DUAL-CURVATURE IMPLICIT REGULARIZATION FOR GRPO VIA SHARPNESS-AWARE MINIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Group Relative Policy Optimization (GRPO) has become a widely used algorithm for reinforcement learning with verifiable rewards (RLVR). Common training recipes adopt small learning rates and extended training schedules, while increasing the learning rate can degrade test performance. To investigate which training quantities are associated with generalization performance, we conduct a controlled 20-seed study of GRPO under a fixed training configuration. The strongest associations with test accuracy are observed among the directional Hessian and Fisher curvature metrics, whereas training reward and gradient norm show weak correlations. Motivated by these findings, we propose KL-SAM, a fully first-order optimization algorithm that combines Sharpness-Aware Minimization (SAM)-style perturbations with a local policy-KL regularizer to target objective and policy sensitivity through a single shared perturbation. A local expansion reveals directional Hessian and Fisher contributions, providing an interpretation as implicit dual-curvature regularization. Experiments on models ranging from 1.5B to 7B parameters across multiple objectives and benchmarks demonstrate the effectiveness of our approach. We will release our code upon publication.

1 INTRODUCTION

Large language models (LLMs) have demonstrated strong potential in mathematical reasoning, code generation, and complex problem solving. Reinforcement learning with verifiable rewards (RLVR) has emerged as a scalable post-training paradigm for improving these capabilities by using automatic signals such as answer matching, unit tests, or formal verifiers as feedback (Lambert et al., 2024). Group Relative Policy Optimization (GRPO) has become a widely used RLVR algorithm for reasoning-oriented LLM post-training, estimating relative advantages from multiple responses sampled for the same prompt. DeepSeekMath introduced GRPO for mathematical reasoning (Shao et al., 2024), and DeepSeek-R1 further demonstrated the effectiveness of large-scale RLVR with GRPO for eliciting advanced reasoning behaviors (Guo et al., 2025).

Despite these advances, improving performance on unseen problems within a limited training budget remains challenging. Strong training rewards need not translate into generalization gains in either RL (Zhang et al., 2018; Cobbe et al., 2019) or RLVR (Wang et al., 2025). GRPO-based training commonly uses small learning rates such as 10^{-6}, with some recipes running for a thousand or more update steps (Liu et al., 2025; Yu et al., 2025; Zheng et al., 2025; Le et al., 2025). Increasing the learning rate can destabilize training or degrade test performance (Rajani et al., 2025; Lee et al., 2026). This motivates regularization that supports generalization at larger learning rates within limited training budgets.

Prior studies have connected local geometry to optimization stability, robustness, and generalization in reinforcement learning (Lyle et al., 2022; Hosseinkhan-Boucher et al., 2025; Lee & Yoon, 2025). Within RLVR, CAPO studies directional objective and policy curvature for update stability (Melo et al., 2025), while GRPO-SG motivates generalization-oriented training through a gradient-norm-based sharpness surrogate (Le et al., 2025). However, direct empirical evidence relating these quantities to generalization differences across independent GRPO runs remains limited. This gap matters for larger-learning-rate regularization: quantities associated with stable optimization need not identify the local sensitivities most relevant to generalization.

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.

We address this issue through a controlled 20-seed study of GRPO under a fixed training configuration, comparing training reward, KL divergence to the reference policy, gradient norm, and directional Hessian and Fisher curvature. The strongest correlations with test accuracy are observed among the directional curvature quantities, which exhibit negative associations, whereas the other metrics show weaker or less consistent associations across Pearson and Spearman correlations.

Motivated by these findings, we propose KL-SAM, which regularizes both objective and policy sensitivity through a single shared perturbation computed from first-order gradients. A local expansion shows that the SAM-perturbed loss (Foret et al., 2021) contributes a directional Hessian term, while a clean-to-perturbed policy KL penalty contributes a directional Fisher term along the same perturbation. This yields an interpretation as *dual-curvature implicit regularization*, without explicit curvature estimation or separate perturbations for the two components. For efficiency, we adopt the output-layer perturbation strategy of logits-SAM (Luo et al., 2026) and augment it with local policy-KL regularization. Experiments across mathematical reasoning models ranging from 1.5B to 7B parameters and GRPO variants show improvements in average test pass@1 over standard training and logits-SAM with modest computational overhead.

Our contributions can be summarized as follows:

- We conduct a controlled 20-seed study of GRPO training, characterizing the associations between directional Hessian and Fisher curvature and generalization performance, with comparisons against standard training metrics.
- We propose KL-SAM, which implements implicit dual-curvature regularization through a single shared perturbation using only first-order gradients. We characterize its local Hessian–Fisher expansion and provide an information-theoretic perspective on policy sensitivity under changes in the training data and its relation to generalization.
- We evaluate KL-SAM across mathematical reasoning models ranging from 1.5B to 7B parameters and GRPO variants, showing improvements in average test pass@1 over standard training and logits-SAM with modest computational overhead.

2 Related Work

RL and LLM reasoning. Reinforcement learning (RL) has become a central tool for post-training optimization of large language models. Early RLHF approaches optimize a learned reward model based on human preferences, typically using policy gradient algorithms such as PPO (Christiano et al., 2017; Stiennon et al., 2020; Ouyang et al., 2022; Schulman et al., 2017). In recent years, reinforcement learning with verifiable rewards (RLVR) has improved reasoning capabilities by replacing learned preference rewards with objective, automatically verifiable signals, such as mathematical answer correctness or code execution outcomes (Lambert et al., 2024; Guo et al., 2025; Team et al., 2025). GRPO further simplifies PPO-style optimization by estimating advantages over groups of completions sampled from the same prompt, thereby eliminating the need for a separate value model while retaining clipped policy updates (Shao et al., 2024). DAPO introduces dynamic sampling (Yu et al., 2025). DR.GRPO corrects length bias (Liu et al., 2025). VAPO improves performance through value-based policy optimization (Yue et al., 2025). GRESO improves rollout efficiency through selective sampling (Zheng et al., 2025). FRPO prevents catastrophic forgetting by optimizing rewards over policies in reachable KL-bounded neighborhoods (Sabbaghi et al., 2026).

CAPO uses directional Hessian and Fisher estimates to mask tokens associated with unstable updates (Melo et al., 2025). GRPO-SG motivates token reweighting through a gradient-norm sharpness surrogate and a robustness-based generalization analysis (Le et al., 2025). QUATRO uses query-adaptive trust-region constraints to stabilize training under aggressive learning rates and policy staleness (Lee et al., 2026).

Sharpness-Aware Minimization. SAM (Foret et al., 2021) is a bilevel optimization method designed to seek flatter minima, and has attracted significant attention due to its consistent performance improvements across a wide range of supervised learning tasks (Foret et al., 2021; Kwon et al., 2021; Kaddour et al., 2022; Liu et al., 2022; Kim et al., 2022; Li & Giannakis, 2023; Nguyen et al., 2023a;b; Truong et al., 2024; Luo et al., 2025b; Phan et al., 2025; Luo et al., 2025a). Singh et al. (2025) propose Functional-SAM for LLM pretraining and demonstrate strong performance. Lee &

Yoon (2025) apply SAM to PPO to improve robustness in both reward and action spaces. SHAPO evaluates policy gradients at adversarially perturbed parameters to promote safe exploration (Mani et al., 2026). logits-SAM (Luo et al., 2026) uses efficient output-layer perturbations to mitigate likelihood displacement in DPO (Rafailov et al., 2023).

3 PRELIMINARIES

Notations. We consider a language model with parameters θ, defining an autoregressive policy $\pi_\theta(y \mid x)$ over output token sequences $y = (y_1, \ldots, y_T)$ given a prompt $x \in \mathcal{X}$. Each token y_t belongs to a vocabulary $\mathcal{V}$ of size N. We write μ for the prompt distribution. Given $x \sim \mu$,

$$\pi_\theta(y \mid x) = \prod_{t=1}^{T} \pi_\theta(y_t \mid x, y_{<t}),$$

where $t = 1, \ldots, T$ indexes token positions within a completion. We use $k = 0, \ldots, K-1$ to index optimization steps.

RLVR objective. Let $R(x, y) \in [0, 1]$ denote the verifiable and bounded reward assigned to completion y for prompt x. The reward-only RLVR objective is

$$J(\theta) \coloneqq \mathbb{E}_{x\sim\mu,\, y\sim\pi_\theta(\cdot|x)} \left[R(x, y)\right].$$

With reference-policy regularization, we write

$$J_{\mathrm{KL}}(\theta) \coloneqq \mathbb{E}_{x\sim\mu,\, y\sim\pi_\theta(\cdot|x)} \left[R(x, y) - \beta_{\mathrm{ref}} D_{\mathrm{KL}}\left(\pi_\theta(\cdot \mid x) \,\|\, \pi_{\mathrm{ref}}(\cdot \mid x)\right)\right], \tag{1}$$

where $\beta_{\mathrm{ref}} \geq 0$ controls the strength of reference-policy regularization.

GRPO surrogate. GRPO estimates the advantage of a completion by normalizing its reward among completions sampled from the old policy for the same prompt. For a prompt x and completion $y \sim \pi_{\mathrm{old}}(\cdot \mid x)$, define

$$\widehat{A}(x, y) \coloneqq \frac{R(x, y) - \mathbb{E}_{y'\sim\pi_{\mathrm{old}}(\cdot|x)}[R(x, y')]}{\sqrt{\mathrm{Var}_{y'\sim\pi_{\mathrm{old}}(\cdot|x)}[R(x, y')] + \epsilon_{\mathrm{std}}}}.$$

The clipped GRPO surrogate is

$$J_{\mathrm{GRPO}}(\theta) \coloneqq \mathbb{E}_{\substack{x\sim\mu \\ y\sim\pi_{\mathrm{old}}(\cdot|x)}} \left[\sum_{t=1}^{T} \min\left(r_t(\theta)\widehat{A}(x, y), \mathrm{clip}\left(r_t(\theta), 1-\varepsilon_{\mathrm{clip}}, 1+\varepsilon_{\mathrm{clip}}\right)\widehat{A}(x, y)\right)\right],$$

where $r_t(\theta)$ is the token-wise importance ratio

$$r_t(\theta) \coloneqq \frac{\pi_\theta(y_t \mid x, y_{<t})}{\pi_{\mathrm{old}}(y_t \mid x, y_{<t})}.$$

Sharpness-Aware Minimization. SAM regularizes training by explicitly penalizing parameter-space sharpness: it seeks parameters that minimize the worst-case loss within an ℓ_2 ball of radius ρ. Let $\mathcal{L}(\theta)$ denote the minimization objective, e.g., the GRPO loss $\mathcal{L}_{\mathrm{GRPO}}(\theta) \coloneqq -J_{\mathrm{GRPO}}(\theta)$ evaluated on the current sampled completions. The SAM objective is

$$\min_{\theta} \max_{\|\epsilon\|_2 \leq \rho} \mathcal{L}(\theta + \epsilon), \tag{2}$$

where $\rho > 0$ controls the perturbation radius.

In practice, the inner maximization is approximated to first order. With $g = \nabla_\theta \mathcal{L}(\theta)$, the perturbation is

$$\epsilon^\star_{\mathrm{sam}} = \rho \frac{g}{\|g\|_2}, \tag{3}$$

and the optimizer takes a descent step using the gradient at the perturbed point, $\nabla_\theta \mathcal{L}(\theta + \epsilon^\star_{\mathrm{sam}})$, treating $\epsilon^\star_{\mathrm{sam}}$ as fixed when differentiating the outer objective.

4 LOCAL CURVATURE QUANTITIES AS GENERALIZATION SIGNALS

Improvements in training reward do not necessarily translate into corresponding gains in test performance in RLVR (Wang et al., 2025). This motivates examining additional quantities associated with generalization. Prior work has studied gradient-norm-based sharpness from a generalization perspective (Le et al., 2025) and directional Hessian and Fisher quantities in relation to optimization stability (Melo et al., 2025). However, direct empirical evidence on how these quantities relate to generalization performance across independently trained GRPO runs remains limited. We therefore conduct a controlled study across 20 random seeds under identical hyperparameter settings, comparing their associations with test pass@1 against those of standard training metrics.

4.1 LOCAL CURVATURE QUANTITIES

Fix a training step k, and let $g = \nabla_\theta \mathcal{L}_k(\theta_k)$ and $H = \nabla^2_\theta \mathcal{L}_k(\theta_k)$ denote the gradient and Hessian of the training objective. For prompts $x \sim \mu$, let F be the trajectory Fisher of π_{θ_k} averaged over the prompt distribution,

$$F := \mathbb{E}_{x\sim\mu,\, y\sim\pi_{\theta_k}(\cdot|x)} \left[\nabla_\theta \log \pi_{\theta_k}(y \mid x) \nabla_\theta \log \pi_{\theta_k}(y \mid x)^\top\right].$$

The quadratic forms $u^\top H u$ and $u^\top F u$ characterize objective and policy curvature along direction u, respectively. To relate these curvature quantities to the effects of a parameter perturbation Δ, following Melo et al. (2025), we consider the second-order approximations

$$m_H(\Delta) := g^\top \Delta + \frac{1}{2}\Delta^\top H \Delta, \tag{4}$$

$$m_F(\Delta) := \frac{1}{2}\Delta^\top F \Delta, \tag{5}$$

which approximate the objective change and local policy KL, respectively. By Taylor expansion,

$$\mathcal{L}_k(\theta_k + \Delta) - \mathcal{L}_k(\theta_k) = m_H(\Delta) + O(\|\Delta\|^3),$$

and

$$\mathbb{E}_{x\sim\mu} D_{\mathrm{KL}}\left(\pi_{\theta_k}(\cdot \mid x) \| \pi_{\theta_k+\Delta}(\cdot \mid x)\right) = m_F(\Delta) + O(\|\Delta\|^3)$$

(a simple proof is provided in Lemma A.2).

4.2 EMPIRICAL CORRELATION WITH GENERALIZATION

With these quantities defined, we examine their associations with test pass@1 in the controlled study described above. We train Qwen2.5-Math-1.5B models (Yang et al., 2024) on the MATH training set (Hendrycks et al., 2021) using GRPO for one epoch under each of 20 random seeds. The correlation analysis uses the final checkpoint of each run. We evaluate pass@1 on MATH500 (Hendrycks et al., 2021; Lightman et al., 2023), AIME24 (Art of Problem Solving, 2024a), AMC (Art of Problem Solving, 2024b), Minerva Math (Lewkowycz et al., 2022), and Olympiad Bench (He et al., 2024). MATH500 is a subset of the MATH test set, and we list it separately to highlight in-domain mathematical generalization. Table 1 reports the correlation between each quantity and pass@1.

From the results in Table 1, we draw the following main conclusions: among all quantities, the second-order curvature of the Hessian and Fisher, both along the gradient direction and along random directions, exhibits a clear negative correlation with test pass@1. Training reward and gradient norm show weak correlations with test pass@1. Notably, the KL divergence to the reference model, which has recently been shown to be highly correlated with the degree of forgetting (Shenfeld et al., 2025; Zhu et al., 2025), instead demonstrates a positive Spearman correlation with test pass@1. This observation further supports the effectiveness of recent GRPO variants (Yu et al., 2025; Liu et al., 2025; Zeng et al., 2025; Sullivan & Koller, 2025; Zheng et al., 2025; Melo et al., 2025) that advocate setting $\beta_{\mathrm{ref}} = 0$ to improve performance.

4.3 AN INFORMATION-THEORETIC GENERALIZATION BOUND

In this subsection, we provide theoretical insight into the role of the defined curvature quantities in generalization. The relationship between objective directional curvature and generalization has

Table 1: Correlation between local curvature quantities and generalization performance across 20 seeds, using the final checkpoint after one training epoch for each run. Each entry reports Pearson/Spearman correlation with pass@1, computed over 4 samples per problem. Gradient and random directions are each normalized to $\|u\|_2 = 1$. For random directions, curvature estimates are averaged over three independently sampled Rademacher directions before computing correlations.

Quantity	Direction	All pass@1	Math500 pass@1
Training reward	-	$+0.112/+0.024$	$+0.140/+0.122$
KL to ref	-	$+0.048/+0.479$	$+0.053/+0.538$
$\|\nabla\mathcal{L}\|$	-	$+0.099/-0.128$	$+0.064/-0.286$
$u^\top Hu$	Gradient	$-0.498/-0.591$	$-0.439/-0.602$
$u^\top Hu$	Random	$-0.483/-0.674$	$-0.478/-0.806$
$u^\top Fu$	Gradient	$-0.710/-0.442$	$-0.731/-0.684$
$u^\top Fu$	Random	$-0.320/-0.358$	$-0.344/-0.600$

been extensively studied in the literature. A representative line of work employs perturbation-based analysis combined with Taylor expansion, often within the PAC-Bayes framework (Foret et al., 2021; Zhuang et al., 2022; Luo et al., 2025b). We therefore focus on establishing a theoretical relationship between policy directional curvature and generalization. For this purpose, we leverage an information-theoretic tool known as conditional mutual information (CMI) (Xu & Raginsky, 2017; Steinke & Zakynthinou, 2020) to bound the generalization error.

Classical CMI bounds usually study the channel from the training selector U to the algorithm output, i.e., the learned parameters θ_U. We instead study the *policy–output channel*: how information about U appears in the trajectories generated by the trained policy. The key advantage is that this output-level leakage can be controlled by trajectory-level KL divergence, whose local second-order expansion is the Fisher quadratic form m_F.

We state the informal result below; the formal setup, statement, and proof are given in Appendix A.1.

Theorem 4.1 (Informal version of Theorem A.3: m_F controls output information leakage)**.** *Let A be the training algorithm and let $\mathrm{gen}(A)$ denote the expected generalization gap. Under the standard supersample construction with dataset size n, let $x_{i,0}$ and $x_{i,1}$ be the two candidate prompts at coordinate i, and let $\Delta_i = \theta_i^1 - \theta_i^0$ be the parameter change induced by switching the i-th selector from $x_{i,0}$ to $x_{i,1}$, with all other selectors fixed. Let $m_{F,i}(\Delta_i)$ denote the trajectory-Fisher curvature along this selector-swap direction. Assume the evaluation loss is bounded in $[0,1]$, and suppose the output-channel regularity conditions in Assumption A.1 hold. Then*

$$|\mathrm{gen}(A)| \le \sqrt{\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[m_{F,i}(\Delta_i) + O(\|\Delta_i\|^3)\right]}.$$

This result shows that m_F controls how distinguishable two counterfactually trained policies are from their generated trajectories. Thus, small trajectory-Fisher curvature implies small output-level information leakage about the training selector, leading to a small CMI-style generalization gap.

5 Method: KL-SAM for GRPO

The preceding analysis relates local geometry to generalization behavior. The quantities $m_H(\Delta)$ and $m_F(\Delta)$ in Eqs. (4)–(5) approximate the objective change and local policy KL induced by a perturbation. A natural question is:

how can we jointly regularize objective and policy sensitivity during optimization?

A classical approach is the trust region method (Conn et al., 2000):

$$\min_{\Delta}\ g^\top \Delta \quad \text{s.t.} \quad m_H(\Delta) \le \delta_H,\ m_F(\Delta) \le \delta_F. \tag{6}$$

Keeping only the Fisher constraint recovers the local quadratic subproblem underlying TRPO (Schulman et al., 2015), which uses Fisher-vector products and conjugate gradients to compute policy

updates. SHAPO (Mani et al., 2026) similarly uses Fisher geometry to construct adversarial policy perturbations, while CAPO (Melo et al., 2025) explicitly estimates Hessian and Fisher directional curvatures under a last-layer approximation. Despite their different implementations, these approaches rely on dedicated curvature computations during training. This motivates a simpler alternative that incorporates the regularizing effects of local geometry directly into the training objective using only first-order gradients.

We propose KL-SAM, a *fully first-order method for implicit Hessian–Fisher curvature regularization*. It uses a single shared perturbation for the SAM-perturbed loss and the clean-to-perturbed policy KL penalty, requiring neither explicit curvature estimates nor Hessian/Fisher-vector products or iterative linear-system solves. The idealized objective is:

$$\min_{\theta} \max_{\|\epsilon\|_2 \le \rho} \mathcal{L}(\theta + \epsilon) + \lambda_{\mathrm{loc}} \mathbb{E}_{x \sim \mu} D_{\mathrm{KL}} \left(\pi_\theta(\cdot \mid x) \| \pi_{\theta+\epsilon}(\cdot \mid x) \right). \tag{7}$$

The additional policy KL term is the key component of our algorithm, which we refer to as the *local KL term*, to distinguish it from the reference-model KL regularization in J_{KL} in Eq. (1). It compares the clean policy π_θ with the locally perturbed policy $\pi_{\theta+\epsilon}$, and is essentially a quadratic form of the Fisher matrix in a neighborhood of the parameter θ. We first present how to compute this objective, and then demonstrate its implicit regularization effect.

Proposition 5.1 (Shared first-order inner maximizer)**.** *For a fixed parameter θ with nonzero loss gradient g, the first-order inner maximizers of SAM and KL-SAM over $\{\epsilon : \|\epsilon\|_2 \le \rho\}$ coincide:*

$$\epsilon^\star_{\text{KL-SAM}} = \epsilon^\star_{\text{SAM}} = \rho \frac{g}{\|g\|_2}.$$

See Proposition A.4 in Appendix A.2.

Because the local KL term has zero first-order variation at the clean policy, KL-SAM reuses the standard SAM perturbation to induce both Hessian and Fisher regularization. Both components are evaluated using the same mini-batch.

For efficient implementation, we adopt the output-layer perturbation strategy of logits-SAM (Luo et al., 2026). Under the same GRPO objective and perturbation configuration, setting $\lambda_{\mathrm{loc}} = 0$ recovers our logits-SAM baseline. KL-SAM augments this perturbed loss with a clean-to-perturbed policy KL penalty, explicitly targeting policy sensitivity along the shared perturbation direction. Thus, the efficient perturbation mechanism is inherited from logits-SAM, while the additional regularization acts on the policy distribution. This implementation keeps the additional time and memory overhead small, as reported in Section 6.4. Pseudocode for KL-SAM is provided in Algorithm 1 in Appendix B.

Theorem 5.2 (Informal version of Theorem A.5: KL-SAM regularizes Hessian–Fisher curvature)**.** *Let $\epsilon^\star$ denote the shared first-order inner maximizer in Proposition 5.1. Up to third-order terms in ρ, KL-SAM satisfies*

$$\mathcal{L}^{\text{KL-SAM}}(\theta) = \mathcal{L}(\theta) + m_H(\epsilon^\star) + \lambda_{\mathrm{loc}} m_F(\epsilon^\star) + O(\rho^3).$$

Theorem 5.2 identifies the two contributions: the perturbed loss captures objective variation, including the directional Hessian term, while the local KL penalty contributes the directional Fisher term. Both arise along the same perturbation computed from first-order gradients.

A limitation is that KL-SAM can only penalize curvature along the SAM perturbation direction, rather than directly constraining the model update direction as in the trust-region formulation in Eq. (6). However, the SAM perturbation direction corresponds to the (normalized) gradient, while the update direction in gradient descent is the negative gradient; the two share the same second-order effect (e.g., $g^\top F g = (-g)^\top F (-g)$). Therefore, an appropriate interpretation is that KL-SAM penalizes the curvature along the original gradient descent update direction, thereby steering the model updates toward directions of lower curvature.

6 Experiments

6.1 Experimental Setup

Models and benchmarks. We conduct experiments with three reasoning models: Qwen2.5-Math-1.5B and -7B (Yang et al., 2024), and DeepSeek-R1-Distill-Qwen-1.5B (Guo et al., 2025). For each

Table 2: Pass@1 results (%) for Qwen2.5-Math-1.5B over three training seeds. Benchmark columns report means, and Avg. reports mean ± standard deviation (SD). Bold denotes the best mean within each objective. Full per-benchmark SDs are reported in Appendix D.

Loss	Method	MATH500	AIME24	AMC23	Minerva	Olympiad	Avg.
GRPO	Baseline	69.55	9.72	48.75	22.00	33.46	$36.70_{\pm 0.23}$
	logits-SAM	70.08	11.94	**52.08**	**24.36**	33.37	$38.37_{\pm 0.28}$
	KL-SAM	**70.62**	11.94	51.88	24.05	**34.26**	$\mathbf{38.55}_{\pm 1.11}$
	KL-only	70.42	**12.22**	47.08	23.31	33.17	$37.24_{\pm 0.60}$
DR.GRPO	Baseline	67.02	12.50	48.33	22.24	31.38	$36.30_{\pm 0.83}$
	logits-SAM	67.12	11.94	47.92	22.37	31.46	$36.16_{\pm 0.47}$
	KL-SAM	**68.82**	11.11	**50.00**	**22.79**	**32.96**	$\mathbf{37.14}_{\pm 0.80}$
	KL-only	64.83	**13.61**	45.21	22.27	30.46	$35.28_{\pm 1.61}$

model, we consider two RL objectives, GRPO (Shao et al., 2024) and DR.GRPO (Liu et al., 2025). All methods are trained on the MATH training set (Hendrycks et al., 2021) and evaluated on five mathematical reasoning benchmarks: MATH500 (Hendrycks et al., 2021; Lightman et al., 2023), AIME24 (Art of Problem Solving, 2024a), AMC23 (Art of Problem Solving, 2024b), Minerva Math (Lewkowycz et al., 2022), and Olympiad Bench (He et al., 2024). We compare KL-SAM with the baseline without SAM or local policy-KL regularization and logits-SAM (Luo et al., 2026), which removes the local KL term. KL-only replaces the perturbed loss with the clean loss while retaining the local-KL penalty and sharing KL-SAM's perturbation radius and KL coefficient. Our implementation is built on the open-source `open-r1` project (Hugging Face, 2025).

Training details. All methods use AdamW (Loshchilov & Hutter, 2017) as the underlying optimizer. With the training budget fixed at one epoch, we search the learning rate for the baseline over $\{1 \times 10^{-6}, 3 \times 10^{-6}, 5 \times 10^{-6}, 1 \times 10^{-5}\}$ and use the selected learning rate for all methods under the corresponding objective: 5×10^{-6} for GRPO and 3×10^{-6} for DR.GRPO. In both cases, we use a constant learning-rate schedule with warmup. The training batch size is 128, and the maximum prompt and completion lengths are set to 512 and 1024, respectively. For each prompt, we sample 8 completions for GRPO training. We search $\rho \in \{1\times10^{-3}, 5\times10^{-3}, 1\times10^{-2}\}$ for both SAM variants and additionally $\lambda_{\text{loc}} \in \{1.0, 5.0, 10.0\}$ for KL-SAM. All other hyperparameters are kept identical to the corresponding baseline. We select the hyperparameters using a 500-problem validation subset of the MATH test set that is disjoint from MATH500; this subset is used only for hyperparameter selection.

Evaluation. We evaluate all models with LightEval (Habib et al., 2023) using the vLLM inference backend (Kwon et al., 2023). During evaluation, we use a maximum model context length of 4096 tokens and allow each model to generate up to 3072 new tokens. We sample completions with temperature 0.6 and top-p 0.95. For each benchmark, we report pass@1 averaged over four sampled completions. The Avg. column reports the average pass@1 across the five benchmarks. Following common practice in the community (Wang et al., 2025; Zheng et al., 2025), we evaluate checkpoints every 100 training steps and at the final step, and report the checkpoint with the best average performance across the five benchmarks.

6.2 Main Results

Across all six model–objective settings, KL-SAM improves both Avg. and MATH500 pass@1 over the baseline (Tables 2 and 3). Averaging the three-seed means for Qwen-1.5B and single-run results for the other models gives an Avg. gain of 2.12 percentage points ($41.06 \rightarrow 43.18$). Three-seed final-checkpoint results for Qwen2.5-Math-1.5B are reported in Table 6 of Appendix D.

Component ablation. KL-SAM outperforms logits-SAM in Avg. in five of six settings, with an average gain of 0.91 percentage points. It also achieves higher pass@1 than logits-SAM on the in-domain MATH500 benchmark across all six model–objective settings. On Qwen2.5-Math-1.5B, gains over logits-SAM are $0.18/0.98$ points under GRPO/DR.GRPO, compared with $1.31/1.86$ points over KL-only (Table 2).

Table 3: Single-run pass@1 results (%) for DeepSeek-R1-Distill-Qwen-1.5B and Qwen2.5-Math-7B. Bold denotes the best result within each model–loss block.

Model	Loss	Method	MATH500	AIME24	AMC23	Minerva	Olympiad	Avg.
DeepSeek-1.5B	GRPO	Baseline	76.20	19.17	61.25	24.82	38.67	44.02
		logits-SAM	76.50	**23.33**	56.25	26.56	39.59	44.45
		KL-SAM	**77.35**	20.83	**63.13**	**28.68**	**41.37**	**46.27**
	DR.GRPO	Baseline	76.90	15.00	55.00	26.10	39.70	42.54
		logits-SAM	78.15	15.83	56.87	26.65	40.78	43.66
		KL-SAM	**79.55**	**17.50**	**58.13**	**28.31**	**42.52**	**45.20**
Qwen-7B	GRPO	Baseline	75.55	15.83	53.13	30.61	38.41	42.70
		logits-SAM	77.15	20.83	**60.63**	32.90	**40.04**	**46.31**
		KL-SAM	**77.25**	**22.50**	56.25	**33.92**	39.70	45.92
	DR.GRPO	Baseline	69.40	**28.33**	58.75	28.03	35.85	44.07
		logits-SAM	73.65	25.83	57.50	28.86	37.26	44.62
		KL-SAM	**73.95**	25.00	**60.00**	**32.35**	**38.56**	**45.97**

6.3 Training Dynamics and Generalization

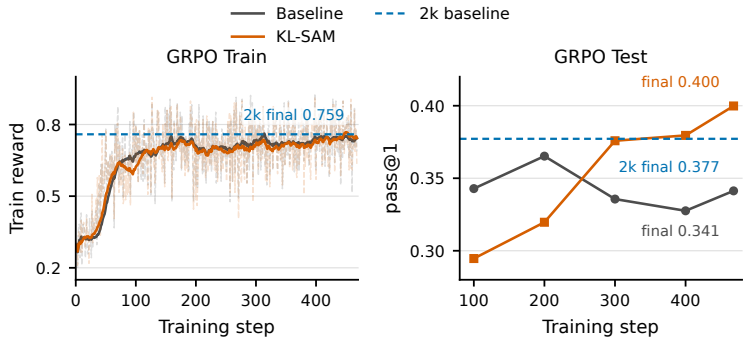


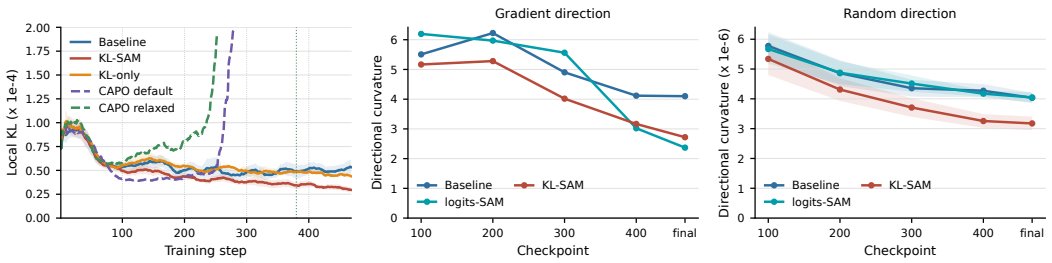
Figure 1: Training reward and test performance dynamics for Qwen2.5-Math-1.5B under GRPO. The 2k baseline denotes the final performance of a GRPO baseline trained for 2k steps with a smaller learning rate of 1×10^{-6}.

Figure 1 provides a closer look at the training dynamics. The GRPO baseline and KL-SAM achieve comparable training rewards, suggesting that the performance difference is not simply due to KL-SAM optimizing the training reward more aggressively. Instead, their test trajectories differ substantially: the baseline tends to plateau in the middle and later stages of training and can even begin to degrade, whereas KL-SAM continues to improve more steadily. Notably, under the relatively aggressive learning rate used in our main experiments, KL-SAM surpasses the 2k-step small-learning-rate baseline within a single training epoch (469 steps).

6.4 Additional Analysis

Curvature dynamics. Figure 2 tracks local policy KL and Hessian curvature during training. Panel (a) shows that KL-SAM reduces the local KL between the clean and perturbed policies, consistent with the directional Fisher regularization in Theorem 5.2. Panels (b–c) report Hessian curvature along gradient and random directions. For the random directions, $\mathbb{E}[u^\top H u] = \mathrm{tr}(H)/d$, where d is the probed parameter dimension, linking the random-direction diagnostic to the dimension-normalized Hessian trace. KL-SAM exhibits lower curvature than the baseline in both directions across the measured checkpoints. Although logits-SAM attains slightly lower gradient-direction curvature than KL-SAM at later checkpoints, its random-direction curvature remains close to the baseline and consistently above KL-SAM.

Comparison with CAPO. We also compare with CAPO token-level masking (Melo et al., 2025), which can be interpreted as a token-level trust-region surrogate for jointly controlling objective



(a) Local KL **(b)** Gradient direction **(c)** Random direction

Figure 2: Local policy and Hessian curvature diagnostics during training: (a) KL divergence between the clean and perturbed policies; (b–c) directional Hessian curvature $u^\top Hu$ along gradient and random directions, each normalized to $\|u\|_2 = 1$. CAPO default uses the original masking threshold, while CAPO relaxed increases the masking threshold by $10\times$.

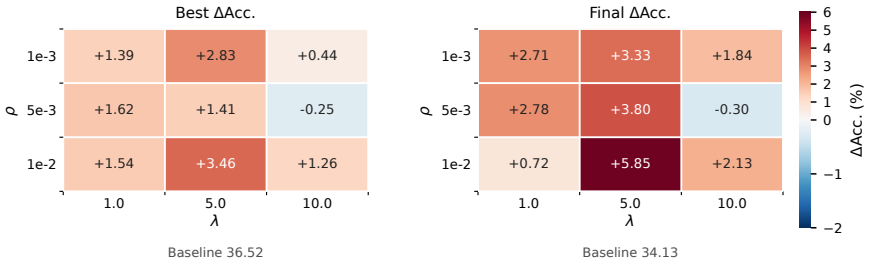


Figure 3: Hyperparameter sensitivity of KL-SAM over perturbation radius ρ and local-KL weight λ_{loc}. Values denote absolute pass@1 improvements over the baseline.

directional curvature and policy directional curvature. In our longer training runs, CAPO does not reliably suppress the local curvature signal over time, and its training can fail after more steps. This contrasts with the original CAPO evaluation, which focuses on a shorter 100-step training horizon. The result suggests that token-level masking alone may be insufficient for long-run GRPO training, motivating our dual-curvature perspective.

Hyperparameter sensitivity. Figure 3 shows that KL-SAM is not sensitive to a single narrow hyperparameter choice: most combinations of ρ and λ_{loc} improve over the baseline. The best region is centered around a moderate local-KL weight, while overly strong regularization can reduce the gain in some settings. This suggests that the local-KL term is useful across a range of perturbation radii, but its strength should be kept in a moderate range.

Computational overhead. Relative to the baseline, logits-SAM adds 3.05% to the average step time and 1.4% to peak memory usage. For KL-SAM, the average step time increases from 13.70s to 14.25s, corresponding to a 4.0% increase, while peak memory increases from 66.53 GiB to 68.03 GiB, corresponding to a 2.3% increase. KL-SAM thus implements implicit dual-curvature regularization using only first-order gradients, with modest additional time and memory overhead.

7 Conclusion

Our controlled 20-seed study links directional Hessian and Fisher curvature to generalization performance in GRPO-based RLVR. Motivated by these findings, KL-SAM uses a single shared perturbation computed from first-order gradients to implement implicit dual-curvature regularization. Experiments on 1.5B–7B models show improved average pass@1 with modest overhead, while a training-dynamics comparison suggests gains at a larger learning rate with fewer updates. Extending these results to larger models and broader training settings remains future work.

AI USE STATEMENT

During the preparation of this manuscript, the authors used AI-assisted tools to improve the clarity and language of the manuscript, assist in identifying relevant literature, support code debugging, and check the correctness of mathematical proofs. All AI-assisted outputs were critically reviewed and verified by the authors, who take full responsibility for the accuracy, integrity, and final content of the manuscript.

REFERENCES

Art of Problem Solving. Aime problems and solutions. https://artofproblemsolving.com/wiki/index.php/AIME_Problems_and_Solutions, 2024a. Accessed: 2026-05-02.

Art of Problem Solving. Amc problems and solutions. https://artofproblemsolving.com/wiki/index.php/AMC_Problems_and_Solutions, 2024b. Accessed: 2026-05-02.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

Karl Cobbe, Oleg Klimov, Chris Hesse, Taehoon Kim, and John Schulman. Quantifying generalization in reinforcement learning. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 1282–1289. PMLR, 2019. URL https://proceedings.mlr.press/v97/cobbe19a.html.

Andrew R Conn, Nicholas IM Gould, and Philippe L Toint. *Trust region methods*. SIAM, 2000.

Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=6Tm1mposlrM.

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.

Nathan Habib, Clémentine Fourrier, Hynek Kydlíček, Thomas Wolf, and Lewis Tunstall. Lighteval: A lightweight framework for llm evaluation, 2023. URL https://github.com/huggingface/lighteval.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, et al. Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3828–3850, 2024.

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. *arXiv preprint arXiv:2103.03874*, 2021.

Rémy Hosseinkhan-Boucher, Onofrio Semeraro, and Lionel Mathelin. Evidence on the regularisation properties of maximum-entropy reinforcement learning. *arXiv preprint arXiv:2501.17115*, 2025.

Hugging Face. Open r1: A fully open reproduction of deepseek-r1, January 2025. URL https://github.com/huggingface/open-r1.

Jean Kaddour, Linqing Liu, Ricardo Silva, and Matt J Kusner. When do flat minima optimizers work? *Advances in Neural Information Processing Systems*, 35:16577–16595, 2022.

Minyoung Kim, Da Li, Shell X Hu, and Timothy Hospedales. Fisher sam: Information geometry and sharpness aware minimisation. In *International Conference on Machine Learning*, pp. 11148–11161. PMLR, 2022.

Jungmin Kwon, Jeongseop Kim, Hyunseo Park, and In Kwon Choi. Asam: Adaptive sharpness-aware minimization for scale-invariant learning of deep neural networks. In *International Conference on Machine Learning*, pp. 5905–5914. PMLR, 2021.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023.

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.

Tue Le, Linh Ngo Van, and Trung Le. Sharpness-guided group relative policy optimization via probability shaping. *arXiv preprint arXiv:2511.00066*, 2025.

Doyeon Lee, Eunyi Lyou, Hyunsoo Cho, Sookyung Kim, Joonseok Lee, and Jaemoo Choi. QUATRO: Query-adaptive trust region policy optimization for LLM fine-tuning. *arXiv preprint arXiv:2602.04620*, 2026. URL https://arxiv.org/abs/2602.04620.

Hyun Kyu Lee and Sung Whan Yoon. Flat reward in policy parameter space implies robust reinforcement learning. In *The Thirteenth International Conference on Learning Representations*, 2025.

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.

Bingcong Li and Georgios B. Giannakis. Enhancing sharpness-aware optimization through variance suppression, 2023. URL https://arxiv.org/abs/2309.15639.

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 *The twelfth international conference on learning representations*, 2023.

Yong Liu, Siqi Mai, Xiangning Chen, Cho-Jui Hsieh, and Yang You. Towards efficient and scalable sharpness-aware minimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12360–12370, 2022.

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

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Haocheng Luo, Mehrtash Harandi, Dinh Phung, and Trung Le. Unveiling m-sharpness through the structure of stochastic gradient noise. *arXiv preprint arXiv:2509.18001*, 2025a.

Haocheng Luo, Tuan Truong, Tung Pham, Mehrtash Harandi, Dinh Phung, and Trung Le. Explicit eigenvalue regularization improves sharpness-aware minimization. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2025b.

Haocheng Luo, Zehang Deng, Thanh-Toan Do, Mehrtash Harandi, Dinh Phung, and Trung Le. Sharpness-aware minimization in logit space efficiently enhances direct preference optimization, 2026. URL https://arxiv.org/abs/2603.18258.

Clare Lyle, Mark Rowland, Will Dabney, Marta Kwiatkowska, and Yarin Gal. Learning dynamics and generalization in deep reinforcement learning. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 14560–14581. PMLR, 2022. URL https://proceedings.mlr.press/v162/lyle22a.html.

Kaustubh Mani, Yann Pequignot, Vincent Mai, and Liam Paull. Shapo: Sharpness-aware policy optimization for safe exploration. *arXiv preprint arXiv:2606.10228*, 2026.

Luckeciano C Melo, Alessandro Abate, and Yarin Gal. Stabilizing policy gradients for sample-efficient reinforcement learning in llm reasoning. *arXiv preprint arXiv:2510.00819*, 2025.

Van-Anh Nguyen, Trung Le, Anh Bui, Thanh-Toan Do, and Dinh Phung. Optimal transport model distributional robustness. *Advances in Neural Information Processing Systems*, 36:24074–24087, 2023a.

Van-Anh Nguyen, Tung-Long Vuong, Hoang Phan, Thanh-Toan Do, Dinh Phung, and Trung Le. Flat seeking bayesian neural networks. *Advances in Neural Information Processing Systems*, 36: 30807–30820, 2023b.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

Hoang Phan, Lam Tran, Quyen Tran, Ngoc Tran, Tuan Truong, Qi Lei, Nhat Ho, Dinh Phung, and Trung Le. Beyond losses reweighting: Empowering multi-task learning via the generalization perspective. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2440–2450, 2025.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

Neel Rajani, Aryo Pradipta Gema, Seraphina Goldfarb-Tarrant, and Ivan Titov. Scalpel vs. hammer: Grpo amplifies existing capabilities, sft replaces them. *arXiv preprint arXiv:2507.10616*, 2025.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 3505–3506, 2020.

Mahdi Sabbaghi, George Pappas, Adel Javanmard, and Hamed Hassani. Robust policy optimization to prevent catastrophic forgetting. *arXiv preprint arXiv:2602.08813*, 2026.

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International conference on machine learning*, pp. 1889–1897. PMLR, 2015.

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.

Idan Shenfeld, Jyothish Pari, and Pulkit Agrawal. Rl's razor: Why online reinforcement learning forgets less. *arXiv preprint arXiv:2509.04259*, 2025.

Sidak Pal Singh, Hossein Mobahi, Atish Agarwala, and Yann Dauphin. Avoiding spurious sharpness minimization broadens applicability of sam. *arXiv preprint arXiv:2502.02407*, 2025.

Thomas Steinke and Lydia Zakynthinou. Reasoning about generalization via conditional mutual information. In *Conference on Learning Theory*, pp. 3437–3452. PMLR, 2020.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. *Advances in neural information processing systems*, 33:3008–3021, 2020.

Michael Sullivan and Alexander Koller. Grpo is secretly a process reward model. *arXiv preprint arXiv:2509.21154*, 2025.

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

Tuan Truong, Quyen Tran, Quan Pham-Ngoc, Nhat Ho, Dinh Phung, and Trung Le. Improving generalization with flat hilbert bayesian inference. *arXiv preprint arXiv:2410.04196*, 2024.

Yiping Wang, Qing Yang, Zhiyuan Zeng, Liliang Ren, Liyuan Liu, Baolin Peng, Hao Cheng, Xuehai He, Kuan Wang, Jianfeng Gao, et al. Reinforcement learning for reasoning in large language models with one training example. *arXiv preprint arXiv:2504.20571*, 2025.

Aolin Xu and Maxim Raginsky. Information-theoretic analysis of generalization capability of learning algorithms. *Advances in neural information processing systems*, 30, 2017.

An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, et al. Qwen2. 5-math technical report: Toward mathematical expert model via self-improvement. *arXiv preprint arXiv:2409.12122*, 2024.

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. *arXiv preprint arXiv:2503.14476*, 2025.

Yu Yue, Yufeng Yuan, Qiying Yu, Xiaochen Zuo, Ruofei Zhu, Wenyuan Xu, Jiaze Chen, Chengyi Wang, TianTian Fan, Zhengyin Du, et al. Vapo: Efficient and reliable reinforcement learning for advanced reasoning tasks. *arXiv preprint arXiv:2504.05118*, 2025.

Anxiang Zeng, Haibo Zhang, Kaixiang Mo, Long Zhang, Shuman Liu, Yanhui Huang, Yawen Liu, Yuepeng Sheng, and Yuwei Huang. Compass-thinker-7b technical report. *arXiv preprint arXiv:2508.08909*, 2025.

Chiyuan Zhang, Oriol Vinyals, Remi Munos, and Samy Bengio. A study on overfitting in deep reinforcement learning. *arXiv preprint arXiv:1804.06893*, 2018. URL https://arxiv.org/abs/1804.06893.

Haizhong Zheng, Yang Zhou, Brian R Bartoldson, Bhavya Kailkhura, Fan Lai, Jiawei Zhao, and Beidi Chen. Act only when it pays: Efficient reinforcement learning for llm reasoning via selective rollouts. *arXiv preprint arXiv:2506.02177*, 2025.

Hanqing Zhu, Zhenyu Zhang, Hanxian Huang, DiJia Su, Zechun Liu, Jiawei Zhao, Igor Fedorov, Hamed Pirsiavash, Zhizhou Sha, Jinwon Lee, et al. The path not taken: Rlvr provably learns off the principals. *arXiv preprint arXiv:2511.08567*, 2025.

Juntang Zhuang, Boqing Gong, Liangzhe Yuan, Yin Cui, Hartwig Adam, Nicha Dvornek, Sekhar Tatikonda, James Duncan, and Ting Liu. Surrogate gap minimization improves sharpness-aware training. *arXiv preprint arXiv:2203.08065*, 2022.

A Formal Theorems and Proofs

A.1 Output-CMI Generalization Bound via Trajectory Fisher

Setup. Let $Z = (X, \cdot) \sim \mathcal{D}$ denote a prompt/task instance, where X is the prompt or initial state. A policy π_θ induces a trajectory distribution P_θ^x, whose density or probability mass function is

$$p_\theta^x(y) \coloneqq \pi_\theta(y \mid x) = \prod_{t=1}^{T} \pi_\theta(y_t \mid x, y_{<t})$$

over completions or trajectories $y = (y_1, \ldots, y_T)$. Let

$$c(z, y) \in [0, 1]$$

be a bounded evaluation loss. Define the stochastic policy loss

$$\ell(\theta, z) \coloneqq \mathbb{E}_{Y \sim P_\theta^{x(z)}}[c(z, Y)].$$

The population and empirical risks are

$$L_{\mathcal{D}}(\theta) = \mathbb{E}_{Z \sim \mathcal{D}}[\ell(\theta, Z)], \qquad L_S(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell(\theta, Z_i).$$

Supersample construction. Let

$$\widetilde{S} = \{(Z_{i,0}, Z_{i,1})\}_{i=1}^n, \qquad Z_{i,0}, Z_{i,1} \overset{\text{i.i.d.}}{\sim} \mathcal{D},$$

and let

$$U = (U_1, \ldots, U_n), \qquad U_i \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(1/2).$$

The selected training set is

$$S_U = \{Z_{i,U_i}\}_{i=1}^n.$$

Let A be a deterministic learning algorithm and write

$$\theta_U = A(S_U).$$

For a fixed coordinate i, define the counterfactual selectors

$$U^{(i \leftarrow b)} = (U_1, \ldots, U_{i-1}, b, U_{i+1}, \ldots, U_n), \qquad b \in \{0, 1\}.$$

The two counterfactually trained parameters are

$$\theta_i^b \coloneqq A(S_{U^{(i \leftarrow b)}}), \qquad b \in \{0, 1\},$$

and the selector-swap direction is

$$\Delta_i \coloneqq \theta_i^1 - \theta_i^0.$$

The supersample generalization gap is

$$\widetilde{\text{gen}}(A) \coloneqq \frac{1}{n} \sum_{i=1}^{n} \left[\ell(\theta_U, Z_{i,1-U_i}) - \ell(\theta_U, Z_{i,U_i})\right].$$

Its expectation equals the usual expected population-minus-training risk:

$$\mathbb{E}_{\widetilde{S},U}[\widetilde{\text{gen}}(A)] = \mathbb{E}_S[L_{\mathcal{D}}(A(S)) - L_S(A(S))].$$

Assumption A.1 (Output-channel regularity)**.** *The following conditions hold.*

(1) For each candidate prompt $x \in \{x_{i,0}, x_{i,1}\}$, the family $\{P_\theta^x\}$ has common support in a neighborhood of the line segment between θ_i^0 and θ_i^1, and the measures $P_{\theta_i^0}^x$ and $P_{\theta_i^1}^x$ are mutually absolutely continuous.

(2) In the same neighborhood, with respect to a common dominating measure μ_x, write $P_\theta^x(dy) = p_\theta^x(y)\mu_x(dy)$. The map $\theta \mapsto \log p_\theta^x(y)$ is three times continuously differentiable, and differentiation under the integral sign up to the required order is justified. The score, Hessian, and third derivative of $\log p_\theta^x(y)$ are dominated by integrable envelopes uniformly over θ in this neighborhood. Moreover, the third-order Taylor remainder constants and the Lipschitz constants of $F_{\mathrm{traj}}(\theta; x)$ are bounded by deterministic constants uniform over all i, $(\widetilde{S}, U_{-i})$, and candidate prompts $x \in \{x_{i,0}, x_{i,1}\}$.

Lemma A.2 (Trajectory KL and Fisher expansion)**.** *Under Assumption A.1, fix a candidate prompt $x \in \{x_{i,0}, x_{i,1}\}$ and a parameter θ in the neighborhood specified there. Define the trajectory score*

$$s_\theta^x(y) \coloneqq \nabla_\theta \log p_\theta^x(y) = \sum_{t=1}^{T} \nabla_\theta \log \pi_\theta(y_t \mid x, y_{<t}),$$

and the trajectory Fisher matrix

$$F_{\mathrm{traj}}(\theta; x) \coloneqq \mathbb{E}_{Y \sim P_\theta^x}\left[s_\theta^x(Y) s_\theta^x(Y)^\top\right].$$

Then, for sufficiently small Δ,

$$D_{\mathrm{KL}}\big(P_\theta^x \,\big\|\, P_{\theta+\Delta}^x\big) = \frac{1}{2}\Delta^\top F_{\mathrm{traj}}(\theta; x)\Delta + O(\|\Delta\|^3).$$

Equivalently,

$$D_{\mathrm{KL}}\big(P_\theta^x \,\big\|\, P_{\theta+\Delta}^x\big) = m_F^{\mathrm{traj}}(\theta, x; \Delta) + O(\|\Delta\|^3),$$

where

$$m_F^{\mathrm{traj}}(\theta, x; \Delta) \coloneqq \frac{1}{2}\Delta^\top F_{\mathrm{traj}}(\theta; x)\Delta.$$

Proof. By Taylor expanding $\log p_{\theta+\Delta}^x(y)$ around θ,

$$\log p_{\theta+\Delta}^x(y) = \log p_\theta^x(y) + \Delta^\top \nabla_\theta \log p_\theta^x(y) + \frac{1}{2}\Delta^\top \nabla_\theta^2 \log p_\theta^x(y)\Delta + O(\|\Delta\|^3).$$

Therefore,

$$\begin{aligned} D_{\mathrm{KL}}(P_\theta^x \| P_{\theta+\Delta}^x) &= \mathbb{E}_{Y \sim P_\theta^x}\left[\log p_\theta^x(Y) - \log p_{\theta+\Delta}^x(Y)\right] \\ &= -\Delta^\top \mathbb{E}_{P_\theta^x}[\nabla_\theta \log p_\theta^x(Y)] - \frac{1}{2}\Delta^\top \mathbb{E}_{P_\theta^x}[\nabla_\theta^2 \log p_\theta^x(Y)]\Delta + O(\|\Delta\|^3). \end{aligned}$$

By Assumption A.1, differentiation under the integral sign is justified. Hence

$$\mathbb{E}_{P_\theta^x}[s_\theta^x(Y)] = \int \nabla_\theta p_\theta^x(y)\, \mu_x(dy) = \nabla_\theta \int p_\theta^x(y)\, \mu_x(dy) = 0.$$

Similarly,

$$0 = \nabla_\theta^2 \int p_\theta^x(y)\, \mu_x(dy) = \int \nabla_\theta^2 p_\theta^x(y)\, \mu_x(dy).$$

Using

$$\nabla_\theta^2 p_\theta^x(y) = p_\theta^x(y)\left[\nabla_\theta^2 \log p_\theta^x(y) + s_\theta^x(y) s_\theta^x(y)^\top\right],$$

we obtain

$$-\mathbb{E}_{P_\theta^x}[\nabla_\theta^2 \log p_\theta^x(Y)] = \mathbb{E}_{P_\theta^x}[s_\theta^x(Y) s_\theta^x(Y)^\top] = F_{\mathrm{traj}}(\theta; x).$$

Thus

$$D_{\mathrm{KL}}(P_\theta^x \| P_{\theta+\Delta}^x) = \frac{1}{2}\Delta^\top F_{\mathrm{traj}}(\theta; x)\Delta + O(\|\Delta\|^3).$$

□

Theorem A.3 (Output-CMI generalization bound via trajectory Fisher)**.** *For each i, condition on $(\widetilde{S}, U_{-i})$. Let*

$$x_{i,b} \coloneqq x(Z_{i,b}), \qquad b \in \{0, 1\}.$$

Define the pair-output observation

$$O_i = (Y_{i,0}, Y_{i,1}),$$

where, conditional on $U_i = b$,

$$Y_{i,0} \sim P_{\theta_i^b}^{x_{i,0}}, \qquad Y_{i,1} \sim P_{\theta_i^b}^{x_{i,1}},$$

independently. Equivalently, the conditional law of O_i given $U_i = b$ is

$$\mathsf{P}_i^b \coloneqq P_{\theta_i^b}^{x_{i,0}} \otimes P_{\theta_i^b}^{x_{i,1}}.$$

Let the output conditional mutual information be

$$I_i \coloneqq I(U_i; O_i \mid \widetilde{S}, U_{-i}).$$

Then

$$\left|\mathbb{E}_{\widetilde{S},U}[\widetilde{\mathrm{gen}}(A)]\right| \le \sqrt{\frac{2}{n}\sum_{i=1}^{n} \mathbb{E}_{\widetilde{S},U_{-i}}[I_i]}.$$

Moreover, suppose Assumption A.1 holds. Assume additionally that there exists $\delta > 0$ such that $\|\Delta_i\| \le \delta$ almost surely for all i. Define the pairwise trajectory Fisher matrix

$$F_i \coloneqq F_{\mathrm{traj}}(\theta_i^0; x_{i,0}) + F_{\mathrm{traj}}(\theta_i^0; x_{i,1}),$$

and

$$m_{F,i}^{\mathrm{traj}}(\Delta_i) \coloneqq \frac{1}{2}\Delta_i^\top F_i \Delta_i.$$

Then there exists a deterministic constant $C_3 < \infty$, depending only on the uniform constants in Assumption A.1, such that

$$I_i \le \frac{1}{2} m_{F,i}^{\mathrm{traj}}(\Delta_i) + \frac{C_3}{4}\|\Delta_i\|^3.$$

Consequently,

$$\left|\mathbb{E}_{\widetilde{S},U}[\widetilde{\mathrm{gen}}(A)]\right| \le \sqrt{\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{\widetilde{S},U_{-i}}\left[m_{F,i}^{\mathrm{traj}}(\Delta_i) + \frac{C_3}{2}\|\Delta_i\|^3\right]}.$$

Proof. First, we prove the CMI part. For each i, define

$$\bar{U}_i \coloneqq 1 - U_i.$$

Using the pair-output observation $O_i = (Y_{i,0}, Y_{i,1})$, define

$$\varphi_i(u, o) \coloneqq c(Z_{i,1-u}, y_{1-u}) - c(Z_{i,u}, y_u), \qquad o = (y_0, y_1).$$

Since $c \in [0, 1]$, we have

$$\varphi_i(u, o) \in [-1, 1].$$

Moreover, under the product measure where $U_i \sim \mathrm{Bernoulli}(1/2)$ is independent of O_i, the conditional mean is zero:

$$\mathbb{E}_{U_i}[\varphi_i(U_i, o)] = \frac{1}{2}\left[c(Z_{i,1}, y_1) - c(Z_{i,0}, y_0)\right] + \frac{1}{2}\left[c(Z_{i,0}, y_0) - c(Z_{i,1}, y_1)\right] = 0.$$

Thus $\varphi_i(U_i, O_i)$ is 1-sub-Gaussian under the product measure.

By the standard variational inequality for mutual information, for each fixed $(\widetilde{S}, U_{-i})$,

$$\left|\mathbb{E}[\varphi_i(U_i, O_i) \mid \widetilde{S}, U_{-i}]\right| \le \sqrt{2I(U_i; O_i \mid \widetilde{S}, U_{-i})} = \sqrt{2I_i}.$$

By construction of O_i, conditional on $(\widetilde{S}, U)$,

$$\mathbb{E}[\varphi_i(U_i, O_i) \mid \widetilde{S}, U] = \ell(\theta_U, Z_{i,1-U_i}) - \ell(\theta_U, Z_{i,U_i}).$$

Therefore,

$$\mathbb{E}[\widetilde{\mathrm{gen}}(A)] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[\varphi_i(U_i, O_i)].$$

Taking expectation and summing over i,

$$|\mathbb{E}[\widetilde{\mathrm{gen}}(A)]| = \left|\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\varphi_i(U_i, O_i)]\right| \leq \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\sqrt{2I_i}].$$

By Cauchy–Schwarz and Jensen's inequality,

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\sqrt{2I_i}] \leq \sqrt{\frac{2}{n}\sum_{i=1}^{n}\mathbb{E}[I_i]}.$$

Therefore,

$$|\mathbb{E}[\widetilde{\mathrm{gen}}(A)]| \leq \sqrt{\frac{2}{n}\sum_{i=1}^{n}\mathbb{E}[I_i]}.$$

Now we bound I_i by a Fisher curvature term. Conditional on $(\widetilde{S}, U_{-i})$, the selector U_i is binary and

$$O_i \mid U_i = b \sim \mathsf{P}_i^b.$$

Therefore,

$$I_i = \mathrm{JS}(\mathsf{P}_i^0, \mathsf{P}_i^1).$$

The Jensen–Shannon divergence is bounded by the symmetrized KL:

$$\mathrm{JS}(\mathsf{P}_i^0, \mathsf{P}_i^1) \leq \frac{1}{4}\left[D_{\mathrm{KL}}(\mathsf{P}_i^0\|\mathsf{P}_i^1) + D_{\mathrm{KL}}(\mathsf{P}_i^1\|\mathsf{P}_i^0)\right].$$

By the product structure of P_i^0 and P_i^1,

$$D_{\mathrm{KL}}(\mathsf{P}_i^0\|\mathsf{P}_i^1) = D_{\mathrm{KL}}(P_{\theta_i^0}^{x_{i,0}}\|P_{\theta_i^1}^{x_{i,0}}) + D_{\mathrm{KL}}(P_{\theta_i^0}^{x_{i,1}}\|P_{\theta_i^1}^{x_{i,1}}),$$

and

$$D_{\mathrm{KL}}(\mathsf{P}_i^1\|\mathsf{P}_i^0) = D_{\mathrm{KL}}(P_{\theta_i^1}^{x_{i,0}}\|P_{\theta_i^0}^{x_{i,0}}) + D_{\mathrm{KL}}(P_{\theta_i^1}^{x_{i,1}}\|P_{\theta_i^0}^{x_{i,1}}).$$

Applying Lemma A.2 under Assumption A.1, we first obtain, for each $x \in \{x_{i,0}, x_{i,1}\}$,

$$D_{\mathrm{KL}}(P_{\theta_i^0}^{x}\|P_{\theta_i^1}^{x}) = \frac{1}{2}\Delta_i^\top F_{\mathrm{traj}}(\theta_i^0; x)\Delta_i + r_{i,x}^{+},$$

with

$$|r_{i,x}^{+}| \leq C_x^{+}\|\Delta_i\|^3.$$

For the reverse KL, we apply Lemma A.2 with base point θ_i^1 and perturbation $-\Delta_i$. This gives

$$D_{\mathrm{KL}}(P_{\theta_i^1}^{x}\|P_{\theta_i^0}^{x}) = \frac{1}{2}(-\Delta_i)^\top F_{\mathrm{traj}}(\theta_i^1; x)(-\Delta_i) + \widetilde{r}_{i,x}^{-},$$

or equivalently,

$$D_{\mathrm{KL}}(P_{\theta_i^1}^{x}\|P_{\theta_i^0}^{x}) = \frac{1}{2}\Delta_i^\top F_{\mathrm{traj}}(\theta_i^1; x)\Delta_i + \widetilde{r}_{i,x}^{-},$$

with

$$|\widetilde{r}_{i,x}^{-}| \leq \widetilde{C}_x^{-}\|\Delta_i\|^3.$$

By the local Lipschitz continuity of $F_{\mathrm{traj}}(\theta; x)$, there exists a constant $L_x < \infty$ such that

$$\left\|F_{\mathrm{traj}}(\theta_i^1; x) - F_{\mathrm{traj}}(\theta_i^0; x)\right\| \leq L_x\|\Delta_i\|.$$

Hence

$$\left|\frac{1}{2}\Delta_i^\top\left[F_{\mathrm{traj}}(\theta_i^1; x) - F_{\mathrm{traj}}(\theta_i^0; x)\right]\Delta_i\right| \leq \frac{L_x}{2}\|\Delta_i\|^3.$$

Therefore, defining

$$r^-_{i,x} \coloneqq \widetilde{r}^-_{i,x} + \frac{1}{2}\Delta_i^\top \left[F_{\mathrm{traj}}(\theta_i^1; x) - F_{\mathrm{traj}}(\theta_i^0; x)\right] \Delta_i,$$

we have

$$D_{\mathrm{KL}}(P^x_{\theta_i^1} \| P^x_{\theta_i^0}) = \frac{1}{2}\Delta_i^\top F_{\mathrm{traj}}(\theta_i^0; x)\Delta_i + r^-_{i,x},$$

where

$$|r^-_{i,x}| \le \left(\widetilde{C}^-_x + \frac{L_x}{2}\right) \|\Delta_i\|^3.$$

Using the product structure of P_i^0 and P_i^1, we get

$$D_{\mathrm{KL}}(\mathsf{P}_i^0 \| \mathsf{P}_i^1) = \sum_{b\in\{0,1\}} D_{\mathrm{KL}}(P^{x_{i,b}}_{\theta_i^0} \| P^{x_{i,b}}_{\theta_i^1}) = m^{\mathrm{traj}}_{F,i}(\Delta_i) + R_i^+,$$

where

$$R_i^+ \coloneqq r^+_{i,x_{i,0}} + r^+_{i,x_{i,1}}.$$

Similarly,

$$D_{\mathrm{KL}}(\mathsf{P}_i^1 \| \mathsf{P}_i^0) = \sum_{b\in\{0,1\}} D_{\mathrm{KL}}(P^{x_{i,b}}_{\theta_i^1} \| P^{x_{i,b}}_{\theta_i^0}) = m^{\mathrm{traj}}_{F,i}(\Delta_i) + R_i^-,$$

where

$$R_i^- \coloneqq r^-_{i,x_{i,0}} + r^-_{i,x_{i,1}}.$$

Since the third-order Taylor remainder constants and the Fisher Lipschitz constants are uniformly bounded over all i, $(\widetilde{S}, U_{-i})$, and candidate prompts $x \in \{x_{i,0}, x_{i,1}\}$, we can absorb all constants into a single deterministic constant $C_3 < \infty$, obtaining

$$|R_i^+| + |R_i^-| \le C_3 \|\Delta_i\|^3.$$

Therefore,

$$\begin{aligned} I_i &\le \frac{1}{4}\left[m^{\mathrm{traj}}_{F,i}(\Delta_i) + R_i^+ + m^{\mathrm{traj}}_{F,i}(\Delta_i) + R_i^-\right] \\ &= \frac{1}{2} m^{\mathrm{traj}}_{F,i}(\Delta_i) + \frac{1}{4}(R_i^+ + R_i^-). \end{aligned}$$

Hence

$$I_i \le \frac{1}{2} m^{\mathrm{traj}}_{F,i}(\Delta_i) + \frac{C_3}{4}\|\Delta_i\|^3.$$

Substituting this into the CMI bound gives

$$\begin{aligned} |\mathbb{E}[\widetilde{\mathrm{gen}}(A)]| &\le \sqrt{\frac{2}{n}\sum_{i=1}^n \mathbb{E}[I_i]} \\ &\le \sqrt{\frac{2}{n}\sum_{i=1}^n \mathbb{E}\left[\frac{1}{2} m^{\mathrm{traj}}_{F,i}(\Delta_i) + \frac{C_3}{4}\|\Delta_i\|^3\right]} \\ &= \sqrt{\frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[m^{\mathrm{traj}}_{F,i}(\Delta_i) + \frac{C_3}{2}\|\Delta_i\|^3\right]}. \end{aligned}$$

This proves the theorem. □

A.2 KL-SAM Analysis

Local KL-SAM setup. For Proposition A.4 and Theorem A.5, fix an optimization step k and write

$$\mathcal{L}_k(\theta) = \mathcal{L}_{\mathrm{GRPO}}(\theta)$$

for the GRPO minimization loss at this step. Let

$$g_k := \nabla_\theta \mathcal{L}_k(\theta_k), \qquad H_k := \nabla_\theta^2 \mathcal{L}_k(\theta_k), \qquad \widehat{g}_k := \frac{g_k}{\|g_k\|_2},$$

and assume $g_k \neq 0$. Define the first-order SAM perturbation

$$\epsilon_{\mathrm{sam},k} := \rho \widehat{g}_k.$$

Let F_k be the trajectory Fisher matrix of π_{θ_k}, averaged over the prompt distribution μ, and define the local clean-vs-perturbed trajectory KL

$$D_{\mathrm{loc},k}(\Delta) := \mathbb{E}_{x\sim\mu} D_{\mathrm{KL}}\left(P^x_{\theta_k} \middle\| P^x_{\theta_k+\Delta}\right).$$

The corresponding KL-SAM perturbed loss is

$$\mathcal{L}_k^{\text{KL-SAM}} := \mathcal{L}_k(\theta_k + \epsilon_{\mathrm{sam},k}) + \lambda_{\mathrm{loc}} D_{\mathrm{loc},k}(\epsilon_{\mathrm{sam},k}).$$

Finally, write

$$m_{H,k}(\Delta) := g_k^\top \Delta + \frac{1}{2}\Delta^\top H_k \Delta, \qquad m_{F,k}(\Delta) := \frac{1}{2}\Delta^\top F_k \Delta.$$

Proposition A.4 (First-order equivalence of SAM and KL-SAM perturbations)**.** *Under the local KL-SAM setup above, suppose $D_{\mathrm{loc},k}$ is differentiable at $\epsilon = 0$. Then*

$$D_{\mathrm{loc},k}(0) = 0, \qquad \nabla_\epsilon D_{\mathrm{loc},k}(\epsilon)\big|_{\epsilon=0} = 0.$$

Consequently, the first-order inner maximizers of SAM and KL-SAM over the ℓ_2 ball coincide, and the unique maximizer is

$$\epsilon^\star_{\text{KL-SAM}} = \epsilon^\star_{\text{SAM}} = \rho \frac{g_k}{\|g_k\|_2}.$$

Proof. By nonnegativity of KL divergence,

$$D_{\mathrm{loc},k}(\epsilon) = \mathbb{E}_{x\sim\mu} D_{\mathrm{KL}}(P^x_{\theta_k} \| P^x_{\theta_k+\epsilon}) \geq 0.$$

Moreover,

$$D_{\mathrm{loc},k}(0) = \mathbb{E}_{x\sim\mu} D_{\mathrm{KL}}(P^x_{\theta_k} \| P^x_{\theta_k}) = 0.$$

Thus $\epsilon = 0$ is a minimizer of $D_{\mathrm{loc},k}$. Since $D_{\mathrm{loc},k}$ is differentiable at 0, it follows that

$$\nabla_\epsilon D_{\mathrm{loc},k}(0) = 0.$$

Therefore the first-order Taylor models of $\mathcal{L}_k(\theta_k + \epsilon)$ and

$$\mathcal{L}_k(\theta_k + \epsilon) + \lambda_{\mathrm{loc}} D_{\mathrm{loc},k}(\epsilon)$$

have the same linear term $g_k^\top \epsilon$. Since $g_k \neq 0$, the unique maximizer of $g_k^\top \epsilon$ over $\|\epsilon\|_2 \leq \rho$ is

$$\epsilon = \rho \frac{g_k}{\|g_k\|_2}.$$

□

Theorem A.5 (KL-SAM implicitly regularizes Hessian–Fisher curvature)**.** *Under the local KL-SAM setup above, assume that $\mathcal{L}_k$ is three times continuously differentiable around θ_k, with locally bounded third derivative, and that Assumption A.1 holds locally with (θ_i^0, θ_i^1) replaced by $(\theta_k, \theta_k + \epsilon)$, uniformly for sufficiently small ϵ and for $x \sim \mu$. Then*

$$\mathcal{L}_k^{\text{KL-SAM}} = \mathcal{L}_k(\theta_k) + \rho\|g_k\|_2 + \frac{\rho^2}{2}\widehat{g}_k^\top \left(H_k + \lambda_{\mathrm{loc}} F_k\right) \widehat{g}_k + O(\rho^3).$$

Equivalently,

$$\mathcal{L}_k^{\text{KL-SAM}} = \mathcal{L}_k(\theta_k) + m_{H,k}(\epsilon_{\mathrm{sam},k}) + \lambda_{\mathrm{loc}} m_{F,k}(\epsilon_{\mathrm{sam},k}) + O(\rho^3).$$

Proof. By Taylor expansion of the GRPO loss around θ_k,

$$\mathcal{L}_k(\theta_k + \epsilon_{\text{sam},k}) = \mathcal{L}_k(\theta_k) + g_k^\top \epsilon_{\text{sam},k} + \frac{1}{2}\epsilon_{\text{sam},k}^\top H_k \epsilon_{\text{sam},k} + O(\rho^3).$$

Since $\epsilon_{\text{sam},k} = \rho\widehat{g}_k$, we have

$$g_k^\top \epsilon_{\text{sam},k} = \rho\|g_k\|_2.$$

Moreover, by Lemma A.2, the local trajectory KL admits the Fisher expansion

$$D_{\text{loc},k}(\epsilon_{\text{sam},k}) = \frac{1}{2}\epsilon_{\text{sam},k}^\top F_k \epsilon_{\text{sam},k} + O(\rho^3).$$

Combining the two expansions gives the result. □

Remark A.6 (Clipped GRPO objectives)**.** *When clipping is active in GRPO objectives, the smoothness assumption above may fail at branch-switching points. In this case, the expansion is understood on regions where the active clipping branch is locally fixed; alternatively, one may replace the clip by a smooth surrogate.*

B Algorithm Details

Algorithm 1 KL-SAM (one training step)

Require: model parameters θ, batch $\mathcal{B}$, perturbation radius ρ, local KL coefficient λ_{loc}

$W \leftarrow$ `lm_head.weight`
Clean forward pass:
  Compute hidden states H, clean logits $Z = \text{linear}(H, W)$, and pre-perturbation RL loss ℓ_{pre}
$g \leftarrow \nabla_W \ell_{\text{pre}}$
$\epsilon \leftarrow \text{stopgrad}(\rho\, g/\|g\|_2)$
Perturbed forward pass:
  $Z^+ \leftarrow \text{linear}(H, W + \epsilon)$
  Compute perturbed RL loss ℓ_{post} from Z^+
  Compute local policy KL $d_{\text{loc}} \leftarrow \mathbb{E}_{x\in\mathcal{B}} D_{\text{KL}}(\pi_\theta(\cdot \mid x)\|\pi_{\theta+\epsilon}(\cdot \mid x))$
  $\ell_{\text{KL-SAM}} \leftarrow \ell_{\text{post}} + \lambda_{\text{loc}} d_{\text{loc}}$
Update:
  Backpropagate $\ell_{\text{KL-SAM}}$
  Optimizer step on θ

Here, $\theta + \epsilon$ denotes perturbing only the output head W.
stopgrad detaches ϵ, treating it as constant during the outer backward pass.
The clean policy is not detached: local-KL gradients flow through both policy branches.

C Implementation Details and Hyperparameters

Distributed training. For distributed training, Qwen2.5-Math-1.5B and DeepSeek-R1-Distill-Qwen-1.5B are trained on $2\times$ A100 GPUs with DeepSpeed ZeRO-2 (Rasley et al., 2020), while Qwen2.5-Math-7B is trained on $4\times$ H100 GPUs with DeepSpeed ZeRO-3.

Hyperparameters. Table 4 lists the hyperparameters used for the results in Tables 2 and 3.

D Additional Multi-Seed Results

Table 5 supplements Table 2 with per-benchmark SDs for Qwen2.5-Math-1.5B. Table 6 reports final-checkpoint performance after one training epoch, using the same hyperparameters as the main results.

Table 4: Hyperparameters used for the main experimental results.

Model	Loss	KL-SAM ρ	KL-SAM λ_{loc}	logits-SAM ρ
Qwen2.5-Math-1.5B	GRPO	1×10^{-2}	5.0	5×10^{-3}
Qwen2.5-Math-1.5B	DR.GRPO	1×10^{-3}	10.0	1×10^{-3}
DeepSeek-R1-Distill-Qwen-1.5B	GRPO	5×10^{-3}	1.0	1×10^{-2}
DeepSeek-R1-Distill-Qwen-1.5B	DR.GRPO	5×10^{-3}	1.0	1×10^{-3}
Qwen2.5-Math-7B	GRPO	5×10^{-3}	1.0	1×10^{-2}
Qwen2.5-Math-7B	DR.GRPO	1×10^{-3}	5.0	1×10^{-2}

Table 5: Pass@1 (%) for Qwen2.5-Math-1.5B, reported as mean $\pm$ SD over three training seeds. Bold denotes the best mean within each objective.

Loss	Method	MATH500	AIME24	AMC23	Minerva	Olympiad	Avg.
GRPO	Baseline	$69.55_{\pm 1.25}$	$9.72_{\pm 2.19}$	$48.75_{\pm 2.04}$	$22.00_{\pm 1.24}$	$33.46_{\pm 1.44}$	$36.70_{\pm 0.23}$
	logits-SAM	$70.08_{\pm 1.08}$	$11.94_{\pm 1.71}$	$\mathbf{52.08}_{\pm 1.06}$	$\mathbf{24.36}_{\pm 1.87}$	$33.37_{\pm 0.49}$	$38.37_{\pm 0.28}$
	KL-SAM	$\mathbf{70.62}_{\pm 1.11}$	$11.94_{\pm 3.07}$	$51.88_{\pm 5.30}$	$24.05_{\pm 0.62}$	$\mathbf{34.26}_{\pm 0.40}$	$\mathbf{38.55}_{\pm 1.11}$
	KL-only	$70.42_{\pm 0.10}$	$\mathbf{12.22}_{\pm 0.79}$	$47.08_{\pm 2.52}$	$23.31_{\pm 0.45}$	$33.17_{\pm 0.18}$	$37.24_{\pm 0.60}$
DR.GRPO	Baseline	$67.02_{\pm 0.61}$	$12.50_{\pm 3.12}$	$48.33_{\pm 3.32}$	$22.24_{\pm 2.06}$	$31.38_{\pm 0.18}$	$36.30_{\pm 0.83}$
	logits-SAM	$67.12_{\pm 1.34}$	$11.94_{\pm 1.57}$	$47.92_{\pm 2.81}$	$22.37_{\pm 0.34}$	$31.46_{\pm 0.59}$	$36.16_{\pm 0.47}$
	KL-SAM	$\mathbf{68.82}_{\pm 0.68}$	$11.11_{\pm 2.39}$	$\mathbf{50.00}_{\pm 2.22}$	$\mathbf{22.79}_{\pm 1.11}$	$\mathbf{32.96}_{\pm 0.76}$	$\mathbf{37.14}_{\pm 0.80}$
	KL-only	$64.83_{\pm 2.56}$	$\mathbf{13.61}_{\pm 1.04}$	$45.21_{\pm 3.28}$	$22.27_{\pm 1.16}$	$30.46_{\pm 2.56}$	$35.28_{\pm 1.61}$

Table 6: Final-checkpoint pass@1 (%) for Qwen2.5-Math-1.5B after one training epoch, reported as mean $\pm$ SD over three training seeds. Bold denotes the best mean within each objective.

Loss	Method	MATH500	AIME24	AMC23	Minerva	Olympiad	Avg.
GRPO	Baseline	$67.62_{\pm 2.23}$	$9.44_{\pm 1.71}$	$43.33_{\pm 1.93}$	$23.19_{\pm 1.08}$	$30.96_{\pm 2.33}$	$34.91_{\pm 0.59}$
	logits-SAM	$70.53_{\pm 0.45}$	$10.83_{\pm 0.68}$	$50.21_{\pm 3.69}$	$\mathbf{24.33}_{\pm 1.93}$	$33.33_{\pm 0.26}$	$37.85_{\pm 0.83}$
	KL-SAM	$\mathbf{70.62}_{\pm 1.11}$	$\mathbf{11.94}_{\pm 3.07}$	$\mathbf{51.88}_{\pm 5.30}$	$24.05_{\pm 0.62}$	$\mathbf{34.26}_{\pm 0.40}$	$\mathbf{38.55}_{\pm 1.11}$
	KL-only	$70.28_{\pm 0.28}$	$11.11_{\pm 0.79}$	$45.62_{\pm 0.51}$	$23.13_{\pm 0.54}$	$33.11_{\pm 0.27}$	$36.65_{\pm 0.24}$
DR.GRPO	Baseline	$68.25_{\pm 1.86}$	$10.28_{\pm 3.36}$	$43.33_{\pm 1.28}$	$\mathbf{24.05}_{\pm 0.57}$	$31.79_{\pm 1.48}$	$35.54_{\pm 1.49}$
	logits-SAM	$67.03_{\pm 1.83}$	$\mathbf{11.94}_{\pm 1.71}$	$42.92_{\pm 2.52}$	$22.64_{\pm 0.48}$	$31.21_{\pm 0.81}$	$35.15_{\pm 1.14}$
	KL-SAM	$\mathbf{69.00}_{\pm 1.54}$	$8.89_{\pm 1.71}$	$\mathbf{49.79}_{\pm 0.78}$	$23.22_{\pm 0.64}$	$\mathbf{33.19}_{\pm 1.05}$	$\mathbf{36.82}_{\pm 0.56}$
	KL-only	$62.23_{\pm 7.00}$	$9.72_{\pm 2.08}$	$43.96_{\pm 4.12}$	$22.95_{\pm 1.22}$	$29.42_{\pm 5.16}$	$33.66_{\pm 3.43}$