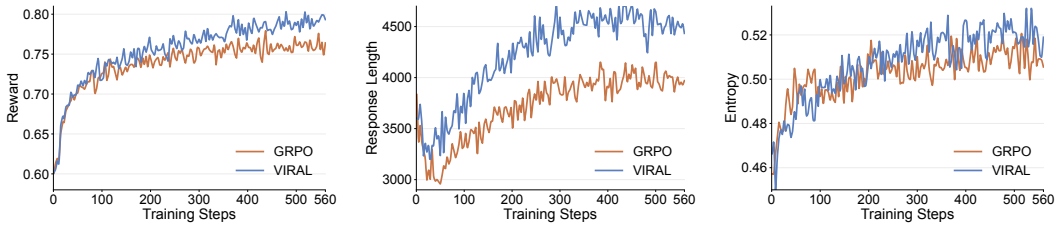
B.6 Discrepancy and accuracy changes before and after optimization

Table 7: Critical-pattern discrepancy and accuracy before and after optimization on 1,000 sampled DeepMath problems.

Model	Before optimization		After optimization	
	$z\,\mathbb{E}_{t\in\mathcal{C}_i}[\mathcal{D}_{i,t} \mid q_i]$	Acc.	$z\,\mathbb{E}_{t\in\mathcal{C}_i}[\mathcal{D}_{i,t} \mid q_i]$	Acc.
Qwen2.5-Bespoke-7B	0.067	56.4	0.054	66.2
LLaMA3.1-Bespoke-8B	0.113	54.2	0.092	60.7
Qwen3-8B	0.0094	72.2	0.0082	76.5

Table 7 compares the critical-pattern discrepancy and reasoning accuracy before and after optimization on 1,000 sampled DeepMath problems. Across all three models, our VIRAL consistently reduces the discrepancy at critical reasoning patterns while improving accuracy. Specifically, the discrepancy decreases by 19.4%, 18.6%, and 12.8% for Qwen2.5-Bespoke-7B, LLaMA3.1-Bespoke-8B, and Qwen3-8B, respectively, accompanied by accuracy gains of 9.8, 6.5, and 4.3 percentage points. Moreover, models with larger discrepancy reductions generally exhibit larger performance improvements. These results support the effectiveness of our VIRAL and indicate that reducing the distributional mismatch at critical reasoning patterns is closely associated with improved reasoning accuracy.

B.7 Training Dynamics



(a) Training reward. (b) Response length. (c) Policy entropy.

Figure 7: Training dynamics of GRPO and VIRAL using Qwen2.5-Bespoke-7B as the base model.

B.8 Analysis of Statistical Significance

Table 8 reports the pairwise statistical comparison between VIRAL and GRPO across different base models and benchmarks. Overall, VIRAL achieves statistically significant improvements in 16 out of 18 model–task combinations, demonstrating that its gains are consistent across both model architectures and reasoning benchmarks. In particular, 9 comparisons reach $p < 0.001$, indicating strong statistical evidence for the observed improvements.

The improvements are especially consistent on LLaMA3.1-Bespoke-8B, where VIRAL significantly outperforms GRPO on all six benchmarks, with gains ranging from $+0.9$ to $+3.8$ points. Similar trends are observed for Qwen2.5-Bespoke-7B and Qwen3-8B, where significant improvements are

Table 8: Pairwise statistical comparison between VIRAL and GRPO. Δ denotes the accuracy improvement of VIRAL over GRPO in percentage points, and Var. denotes the variance across repeated evaluation runs. *, **, and *** indicate $p < 0.05$, $p < 0.01$, and $p < 0.001$, respectively; ns denotes not significant.

Base Model	**Task**	**Δ (pt)**	**Var.**	**p / sig.**
Qwen2.5-Bespoke-7B	MATH500	+2.2	0.24	$< 0.0001^{***}$
	AIME24	+3.4	0.55	$< 0.0001^{***}$
	AIME25	+2.5	0.49	0.0006^{***}
	AIME26	+2.7	0.61	0.0014^{**}
	AMC23	+1.2	0.31	0.0073^{**}
	Minerva	-0.7	0.27	0.1036^{ns}
Qwen3-8B	MATH500	+2.9	0.19	$< 0.0001^{***}$
	AIME24	+1.7	0.46	0.0248^{*}
	AIME25	+3.1	0.42	$< 0.0001^{***}$
	AIME26	+1.1	0.52	0.1529^{ns}
	AMC23	+0.9	0.25	0.0084^{**}
	Minerva	+1.5	0.22	$< 0.0001^{***}$
LLaMA3.1-Bespoke-8B	MATH500	+3.7	0.29	$< 0.0001^{***}$
	AIME24	+2.9	0.63	0.0003^{***}
	AIME25	+2.3	0.57	0.0016^{**}
	AIME26	+2.5	0.69	0.0011^{**}
	AMC23	+3.8	0.35	$< 0.0001^{***}$
	Minerva	+0.9	0.30	0.0134^{*}

obtained on five out of six benchmarks. The only non-significant cases are Minerva with Qwen2.5-Bespoke-7B ($\Delta = -0.7$, $p = 0.1036$) and AIME26 with Qwen3-8B ($\Delta = +1.1$, $p = 0.1529$).

Moreover, the relatively small variances across repeated runs (0.19–0.69) suggest that the observed improvements are stable rather than being driven by large run-to-run fluctuations. These results provide statistical evidence that VIRAL consistently improves over GRPO across diverse base models and reasoning tasks.

C PRELIMINARIES KNOWLEDGE.

C.1 GROUP RELATIVE POLICY OPTIMIZATION

Reinforcement Learning with Verifiable Rewards (RLVR) has emerged as a dominant paradigm for eliciting reasoning capabilities from large language models, and Group Relative Policy Optimization (GRPO) (Shao et al., 2024) is arguably its most widely adopted instantiation. GRPO inherits the clipped surrogate objective of Proximal Policy Optimization (PPO) (Schulman et al., 2017), but dispenses with the learned value network that PPO relies on for advantage estimation. Instead of regressing a per-token baseline, GRPO samples a *group* of responses for each prompt and treats the group's empirical reward statistics as the baseline. This design removes an entire model from the training loop, halving memory footprint and eliminating the instability associated with critic learning, while retaining—and in practice often exceeding—the reasoning gains of PPO-style fine-tuning.

We now fix notation. Let Q be the distribution over questions, π_θ the policy under optimization, $\pi_{\theta_{\text{old}}}$ the behavior policy used to collect rollouts, and $\pi_{\theta_{\text{ref}}}$ a frozen reference policy (typically the initial checkpoint). For a question $q \sim Q$, GRPO draws N responses $\{\mathfrak{o}_i\}_{i=1}^{N}$ independently from $\pi_{\theta_{\text{old}}}(\cdot \mid q)$, and we write $\mathfrak{o}_i^t$ for the t-th token of $\mathfrak{o}_i$ and $\mathfrak{o}_i^{<t}$ for its prefix. Denoting the token-level importance ratio by

$$\rho_i^t(\theta) = \frac{\pi_\theta\big(\mathfrak{o}_i^t \mid q, \mathfrak{o}_i^{<t}\big)}{\pi_{\theta_{\text{old}}}\big(\mathfrak{o}_i^t \mid q, \mathfrak{o}_i^{<t}\big)}, \tag{6}$$

the GRPO objective maximized with respect to θ reads

$$J_{\text{GRPO}}(\theta) = \mathbb{E}_{q\sim Q,\ \{\mathfrak{o}_i\}_{i=1}^{N}\sim\pi_{\theta_{\text{old}}}(\cdot|q)}\Bigg[\frac{1}{N}\sum_{i=1}^{N}\frac{1}{|\mathfrak{o}_i|}\sum_{t=1}^{|\mathfrak{o}_i|} \min\Big(\rho_i^t(\theta)\,A_i,\ \text{clip}\big(\rho_i^t(\theta),\,1-\epsilon,\,1+\epsilon\big)\,A_i\Big)\ -\ \beta\, D_{\text{KL}}\big(\pi_\theta \,\|\, \pi_{\theta_{\text{ref}}}\big)\Bigg], \tag{7}$$

where ϵ bounds the trust region through clipping and β weights the KL penalty that anchors π_θ to the reference policy. The $\min(\cdot,\cdot)$ operator, carried over from PPO, yields a pessimistic lower bound on the unclipped surrogate and thereby discourages destructively large policy updates.

The distinguishing ingredient of GRPO is how the advantage A_i is formed. Each response $\mathfrak{o}_i$ receives a scalar reward r_i from a verifier (e.g., an exact-match check against the reference answer), and the resulting group rewards $\{r_1, \ldots, r_N\}$ are standardized within the group:

$$A_i = \frac{r_i - \mu}{\sigma}, \qquad \mu = \frac{1}{N}\sum_{j=1}^{N} r_j, \qquad \sigma = \sqrt{\frac{1}{N}\sum_{j=1}^{N}\big(r_j - \mu\big)^2}. \tag{8}$$

Standardization renders the advantage invariant to the scale and offset of the reward, so that a response is credited only insofar as it outperforms its siblings sampled for the same question. Crucially, A_i is a *sequence-level* quantity: it is shared verbatim by every token $\mathfrak{o}_i^t$ in Eq. equation 7, regardless of that token's individual contribution to correctness.

In the RLVR setting considered here, every sampled response $\mathfrak{o} \in \{\mathfrak{o}_i\}_{i=1}^{N}$ comprises a chain-of-thought (CoT) c followed by a final answer, and the verifier inspects only the latter. Consequently, all tokens of the reasoning trace inherit a single outcome-based credit signal—an observation that motivates the token-level supervision we develop in the remainder of this paper.

C.1.1 ON-POLICY DISTILLATION AND ON-POLICY SELF-DISTILLATION.

On-policy distillation (OPD) transfers knowledge from a teacher model while preserving the state distribution induced by the student policy. Given a query q_i, the student policy $\pi_{\theta_{\text{S}}}$ samples an on-policy response $\hat{y}_i^{\text{S}} = (y_{i,1}^{\text{S}}, \ldots, y_{i,T}^{\text{S}})$. Instead of requiring the student to imitate a separately generated teacher trajectory, OPD evaluates the teacher distribution along the prefixes actually visited by the student. At decoding step t, the student and teacher distributions are respectively

$$p_{i,t}^{\text{S}} = \pi_{\theta_{\text{S}}}\left(\cdot \mid q_i, \hat{y}_{i,<t}^{\text{S}}\right), \qquad p_{i,t}^{\text{T}} = \pi_{\theta_{\text{T}}}\left(\cdot \mid q_i, \hat{y}_{i,<t}^{\text{S}}\right). \tag{9}$$

A typical OPD objective minimizes their token-level distribution discrepancy over student-generated trajectories:

$$\mathcal{L}_{\mathrm{OPD}} = \mathbb{E}_{\hat{y}_i^{\mathrm{S}} \sim \pi_{\theta_{\mathrm{S}}}} \left[\frac{1}{T} \sum_{t=1}^{T} D_{\mathrm{KL}} \left(p_{i,t}^{\mathrm{S}} \,\|\, \mathrm{sg}\left[p_{i,t}^{\mathrm{T}} \right] \right) \right], \tag{10}$$

where $\mathrm{sg}[\cdot]$ denotes the stop-gradient operator. Equivalently, for the sampled token $y_{i,t}^{\mathrm{S}}$, the teacher provides a token-wise learning signal through

$$\Delta_{i,t}^{\mathrm{OPD}} = \log \pi_{\theta_{\mathrm{T}}} \left(y_{i,t}^{\mathrm{S}} \mid q_i, \hat{y}_{i,<t}^{\mathrm{S}} \right) - \log \pi_{\theta_{\mathrm{S}}} \left(y_{i,t}^{\mathrm{S}} \mid q_i, \hat{y}_{i,<t}^{\mathrm{S}} \right). \tag{11}$$

On-policy self-distillation (OPSD) extends this formulation by constructing the teacher from the model itself rather than relying on an independent external model. Specifically, the teacher is provided with privileged information s_i, such as a ground-truth solution, that is unavailable to the student during generation. While the student distribution remains $\pi_{\theta_{\mathrm{S}}}(\cdot \mid q_i, \hat{y}_{i,<t}^{\mathrm{S}})$, the privileged teacher evaluates the same student-generated prefix as

$$p_{i,t}^{\mathrm{T}} = \pi_{\theta_{\mathrm{T}}} \left(\cdot \mid q_i, s_i, \hat{y}_{i,<t}^{\mathrm{S}} \right). \tag{12}$$

The resulting token-level self-distillation signal is therefore

$$\Delta_{i,t}^{\mathrm{OPSD}} = \log \pi_{\theta_{\mathrm{T}}} \left(y_{i,t}^{\mathrm{S}} \mid q_i, s_i, \hat{y}_{i,<t}^{\mathrm{S}} \right) - \log \pi_{\theta_{\mathrm{S}}} \left(y_{i,t}^{\mathrm{S}} \mid q_i, \hat{y}_{i,<t}^{\mathrm{S}} \right). \tag{13}$$

In this manner, OPSD transforms privileged solution information into dense supervision defined directly over the student's on-policy reasoning trajectory.

C.2 Monotonic Semantic Alignment

For each reasoning pattern, we first obtain a semantic representation by aggregating the hidden representations of its constituent tokens. Specifically, for the k-th teacher reasoning pattern $R_{i,k}^{\mathrm{T}}$ and the j-th student reasoning pattern $R_{i,j}^{\mathrm{S}}$, we define

$$\mathbf{z}_{i,k}^{\mathrm{T}} = \mathrm{Pool} \left(\left\{ \mathbf{h}_{i,t}^{\mathrm{T}} \mid o_{i,t}^{\mathrm{T}} \in R_{i,k}^{\mathrm{T}} \right\} \right), \tag{14}$$

and

$$\mathbf{z}_{i,j}^{\mathrm{S}} = \mathrm{Pool} \left(\left\{ \mathbf{h}_{i,t}^{\mathrm{S}} \mid o_{i,t}^{\mathrm{S}} \in R_{i,j}^{\mathrm{S}} \right\} \right), \tag{15}$$

where $\mathbf{h}_{i,t}^{\mathrm{T}}$ and $\mathbf{h}_{i,t}^{\mathrm{S}}$ denote the hidden representations of the corresponding tokens, and $\mathrm{Pool}(\cdot)$ denotes mean pooling.

We then measure the semantic similarity between each teacher–student reasoning-pattern pair using cosine similarity:

$$s_{i,k,j} = \mathrm{Cos} \left(\mathbf{z}_{i,k}^{\mathrm{T}}, \mathbf{z}_{i,j}^{\mathrm{S}} \right). \tag{16}$$

To preserve the sequential structure of the reasoning trajectories, the matching is performed under a monotonic constraint. Let ℓ_{k-1} denote the index of the most recently matched student reasoning pattern before $R_{i,k}^{\mathrm{T}}$. The candidate set for the current teacher reasoning pattern is defined as

$$\mathcal{A}_i(k) = \{ j \mid \ell_{k-1} \le j \le g \}, \tag{17}$$

with $\ell_0 = 1$. Among all admissible student reasoning patterns, we select the one with the maximum semantic similarity:

$$j_k^* = \arg \max_{j \in \mathcal{A}_i(k)} s_{i,k,j}. \tag{18}$$

Accordingly, for two successfully matched teacher reasoning patterns $k_1 < k_2$, their matched student indices satisfy

$$\phi_i(k_1) \le \phi_i(k_2), \tag{19}$$

thereby preserving their relative reasoning order.

However, selecting the most similar candidate does not necessarily imply that a meaningful semantic correspondence exists. In particular, a reasoning pattern introduced or substantially rewritten by the

teacher may have no valid counterpart in the student response. We therefore further examine the similarity of the best candidate using a threshold δ:

$$\phi_i(k) = \begin{cases} j_k^*, & s_{i,k,j_k^*} \geq \delta, \\ \varnothing, & s_{i,k,j_k^*} < \delta. \end{cases} \tag{20}$$

Thus, when the most similar student reasoning pattern still exhibits insufficient semantic similarity to the teacher pattern, we regard the corresponding reasoning pattern as absent from the student response rather than forcibly establishing an unreliable correspondence.

For a successful match, i.e., $\phi_i(k) = j_k^*$, we update $\ell_k = j_k^*$ and continue the monotonic search from this position. When $\phi_i(k) = \varnothing$, the monotonic boundary remains unchanged, i.e., $\ell_k = \ell_{k-1}$, allowing subsequent teacher reasoning patterns to continue searching from the latest valid correspondence.

D Proof of Theorem 3.1

We first provide the assumptions and auxiliary results required for Theorem 3.1. All auxiliary constructions introduced below are used only for theoretical analysis and do not modify the training procedure of VIRAL.

Theorem D.1. *Let q_i denote the problem, $\mathcal{D}_{i,t}^{S\to T}$ the normalized teacher–student discrepancy at reasoning pattern t, and*

$$\mathcal{D}_{i,t} = \mathbb{E}\big[\mathcal{D}_{i,t}^{S\to T} \mid q_i\big].$$

Let ϵ denote the prediction error of the auxiliary teacher-guided reference policy defined below. Then

$$\Pr\left(\hat{e}_i^S \neq e_i \mid q_i\right) \leq \epsilon + \min\left\{ \sqrt{z \log 2 \sum_{t\in\mathcal{C}_i} \mathcal{D}_{i,t}},\ \sqrt{1 - 4^{-z\sum_{t\in\mathcal{C}_i} \mathcal{D}_{i,t}}} \right\}. \tag{21}$$

D.1 Setting and Assumptions

Fix a problem q_i with ground-truth answer e_i. Let

$$y = (y_1, \ldots, y_L)$$

denote a student reasoning trajectory, and let $\hat{e}(y)$ denote the answer extracted from y. For token position k, define the student and privileged-teacher next-token distributions as

$$p_k^S = \pi_{\theta_S}(\cdot \mid q_i, y_{<k}), \qquad p_k^T = \pi_{\theta_T}(\cdot \mid q_i, c_i, y_{<k}), \tag{22}$$

where c_i denotes the privileged context available to the teacher.

We further define

$$m_k = \frac{1}{2}(p_k^S + p_k^T), \qquad J_k = D_{\mathrm{JS}}\big(p_k^T \,\|\, p_k^S\big). \tag{23}$$

For a reasoning pattern t containing z token positions, define

$$\mathcal{D}_{i,t}^{S\to T} = \frac{1}{z\log 2} \sum_{v=1}^{z} J_{i,t,v}, \tag{24}$$

and

$$\mathcal{D}_{i,t} = \mathbb{E}\left[\mathcal{D}_{i,t}^{S\to T} \mid q_i\right]. \tag{25}$$

Assumption D.2. At every token position considered in the theoretical analysis, the student and privileged-teacher next-token distributions are evaluated on the same student reasoning prefix $y_{<k}$. The teacher additionally observes the privileged context c_i.

Remark D.3. The analysis only requires that c_i be fixed for the corresponding problem instance and not depend on future tokens $y_{\geq k}$. Thus, the formulation covers both direct ground-truth conditioning and the student-anchored teacher refinement used by VIRAL. The role of Assumption D.2 is to ensure that $D_{\mathrm{JS}}(p_k^T \| p_k^S)$ measures a local distributional difference under the same reasoning state rather than a difference caused by distinct prefixes.

Assumption D.4. For a fixed problem q_i, the theoretical comparison is performed after the critical reasoning patterns have been identified. The selected set $\mathcal{C}_i$ and its pattern boundaries are held fixed when comparing the student and auxiliary reference policies.

Remark D.5. Assumption D.4 is used only to isolate the effect of teacher–student discrepancy for a given selected set. It does not imply that $\mathcal{C}_i$ is fixed during training. In practice, critical patterns are re-identified from each sampled trajectory.

Assumption D.6. Each selected reasoning pattern contains z tokens, trajectories have a finite maximum length L, and $\hat{e}(y)$ is a deterministic function of the generated trajectory.

The fixed pattern length is introduced only for notational simplicity. The same arguments extend to variable-length reasoning patterns by replacing z with the corresponding pattern length z_t.

D.2 Auxiliary Teacher-Guided Reference Policy

Let K_i denote the token positions belonging to the selected reasoning patterns $\mathcal{C}_i$. Define the auxiliary policy $\tilde{\pi}_i$ by

$$\tilde{p}_k = \begin{cases} m_k, & k \in K_i, \\ p_k^S, & k \notin K_i. \end{cases} \tag{26}$$

Thus, $\tilde{\pi}_i$ follows the student everywhere except on the selected reasoning patterns, where it uses the midpoint distribution between the student and the privileged teacher. This policy is introduced only for analysis.

Let P_S and $\tilde{P}$ denote the trajectory distributions induced by π_{θ_S} and $\tilde{\pi}_i$, respectively. Define

$$\epsilon = \Pr_{\tilde{P}}\big(\hat{e}(y) \neq e_i \mid q_i\big). \tag{27}$$

D.2.1 Auxiliary Lemmas

Lemma D.7 (Pattern-level discrepancy accumulation)**.** *Under Assumptions D.2–D.6,*

$$\mathbb{E}\left[\sum_{k\in K_i} J_k \,\middle|\, q_i\right] = z\log 2 \sum_{t\in\mathcal{C}_i} \mathcal{D}_{i,t}. \tag{28}$$

Proof. For each selected reasoning pattern t, Eq. equation 24 gives

$$\sum_{v=1}^{z} J_{i,t,v} = z\log 2\, \mathcal{D}_{i,t}^{S\to T}.$$

Summing over $t \in \mathcal{C}_i$,

$$\sum_{k\in K_i} J_k = z\log 2 \sum_{t\in\mathcal{C}_i} \mathcal{D}_{i,t}^{S\to T}.$$

Since $\mathcal{C}_i$ is fixed under Assumption D.4, taking conditional expectations and applying linearity of expectation yields

$$\mathbb{E}\left[\sum_{k\in K_i} J_k \,\middle|\, q_i\right] = z\log 2 \sum_{t\in\mathcal{C}_i} \mathbb{E}\left[\mathcal{D}_{i,t}^{S\to T} \mid q_i\right],$$

which equals Eq. equation 28. □

Lemma D.8 (Trajectory divergence bound)**.** *Under Assumptions D.2–D.6,*

$$D_{\mathrm{KL}}\big(P_S \,\|\, \tilde{P}\big) \le 2z\log 2 \sum_{t\in\mathcal{C}_i} \mathcal{D}_{i,t}. \tag{29}$$

Proof. By the autoregressive chain rule for KL divergence,

$$D_{\mathrm{KL}}\big(P_S \,\|\, \tilde{P}\big) = \mathbb{E}_{P_S}\left[\sum_{k=1}^{L} D_{\mathrm{KL}}\big(p_k^S \,\|\, \tilde{p}_k\big) \,\middle|\, q_i\right]. \tag{30}$$

The two policies coincide for $k \notin K_i$, hence

$$D_{\mathrm{KL}}\big(P_S \,\|\, \tilde{P}\big) = \mathbb{E}_{P_S}\left[\sum_{k\in K_i} D_{\mathrm{KL}}\big(p_k^S \,\|\, m_k\big) \;\middle|\; q_i\right]. \tag{31}$$

By definition of Jensen–Shannon divergence,

$$J_k = \frac{1}{2} D_{\mathrm{KL}}\big(p_k^S \,\|\, m_k\big) + \frac{1}{2} D_{\mathrm{KL}}\big(p_k^T \,\|\, m_k\big). \tag{32}$$

Since KL divergence is non-negative,

$$D_{\mathrm{KL}}\big(p_k^S \,\|\, m_k\big) \le 2J_k. \tag{33}$$

Therefore,

$$D_{\mathrm{KL}}\big(P_S \,\|\, \tilde{P}\big) \le 2\,\mathbb{E}\left[\sum_{k\in K_i} J_k \;\middle|\; q_i\right].$$

Applying Lemma D.7 proves Eq. equation 29. □

Lemma D.9 (Prediction-error difference)**.** *Let*

$$E_i = \{y : \hat{e}(y) \neq e_i\}.$$

Then

$$\Pr_{P_S}(E_i \mid q_i) \le \epsilon + D_{\mathrm{TV}}\big(P_S, \tilde{P}\big). \tag{34}$$

Proof. By the definition of total variation distance,

$$P_S(E_i \mid q_i) - \tilde{P}(E_i \mid q_i) \le D_{\mathrm{TV}}\big(P_S, \tilde{P}\big).$$

Since $\tilde{P}(E_i \mid q_i) = \epsilon$ by Eq. equation 27, the result follows. □

D.2.2 Proof of Theorem 3.1

Proof. By Lemma D.8,

$$D_{\mathrm{KL}}\big(P_S \,\|\, \tilde{P}\big) \le 2z \log 2 \sum_{t\in\mathcal{C}_i} \mathcal{D}_{i,t}.$$

Pinsker's inequality therefore gives

$$D_{\mathrm{TV}}\big(P_S, \tilde{P}\big) \le \sqrt{\frac{1}{2} D_{\mathrm{KL}}\big(P_S \,\|\, \tilde{P}\big)} \tag{35}$$

$$\le \sqrt{z \log 2 \sum_{t\in\mathcal{C}_i} \mathcal{D}_{i,t}}. \tag{36}$$

Meanwhile, the Bretagnolle–Huber inequality gives

$$D_{\mathrm{TV}}\big(P_S, \tilde{P}\big) \le \sqrt{1 - \exp\Big(-D_{\mathrm{KL}}(P_S \,\|\, \tilde{P})\Big)} \tag{37}$$

$$\le \sqrt{1 - \exp\left(-2z \log 2 \sum_{t\in\mathcal{C}_i} \mathcal{D}_{i,t}\right)} \tag{38}$$

$$= \sqrt{1 - 4^{-z\sum_{t\in\mathcal{C}_i} \mathcal{D}_{i,t}}}. \tag{39}$$

Combining Eqs. equation 36 and equation 39,

$$D_{\mathrm{TV}}\big(P_S, \tilde{P}\big) \le \min\left\{\sqrt{z \log 2 \sum_{t\in\mathcal{C}_i} \mathcal{D}_{i,t}},\ \sqrt{1 - 4^{-z\sum_{t\in\mathcal{C}_i} \mathcal{D}_{i,t}}}\right\}. \tag{40}$$

Finally, Lemma D.9 yields

$$\Pr\left(\hat{e}_i^S \neq e_i \mid q_i\right) \le \epsilon + D_{\mathrm{TV}}\big(P_S, \tilde{P}\big),$$

which establishes Eq. equation 21. □

D.3 SELECTION-AWARE ANALYSIS

Theorem 3.1 characterizes the effect of teacher–student discrepancy for an arbitrary fixed set of selected reasoning patterns $\mathcal{C}_i$. It does not, by itself, specify which patterns should be selected. We now analyze this question separately.

Our analysis proceeds in two stages. First, under an ideal correctness-conditioned teacher, we show directly that teacher guidance produces non-negative improvement and that the gain is concentrated at reasoning states whose continuation affects final correctness. Second, for the practical information-gain score used by VIRAL, we state an explicit and empirically testable consistency condition relating the score to the true correction gain. Under this condition, high-information patterns maximize a guaranteed first-order improvement lower bound.

Importantly, the second stage is a conditional guarantee rather than an assumption-free identification of the practical information-gain score with the true correction gain.

D.3.1 CORRECTION GAIN OF TEACHER GUIDANCE

For a student prefix $y_{<k}$, define the student's continuation success probability

$$V^S(y_{<k}) = \Pr_{y' \sim \pi_{\theta_S}} (\hat{e}(y') = e_i \mid q_i,\, y'_{<k} = y_{<k})\,. \tag{41}$$

For a complete trajectory y, define

$$V^S(y) = \mathbb{1}[\hat{e}(y) = e_i].$$

Let

$$\mathrm{Acc}(\pi) = \Pr_{y \sim \pi} \big(\hat{e}(y) = e_i \mid q_i\big).$$

Then

$$\mathrm{Acc}(\pi_{\theta_S}) = V^S(\emptyset).$$

Definition D.10 (Token-level correction gain)**.** For a prefix $y_{<k}$, define

$$\Delta_k(y_{<k}) = \mathbb{E}_{a \sim p_k^T}\left[V^S(y_{<k} \circ a)\right] - V^S(y_{<k}). \tag{42}$$

Thus, $\Delta_k(y_{<k})$ measures the change in the student's expected final correctness when the next token is sampled from the privileged teacher rather than the student, while all subsequent tokens are evaluated using the student's continuation value.

For a reasoning pattern t, define its expected correction gain as

$$\Gamma_{i,t} = \mathbb{E}_{y \sim P_S}\left[\sum_{v=1}^{z} \Delta_{(t,v)}\big(y_{<(t,v)}\big) \;\middle|\; q_i\right]. \tag{43}$$

For $\alpha \in [0,1]$, define the analytical interpolation policy $\pi_{\alpha,\mathcal{C}_i}$ by

$$p_k^\alpha = \begin{cases} (1-\alpha)p_k^S + \alpha p_k^T, & k \in K_i, \\ p_k^S, & k \notin K_i. \end{cases} \tag{44}$$

Let P_α denote its trajectory distribution. Note that

$$\pi_{0,\mathcal{C}_i} = \pi_{\theta_S}, \qquad \pi_{1/2,\mathcal{C}_i} = \tilde{\pi}_i.$$

The interpolation policy is used only to characterize the local direction of teacher-guided optimization. The actual VIRAL objective realizes teacher guidance through advantage modulation rather than explicit policy mixing.

Lemma D.11 (Exact teacher-guided performance decomposition)**.** *For every $\alpha \in [0,1]$,*

$$\mathrm{Acc}\big(\pi_{\alpha,\mathcal{C}_i}\big) - \mathrm{Acc}\big(\pi_{\theta_S}\big) = \alpha\, \mathbb{E}_{y \sim P_\alpha}\left[\sum_{k \in K_i} \Delta_k(y_{<k})\right]. \tag{45}$$

Moreover,

$$\frac{d}{d\alpha}\mathrm{Acc}\big(\pi_{\alpha,\mathcal{C}_i}\big)\bigg|_{\alpha=0} = \sum_{t \in \mathcal{C}_i} \Gamma_{i,t}. \tag{46}$$

Proof. For any complete trajectory y, telescoping gives

$$\mathbb{1}[\hat{e}(y) = e_i] - V^S(\emptyset) = \sum_{k=1}^{L} \left[V^S(y_{\leq k}) - V^S(y_{<k}) \right]. \tag{47}$$

Taking expectation under P_α and conditioning on $y_{<k}$,

$$\mathbb{E}_{P_\alpha} \left[V^S(y_{\leq k}) \mid y_{<k} \right] = \sum_a p_k^\alpha(a) V^S(y_{<k} \circ a). \tag{48}$$

By definition of V^S,

$$\sum_a p_k^S(a) V^S(y_{<k} \circ a) = V^S(y_{<k}). \tag{49}$$

Hence, for $k \in K_i$,

$$\sum_a p_k^\alpha(a) V^S(y_{<k} \circ a) = (1-\alpha) V^S(y_{<k}) + \alpha \sum_a p_k^T(a) V^S(y_{<k} \circ a) \tag{50}$$

$$= V^S(y_{<k}) + \alpha \Delta_k(y_{<k}), \tag{51}$$

while for $k \notin K_i$ the conditional difference is zero.

Substituting into Eq. equation 47 and summing over k proves Eq. equation 45.

For the derivative at $\alpha = 0$, divide Eq. equation 45 by α. Since the vocabulary and horizon are finite, P_α converges to P_S as $\alpha \to 0$. Moreover, the summand is bounded. Therefore,

$$\frac{d}{d\alpha} \mathrm{Acc}\big(\pi_{\alpha,\mathcal{C}_i}\big)\Big|_{\alpha=0} = \mathbb{E}_{P_S} \left[\sum_{k \in K_i} \Delta_k(y_{<k}) \right] \tag{52}$$

$$= \sum_{t \in \mathcal{C}_i} \Gamma_{i,t}, \tag{53}$$

which proves Eq. equation 46. □

D.3.2 Ideal Correctness-Conditioned Teacher

We next characterize the correction gain when the teacher coincides with the student's posterior distribution conditioned on eventual correctness.

Assumption D.12 (Ideal correctness-conditioned teacher)**.** For every prefix with $V^S(y_{<k}) > 0$,

$$p_k^T(a) = \Pr_{\pi_{\theta_S}} \left(y_k = a \mid q_i,\, y_{<k},\, \hat{e}(y) = e_i \right). \tag{54}$$

For prefixes with $V^S(y_{<k}) = 0$, define $p_k^T = p_k^S$.

Lemma D.13 (Correctness conditioning yields non-negative gain)**.** *Under Assumption D.12, let*

$$V = V^S(y_{<k}), \qquad V_a = V^S(y_{<k} \circ a).$$

For every prefix with $V > 0$,

$$\Delta_k(y_{<k}) = \frac{\mathrm{Var}_{a \sim p_k^S}[V_a]}{V} \geq V D_{\mathrm{KL}}\big(p_k^T \,\|\, p_k^S\big) \geq 0. \tag{55}$$

For $V = 0$, $\Delta_k(y_{<k}) = 0$ by construction.

Proof. By Bayes' rule,

$$p_k^T(a) = p_k^S(a) \frac{V_a}{V}. \tag{56}$$

Therefore,

$$\Delta_k = \sum_a \big(p_k^T(a) - p_k^S(a)\big)V_a \tag{57}$$

$$= \frac{1}{V}\sum_a p_k^S(a)(V_a - V)V_a. \tag{58}$$

Using

$$\sum_a p_k^S(a)V_a = V,$$

we obtain

$$\Delta_k = \frac{1}{V}\sum_a p_k^S(a)(V_a - V)^2 \tag{59}$$

$$= \frac{\mathrm{Var}_{a\sim p_k^S}[V_a]}{V} \geq 0. \tag{60}$$

To obtain the KL lower bound, note that

$$x - y \geq y \log \frac{x}{y}, \qquad x, y > 0.$$

Applying this inequality with $x = V_a$ and $y = V$ gives

$$\Delta_k = \sum_a p_k^T(a)(V_a - V) \tag{61}$$

$$\geq V \sum_a p_k^T(a) \log \frac{V_a}{V}. \tag{62}$$

By Eq. equation 56,

$$\frac{V_a}{V} = \frac{p_k^T(a)}{p_k^S(a)}$$

whenever $p_k^T(a) > 0$. Hence,

$$\Delta_k \geq V \sum_a p_k^T(a) \log \frac{p_k^T(a)}{p_k^S(a)} \tag{63}$$

$$= V D_{\mathrm{KL}}\big(p_k^T \,\|\, p_k^S\big), \tag{64}$$

which completes the proof. □

Theorem D.14 (Policy improvement under an ideal correctness-conditioned teacher)**.** *Under Assumption D.12, for any fixed selected set $\mathcal{C}_i$ and any $\alpha \in [0, 1]$,*

$$\mathrm{Acc}\big(\pi_{\alpha,\mathcal{C}_i}\big) \geq \mathrm{Acc}\big(\pi_{\theta_S}\big). \tag{65}$$

The inequality is strict whenever P_α assigns positive probability to a selected prefix for which

$$\mathrm{Var}_{a\sim p_k^S}[V_a] > 0.$$

Proof. By Lemma D.13,

$$\Delta_k(y_{<k}) \geq 0$$

for every prefix. Applying Lemma D.11,

$$\mathrm{Acc}\big(\pi_{\alpha,\mathcal{C}_i}\big) - \mathrm{Acc}\big(\pi_{\theta_S}\big) = \alpha \mathbb{E}_{P_\alpha}\left[\sum_{k\in K_i} \Delta_k(y_{<k})\right] \geq 0.$$

Strict improvement follows whenever the integrand is strictly positive with non-zero probability. □

Remark D.15 (Where does the improvement occur?). Under the ideal teacher,

$$D_{\mathrm{KL}}\big(p_k^T \,\|\, p_k^S\big) = D_{\mathrm{KL}}\left(p(y_k \mid y_{<k}, \hat{e} = e_i) \,\|\, p(y_k \mid y_{<k})\right). \tag{66}$$

This quantity measures how much conditioning on eventual correctness changes the distribution of the next reasoning step. Lemma D.13 shows that such correctness-sensitive positions have non-negative correction gain.

The variance representation provides an equivalent interpretation: Δ_k is large when different plausible next tokens lead to substantially different probabilities of eventual success. Conversely, if $V_a \approx V$ for all plausible next tokens, then $\Delta_k \approx 0$. Thus the ideal analysis identifies correctness-sensitive branching points as the positions at which teacher guidance is useful.

D.3.3 Connecting Practical Information Gain to Correction Gain

The previous result is assumption-free with respect to pattern selection once the teacher is taken to be the ideal correctness-conditioned posterior. However, the practical information-gain score used by VIRAL is not identical by definition to Eq. equation 66. We therefore make the relationship between the practical score and the true correction gain explicit rather than assuming that the two quantities are interchangeable.

For each candidate student reasoning pattern t, let $\varsigma_{i,t}$ denote the practical information-gain score used by VIRAL:

$$\varsigma_{i,t} = \begin{cases} \mathrm{IG}_{\mathrm{truth}}\big(R^T_{i,\psi(t)}\big), & t \in \hat{\mathcal{C}}_i^S, \\ \mathrm{IG}\big(R^S_{i,t}\big), & t \in \mathcal{C}_i^S, \end{cases} \tag{67}$$

where $\psi(t)$ denotes the aligned teacher pattern.

Assumption D.16 (Information-gain–gain consistency)**.** There exist constants $\gamma > 0$, $\bar{\gamma} \geq \gamma$, and $\eta \geq 0$ such that, for each candidate pattern t,

$$\gamma \varsigma_{i,t} - \eta \leq \Gamma_{i,t} \leq \bar{\gamma}[\varsigma_{i,t}]_+ + \eta. \tag{68}$$

Remark D.17. Assumption D.16 is the only additional condition linking the practical selection score of VIRAL to the true correction gain. Theorem D.1, Lemma D.11, and Theorem D.14 do not depend on this assumption.

The assumption is empirically testable: $\Gamma_{i,t}$ can be estimated through Monte-Carlo continuations from the corresponding prefix, while $\varsigma_{i,t}$ is directly computed by the selection mechanism. We provide such an empirical verification in Appendix D.3.5.

Theorem D.18 (First-order guarantee for information-gain-based selection)**.** *Under Assumptions D.2–D.6 and D.16, for any selected set $\mathcal{C}_i$,*

$$\frac{d}{d\alpha}\mathrm{Acc}\big(\pi_{\alpha,\mathcal{C}_i}\big)\Big|_{\alpha=0} = \sum_{t\in\mathcal{C}_i} \Gamma_{i,t} \geq \sum_{t\in\mathcal{C}_i} \left(\gamma \varsigma_{i,t} - \eta\right). \tag{69}$$

Proof. The equality follows directly from Eq. equation 46. Applying the lower bound in Assumption D.16 to each selected pattern gives

$$\sum_{t\in\mathcal{C}_i} \Gamma_{i,t} \geq \sum_{t\in\mathcal{C}_i} \left(\gamma \varsigma_{i,t} - \eta\right),$$

which proves the result. □

Corollary D.19 (Consequences for pattern selection)**.** *Under the assumptions of Theorem D.18, the following statements hold.*

(i) ***Thresholded high-information patterns have positive first-order gain.*** *If every selected pattern satisfies*

$$\varsigma_{i,t} \geq \tau$$

for some

$$\tau > \frac{\eta}{\gamma},$$

then

$$\frac{d}{d\alpha}\mathrm{Acc}\big(\pi_{\alpha,\mathcal{C}_i}\big)\Big|_{\alpha=0} \geq |\mathcal{C}_i|(\gamma\tau - \eta) > 0. \tag{70}$$

Consequently, by differentiability, there exists $\alpha_0 > 0$ such that

$$\mathrm{Acc}\big(\pi_{\alpha,\mathcal{C}_i}\big) > \mathrm{Acc}\big(\pi_{\theta_S}\big)$$

for all sufficiently small $\alpha \in (0, \alpha_0)$.

(ii) ***Top-K information-gain selection maximizes the guaranteed first-order lower bound.*** *Among all candidate sets satisfying $|\mathcal{C}_i| = K$, the lower bound*

$$\sum_{t\in\mathcal{C}_i}(\gamma\varsigma_{i,t} - \eta) \tag{71}$$

is maximized by selecting the K patterns with the largest information-gain scores.

(iii) ***Non-positive-information patterns have only limited guaranteed benefit.*** *If $\mathcal{C}_i' \supseteq \mathcal{C}_i$ and every newly added pattern satisfies $\varsigma_{i,t} \leq 0$, then*

$$\sum_{t\in\mathcal{C}_i'\setminus\mathcal{C}_i}\Gamma_{i,t} \leq \eta\,|\mathcal{C}_i' \setminus \mathcal{C}_i|. \tag{72}$$

Thus, under Assumption D.16, low-information patterns are not guaranteed to provide a correction gain comparable to high-information patterns.

Proof. For (i), Theorem D.18 gives

$$\frac{d}{d\alpha}\mathrm{Acc}\big(\pi_{\alpha,\mathcal{C}_i}\big)\Big|_{\alpha=0} \geq \sum_{t\in\mathcal{C}_i}(\gamma\tau - \eta),$$

which equals $|\mathcal{C}_i|(\gamma\tau - \eta) > 0$. A positive derivative at $\alpha = 0$, together with differentiability, implies strict improvement for sufficiently small positive α.

For (ii), with $|\mathcal{C}_i| = K$ fixed,

$$\sum_{t\in\mathcal{C}_i}(\gamma\varsigma_{i,t} - \eta) = \gamma\sum_{t\in\mathcal{C}_i}\varsigma_{i,t} - K\eta.$$

Since $\gamma > 0$ and $K\eta$ is constant across all size-K sets, the expression is maximized by the K largest values of $\varsigma_{i,t}$.

For (iii), the upper bound in Assumption D.16 gives, for every newly added pattern with $\varsigma_{i,t} \leq 0$,

$$\Gamma_{i,t} \leq \bar{\gamma}[\varsigma_{i,t}]_+ + \eta = \eta.$$

Summing over the added patterns proves Eq. equation 72. □

D.3.4 Relation to Theorem 3.1

The two analyses characterize complementary aspects of VIRAL.

Theorem 3.1 considers an arbitrary fixed selected set $\mathcal{C}_i$ and shows that the gap between the current student and the teacher-guided reference policy is controlled by

$$\sum_{t\in\mathcal{C}_i}\mathcal{D}_{i,t}.$$

Thus, once the relevant reasoning patterns have been selected, reducing teacher–student discrepancy on those patterns tightens the student-error bound.

The selection-aware analysis instead addresses where teacher guidance should be applied. Under the ideal correctness-conditioned teacher, Theorem D.14 proves that teacher guidance is non-degrading and yields strict improvement at correctness-sensitive branching points. For the practical

information-gain score used by VIRAL, Theorem D.18 gives a conditional first-order guarantee under the empirically testable information-gain–gain consistency condition.

Therefore, the theoretical roles of the two components are distinct: information gain determines *where* supervision is preferentially allocated, whereas the discrepancy term determines *how strongly* the selected student reasoning patterns remain aligned with the privileged teacher.

We emphasize that $\pi_{\alpha,\mathcal{C}_i}$ is an analytical interpolation policy. The actual VIRAL optimization does not explicitly mix the student and teacher distributions. Rather, its advantage modulation encourages the same local preference: stronger positive supervision is assigned to selected patterns that are both correctness-relevant and sufficiently consistent with the teacher. Accordingly, the analysis provides a local theoretical justification for the optimization direction encouraged by VIRAL, rather than an exact equivalence between the analytical interpolation and the implemented policy update.

D.3.5 Empirical Verification of the Information-Gain–Gain Consistency Condition

Assumption D.16 concerns the relationship between the practical information-gain score $\varsigma_{i,t}$ and the correction gain $\Gamma_{i,t}$. We therefore evaluate this relationship directly.

We sample 1000 DeepMath problems and obtain one student rollout per problem using Qwen2.5-Bespoke-7B. Each rollout is segmented into reasoning patterns. For every candidate pattern t, we estimate its correction gain by Monte-Carlo continuation.

Specifically, from the student prefix immediately preceding pattern t, we construct two alternatives: one in which the pattern is generated using the privileged teacher and one in which it is generated using the student. From the resulting prefixes, we sample $M = 16$ student continuations and measure the difference in empirical correctness. This yields an estimate $\hat{\Gamma}_{i,t}$.

We then compare $\hat{\Gamma}_{i,t}$ against the practical information-gain score $\varsigma_{i,t}$. Table 9 reports (a) the average estimated correction gain across information-gain deciles and (b) the accuracy of the analytical teacher-guided reference under top-K, random-K, and bottom-K pattern selection with matched K. A monotonic increase of $\hat{\Gamma}_{i,t}$ with $\varsigma_{i,t}$, together with superior top-K performance, provides direct empirical support for the consistency condition used in Theorem D.18.

Table 9: Empirical verification of the information-gain–gain consistency condition on 1,000 DeepMath problems using Qwen2.5-Bespoke-7B. Panel (a) reports the practical information-gain score $\varsigma_{i,t}$ and estimated correction gain $\hat{\Gamma}_{i,t}$ across information-gain deciles. Panel (b) compares the accuracy (%) of the teacher-guided reference under top-K, random-K, and bottom-K pattern selection with matched K.

(a) Information gain vs. estimated correction gain

IG Decile	Mean $\varsigma_{i,t}$	Mean $\hat{\Gamma}_{i,t}$
1	-0.081	-0.003
2	-0.043	-0.001
3	-0.018	0.001
4	0.006	0.004
5	0.029	0.008
6	0.057	0.014
7	0.091	0.022
8	0.137	0.033
9	0.214	0.049
10	0.361	0.071

(b) Teacher-guided reference accuracy under different selections

Selection	Number of selected patterns K				Avg.
	2	4	6	8	
Bottom-K	56.5	56.7	56.8	56.9	56.7
Random-K	57.8	59.1	60.0	60.6	59.4
Top-K	**60.2**	**62.4**	**64.0**	**65.1**	**62.9**

Note. $\hat{\Gamma}_{i,t}$ is estimated using $M = 16$ Monte-Carlo student continuations from teacher- and student-generated alternatives at each candidate reasoning pattern. All selection strategies use the same number K of reasoning patterns.

E VISUALIZATION AND EXAMPLES OF REWARD AND PENALTY

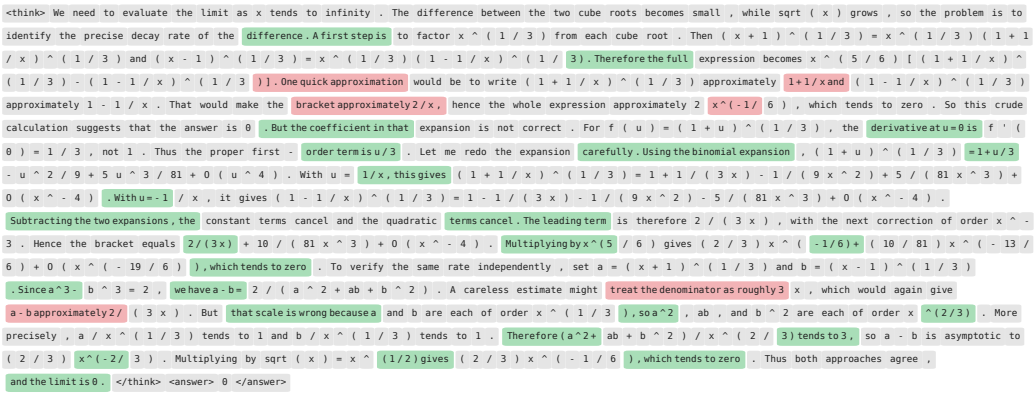


Figure 8: Visualization of token-level reward and penalty signals during reasoning. Green and red highlights denote positively and negatively adjusted reasoning tokens, respectively.

F PROMPT OF TEACHER MODEL

Privileged Teacher Prompt

```
Problem: {question}

Verified correct answer/reference solution:
{ref_solution}

You are a privileged reasoning teacher. The student's current
reasoning prefix will be provided below. Please refine the reasoning
based on this prefix: preserve the valid reasoning structure,
correct or supplement any erroneous or incomplete parts, and continue
improving the trajectory so that the final reasoning is consistent
with the verified correct answer. Do not restart the solution from
scratch or ignore the student's existing reasoning prefix.

Student reasoning prefix begins below:
{student_solution}
```

Figure 9: Prompt template for the privileged teacher.

F.1 EXAMPLE OF VIRAL

Example 1: Number Theory

Student

Prompt

`Problem: Find all integer solutions` (x, y, z) `to`

$$x^2 + y^2 = 3z^2.$$

`Please reason step by step, and put your final answer within \boxed{}.`

Response

Since squares modulo 3 can only be 0 or 1,

$$x^2 + y^2 \equiv 0 \pmod 3$$

implies that both x and y are divisible by 3.
Let

$$x = 3a, \qquad y = 3b.$$

Then

$$9a^2 + 9b^2 = 3z^2,$$

so

$$z^2 = 3(a^2 + b^2).$$

Hence z is also divisible by 3.
...
Thus any nonzero solution produces another solution of the same form with all variables divided by 3. Repeating this gives an infinite descent, which is impossible for nonzero integers. Therefore,

$$\boxed{(x, y, z) = (0, 0, 0)}.$$

Privileged Teacher

Prompt

Problem: Find all integer solutions (x, y, z) to

$$x^2 + y^2 = 3z^2.$$

Verified correct answer/reference solution:

Verified Reference Solution

Consider the equation modulo 3. Since the quadratic residues modulo
...
The new equation has the same form as the original one. Hence any nontrivial solution would generate an infinite descent. Therefore,

$$\boxed{(x, y, z) = (0, 0, 0)}.$$

You are a privileged reasoning teacher. The student's current reasoning prefix will be provided below. Please refine the reasoning based on this prefix: preserve the valid reasoning structure, correct or supplement any erroneous or incomplete parts, and continue improving the trajectory so that the final reasoning is consistent with the verified correct answer. Do not restart the solution from scratch or ignore the student's existing reasoning prefix.

Student reasoning prefix begins below:

Student Reasoning Prefix

Since squares modulo 3 can only be 0 or 1,

$$x^2 + y^2 \equiv 0 \pmod 3,$$

so both x and y must be divisible by 3.
Let

$$x = 3a, \qquad y = 3b.$$

Then

$$9a^2 + 9b^2 = 3z^2.$$

...

Response

Dividing by 3 gives

$$3(a^2 + b^2) = z^2.$$

Therefore $3 \mid z$, so write $z = 3c$. We obtain

$$a^2 + b^2 = 3c^2,$$

which has exactly the same structure as the original equation.
...
Thus a nonzero solution would generate an infinite sequence of solutions divisible by arbitrarily high powers of 3, which is impossible for nonzero integers. Hence,

$$\boxed{(x, y, z) = (0, 0, 0)}.$$

Example 2: Stochastic Calculus

Student

Prompt

Problem: Find the value of $\beta \in \mathbb{R}$ for which

$$2W_t^3 + \beta t W_t$$

is a martingale, where W_t is a standard Wiener process.
Please reason step by step, and put your final answer within \boxed{}.

Response

Let

$$X_t = 2W_t^3 + \beta t W_t.$$

For X_t to be a martingale, the drift term should vanish.
Using Itô's formula,

$$d(W_t^3) = 3W_t^2\, dW_t + 3W_t\, dt,$$

so

$$d(2W_t^3) = 6W_t^2\, dW_t + 6W_t\, dt.$$

Also,

$$d(tW_t) = W_t\, dt + t\, dW_t.$$

Therefore,

$$dX_t = (6W_t^2 + \beta t)dW_t + (6 + \beta)W_t dt.$$

...
The drift vanishes when

$$6 + \beta = 0.$$

Hence,

$$\boxed{\beta = -6}.$$

Privileged Teacher

Prompt

Problem: Find the value of $\beta \in \mathbb{R}$ for which

$$2W_t^3 + \beta t W_t$$

is a martingale, where W_t is a standard Wiener process.

Verified correct answer/reference solution:

Verified Reference Solution

Let

...

The drift must vanish, so

$$6 + \beta = 0.$$

Therefore,

$$\boxed{\beta = -6}.$$

You are a privileged reasoning teacher. The student's current reasoning prefix will be provided below. Please refine the reasoning based on this prefix: preserve the valid reasoning structure, correct or supplement any erroneous or incomplete parts, and continue improving the trajectory so that the final reasoning is consistent with the verified correct answer. Do not restart the solution from scratch or ignore the student's existing reasoning prefix.

Student reasoning prefix begins below:

Student Reasoning Prefix

Let

$$X_t = 2W_t^3 + \beta t W_t.$$

For X_t to be a martingale, its drift term should vanish.
Using Itô's formula,

$$d(2W_t^3) = 6W_t^2 dW_t + 6W_t dt.$$

Also,

$$d(\beta t W_t) = \beta t dW_t + \beta W_t dt.$$

...

Response

Combining the two terms gives

$$dX_t = (6W_t^2 + \beta t)dW_t + (6 + \beta)W_t dt.$$

Thus the finite-variation component is

$$(6 + \beta)W_t dt.$$

...
For this drift to vanish,

$$6 + \beta = 0.$$

Hence,

$$\boxed{\beta = -6}.$$

Example 3: Telescoping Series

Student

Prompt

Problem: Let

$$f(x) = \lim_{N\to\infty} \sum_{n=1}^{N} \arctan\left(\frac{x}{n(n+1)+x^2}\right).$$

Find $f'(2)$.
Please reason step by step, and put your final answer within \boxed{}.

Response

The denominator suggests using the tangent subtraction identity. Let

$$A_n = \arctan\left(\frac{x}{n}\right), \qquad B_n = \arctan\left(\frac{x}{n+1}\right).$$

Then

$$\tan(A_n - B_n) = \frac{\frac{x}{n} - \frac{x}{n+1}}{1 + \frac{x^2}{n(n+1)}} = \frac{x}{n(n+1)+x^2}.$$

Therefore,

$$\arctan\left(\frac{x}{n(n+1)+x^2}\right) = \arctan\left(\frac{x}{n}\right) - \arctan\left(\frac{x}{n+1}\right).$$

Hence the sum telescopes:

$$f_N(x) = \arctan(x) - \arctan\left(\frac{x}{N+1}\right).$$

...
Taking $N \to \infty$ gives

$$f(x) = \arctan(x),$$

so

$$f'(x) = \frac{1}{1+x^2}.$$

Therefore,

$$\boxed{f'(2) = \frac{1}{5}}.$$

Privileged Teacher

Prompt

Problem: Let

$$f(x) = \lim_{N\to\infty} \sum_{n=1}^{N} \arctan\left(\frac{x}{n(n+1)+x^2}\right).$$

Find $f'(2)$.

Verified correct answer/reference solution:

Verified Reference Solution

Using

$$\arctan\left(\frac{x}{n(n+1)+x^2}\right) = \arctan\left(\frac{x}{n}\right) - \arctan\left(\frac{x}{n+1}\right),$$

the finite sum telescopes:

$$f_N(x) = \arctan(x) - \arctan\left(\frac{x}{N+1}\right).$$

...

Taking $N \to \infty$ gives

$$f(x) = \arctan(x), \qquad f'(x) = \frac{1}{1+x^2}.$$

Hence,

$$\boxed{f'(2) = \frac{1}{5}}.$$

You are a privileged reasoning teacher. The student's current reasoning prefix will be provided below. Please refine the reasoning based on this prefix: preserve the valid reasoning structure, correct or supplement any erroneous or incomplete parts, and continue improving the trajectory so that the final reasoning is consistent with the verified correct answer. Do not restart the solution from scratch or ignore the student's existing reasoning prefix.

Student reasoning prefix begins below:

Student Reasoning Prefix

The denominator suggests using the tangent subtraction identity. Let

$$A_n = \arctan\left(\frac{x}{n}\right), \qquad B_n = \arctan\left(\frac{x}{n+1}\right).$$

Then

$$\tan(A_n - B_n) = \frac{\frac{x}{n} - \frac{x}{n+1}}{1 + \frac{x^2}{n(n+1)}} = \frac{x}{n(n+1)+x^2}.$$

Therefore,

$$\arctan\left(\frac{x}{n(n+1)+x^2}\right) = \arctan\left(\frac{x}{n}\right) - \arctan\left(\frac{x}{n+1}\right).$$

...

Response

Thus the partial sum telescopes to

$$f_N(x) = \arctan(x) - \arctan\left(\frac{x}{N+1}\right).$$

Since

$$\arctan\left(\frac{x}{N+1}\right) \to 0 \qquad \text{as } N \to \infty,$$

we obtain

$$f(x) = \arctan(x).$$

...
Therefore,

$$f'(x) = \frac{1}{1+x^2},$$

and evaluating at $x = 2$ gives

$$\boxed{f'(2) = \frac{1}{5}}.$$