EXPLORATION PRIOR: SEMANTICALLY GUIDED EXPLORATION FOR REINFORCEMENT LEARNING OF LANGUAGE MODEL REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement learning with verifiable rewards (RLVR) substantially improves the reasoning capabilities of large language models, yet effective exploration remains challenging as policy optimization can progressively concentrate probability mass within a narrow region of the continuation space. We introduce EXPLORATION PRIOR (EP), a framework that uses the model's natural-language conditioning capability to provide semantic guidance during RL training. EP uses lightweight, answer-free path-switching instructions to construct an intervention-conditioned continuation distribution as an exploration prior. Rather than specifying a correct reasoning path, the intervention shifts the continuation distribution toward alternative reasoning directions, while task rewards determine which sampled continuations are reinforced through policy optimization. From a probabilistic-inference perspective, the reward tilts this prior toward high-reward continuations, yielding a posterior that is approximated through reverse-KL-regularized policy optimization. Across five mathematical reasoning benchmarks and three model backbones, EP consistently outperforms standard RLVR and representative exploration- and prior-based baselines. On Qwen3-4B, EP improves the mean Avg@16 from 46.6% to 51.9% and the mean Pass@16 from 67.8% to 71.5% over GRPO. Moreover, its margin over GRPO in Pass@K becomes more pronounced at larger sampling budgets, indicating broader coverage of successful reasoning trajectories.

1 INTRODUCTION

Reinforcement learning with verifiable rewards (RLVR) has substantially improved the reasoning capabilities of large language models (LLMs) (Shao et al., 2024; Guo et al., 2025). However, effective exploration remains challenging in reasoning tasks. The large token-level action space and long reasoning horizon induce a vast space of intermediate reasoning states and trajectories, while policy optimization can progressively concentrate probability mass within a limited region of the continuation space. Recent studies have accordingly observed decreasing policy entropy and increasingly homogeneous rollouts during RLVR training (Yu et al., 2025; Cui et al., 2025; Wu et al., 2026). Such concentration can make reasoning trajectories that receive low probability under the current policy increasingly difficult to discover, even when they may yield high reward.

Existing approaches promote exploration through mechanisms such as entropy-aware policy updates, selective updates to high-entropy tokens, trajectory-level diversity objectives, and set-level optimization (Cui et al., 2025; Wang et al., 2025; Chen et al., 2025; Hu et al., 2026). Although effective, these methods generally encourage broader exploration while largely leaving the semantic direction of exploration implicit. In this work, we investigate a complementary source of exploration already available within the language model: its ability to reshape the distribution over reasoning continuations through natural-language guidance.

We introduce EXPLORATION PRIOR (EP), which uses lightweight, answer-free path-switching guidance to construct a context-dependent prior for reasoning exploration. Rather than specifying a correct reasoning path, the intervention reshapes the relative likelihoods of candidate continuations

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.

through natural-language conditioning. The intervention is used only to construct the prior, while training trajectories remain sampled from the current policy.

From a probabilistic-inference perspective, the task reward tilts the intervention-conditioned prior toward high-reward continuations, yielding a posterior that is approximated through reverse-KL-regularized policy optimization. We further anneal the exploration-prior coefficient over training, placing greater emphasis on semantically guided exploration early in optimization and gradually shifting emphasis toward the task reward.

Across five mathematical reasoning benchmarks and three model backbones, EP consistently outperforms standard RLVR and representative exploration- and prior-based baselines. improves the mean Avg@16 from 46.6% to 51.9% and the mean Pass@16 from 67.8% to 71.5% over GRPO. Its advantage in Pass@K becomes more pronounced at larger sampling budgets, together with higher rollout diversity, indicating broader coverage of successful reasoning trajectories.

Our main contributions are:

- We introduce EXPLORATION PRIOR, a framework that leverages natural-language conditioning to construct answer-free, context-dependent priors for reasoning exploration in RLVR.
- We formulate EP from a probabilistic-inference perspective, in which task rewards tilt an intervention-conditioned continuation prior toward high-reward continuations, yielding a practical reverse-KL-regularized RL objective with annealed prior strength.
- We demonstrate consistent improvements across five mathematical reasoning benchmarks and three model backbones, together with evidence that EP increases rollout diversity and improves successful-solution coverage as the inference-time sampling budget grows.

2 RELATED WORK

2.1 EXPLORATION IN REINFORCEMENT LEARNING FOR REASONING

Reinforcement learning with verifiable rewards (RLVR) can reduce policy entropy and produce increasingly homogeneous rollouts during training (Yu et al., 2025; Cui et al., 2025; Wu et al., 2026), motivating methods that explicitly encourage exploration. Existing approaches promote exploration through several mechanisms. One line modifies policy optimization using uncertainty- or entropy-related signals. Clip-Cov and KL-Cov constrain updates associated with substantial entropy reduction (Cui et al., 2025), while Wang et al. (2025) selectively update a small fraction of high-entropy tokens to better balance exploration and task performance. These methods identify where additional exploration may be beneficial, while largely leaving the semantic direction of exploration implicit. A second line introduces trajectory-level objectives. Pass@k Training optimizes set-level success (Chen et al., 2025), DIVER promotes diversity across reasoning trajectories (Hu et al., 2026), and SAGE reshapes the reverse-KL anchor using entropy and surprisal signals (Lee et al., 2026). A complementary line modifies the reasoning process itself. E3 performs in-context exploration through iterative generation, verification, and refinement over multiple hypotheses (Setlur et al., 2026), while Length-Incentivized Exploration encourages broader state coverage through length-based rewards and redundancy penalties (**?**). In contrast, EP uses natural-language semantic guidance to construct a context-dependent prior over reasoning continuations, thereby reshaping the continuation distribution through semantic conditioning. The intervention is used only to construct the prior, while task rewards provide the learning signal for policy updates on trajectories sampled from the current policy.

2.2 KL-REGULARIZED RL AND POLICY PRIORS

KL regularization is widely used in language-model post-training to balance reward optimization against deviation from a reference policy, and admits a probabilistic-inference interpretation in which reward tilts a prior distribution toward desirable behaviors (Levine, 2018; Ziegler et al., 2019; Stiennon et al., 2020; Ouyang et al., 2022). In reasoning-oriented RL, however, strong regularization toward a fixed reference can restrict policy exploration, and recent RLVR methods often remove this constraint (Shao et al., 2024; Yu et al., 2025; Hu et al., 2025). A related line of work constructs

auxiliary distributions through teacher or context conditioning. On-policy distillation matches student predictions to a teacher on student-generated trajectories (Lu & Lab, 2025), while recent self- and context-distillation methods derive teacher distributions from privileged reasoning traces, experiential knowledge, behavioral instructions, environment feedback, or self-generated successful solutions (Zhao et al., 2026; Ye et al., 2026; Hübotter et al., 2026). Several approaches further combine such dense supervision with RLVR objectives (Hübotter et al., 2026; Yang et al., 2026; Team et al., 2026). EP differs in the role assigned to the conditioned distribution. Rather than treating it as a fixed reference to remain close to or a teacher target to imitate, EP uses an answer-free conditioned distribution as a policy prior for exploration, with task reward determining which induced behaviors are reinforced. EP differs in how it uses the conditioned distribution. Rather than treating it as a fixed reference for the policy to stay close to or as a teacher distribution to imitate, EP uses an answer-free conditioned distribution as an exploration prior. Training trajectories are still sampled from the current policy, and task rewards provide the learning signal for policy optimization.

3 PRELIMINARIES

Reinforcement Learning with Verifiable Rewards. Let $\mathcal{D}$ denote the training distribution over problems. Given $x \sim \mathcal{D}$, a policy π_θ generates a reasoning trajectory $y \sim \pi_\theta(\cdot \mid x)$ and receives a verifiable reward $R(x, y)$. RLVR maximizes

$$\mathcal{J}_{\mathrm{RL}}(\theta) = \mathbb{E}_{x\sim\mathcal{D},\, y\sim\pi_\theta(\cdot|x)}\left[R(x, y)\right],$$

typically using policy-gradient methods (Williams & Peng, 1991).

GRPO (Shao et al., 2024) is a critic-free policy optimization method widely used for RLVR. For each problem x, it samples G trajectories $\{y_i\}_{i=1}^{G}$ from a rollout policy $\pi_{\theta_{\mathrm{old}}}$. Let $R_i := R(x, y_i)$. The group-relative advantage is

$$A_i = \frac{R_i - \mathrm{mean}(\{R_j\}_{j=1}^{G})}{\mathrm{std}(\{R_j\}_{j=1}^{G})}. \tag{1}$$

GRPO then performs PPO-style clipped updates using the token-level importance ratio

$$r_{i,t}(\theta) = \frac{\pi_\theta(y_{i,t} \mid x, y_{i,<t})}{\pi_{\theta_{\mathrm{old}}}(y_{i,t} \mid x, y_{i,<t})}. \tag{2}$$

Our implementation is related to on-policy self-distillation in that the conditioned distribution is evaluated on prefixes sampled by the rollout policy. We review the corresponding objective in Appendix B.

4 METHOD

4.1 CONSTRUCTING THE EXPLORATION PRIOR

Large language models can substantially alter their continuation distributions in response to changes in natural-language context. We leverage this conditioning capability to construct a continuation distribution that differs from the policy's default continuation behavior and use it as an exploration prior.

Given a problem x and an on-policy reasoning trajectory $y = (y_{<k}, y_{k:T})$ sampled from the current policy π_θ, we select an intermediate reasoning position k and append an answer-free natural-language intervention e to the prefix $y_{<k}$. The intervention contains no reference solution, correct answer, or external feedback. We then evaluate a frozen snapshot of the policy under the new context and define the resulting conditional distribution as

$$\pi_{\bar\theta}^{e}(\cdot \mid x, y_{<k}) \triangleq \pi_{\bar\theta}(\cdot \mid x, y_{<k}, e), \tag{3}$$

where $\bar\theta$ denotes the frozen policy snapshot used to construct the prior, and gradients are stopped through this distribution. At the beginning of each policy update, we set $\bar\theta \leftarrow \theta$ and keep $\bar\theta$ fixed throughout the update.

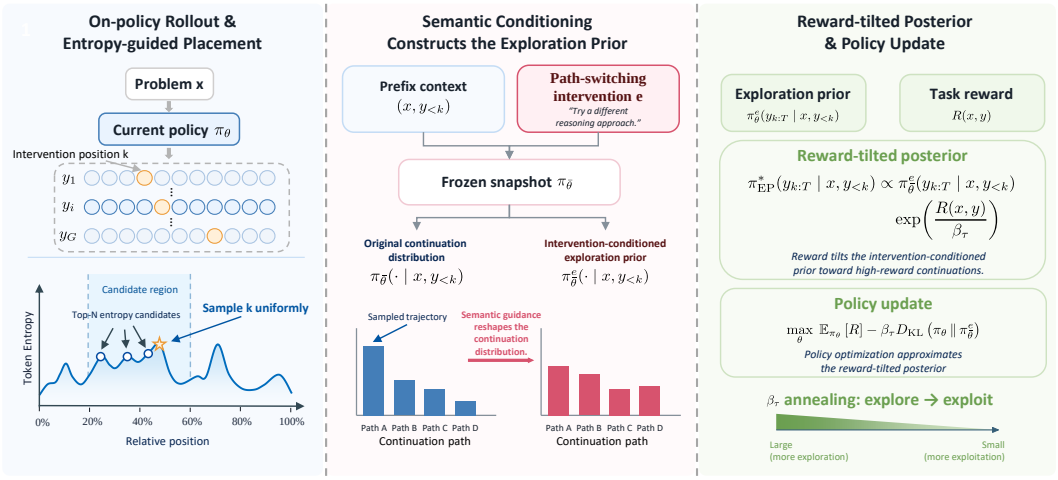

Figure 1: **Overview of Exploration Prior (EP). (1)** The current policy generates on-policy reasoning trajectories. Within a predefined candidate region, we select high-entropy intervention candidates and uniformly sample the intervention position k from them. **(2)** An answer-free path-switching intervention conditions a frozen policy snapshot to construct the exploration prior, thereby reshaping the relative distribution over possible continuations. **(3)** Task rewards tilt the exploration prior toward high-reward continuations, yielding a posterior target that is approximated through reverse-KL policy optimization. Prior regularization is applied only to the continuation suffix and is gradually annealed over the course of training. The intervention is used solely to condition the exploration prior, while all training rollouts remain on-policy.

In our implementation, we use path switching as a simple and effective form of semantically meaningful reasoning guidance that encourages reconsideration of the current reasoning trajectory. The intervention does not need to specify which reasoning path is correct or preferable. Instead, it induces a context-dependent shift in the relative likelihoods of candidate continuations, while the task reward determines which of the resulting behaviors are reinforced. We further discuss this interaction in Section 4.2.

Entropy-guided Intervention Placement. Prior work has shown that token-level predictive entropy provides a useful signal for identifying uncertain decision points in reasoning trajectories (Wang et al., 2025). We therefore use token entropy as a heuristic for selecting positions at which to construct the exploration prior. For each trajectory, we restrict candidate intervention positions to the relative range 0.2–0.6 of the generated response. For each candidate position t, we compute

$$H_t = -\sum_{v \in \mathcal{V}} \pi_{\bar{\theta}}(v \mid x, y_{<t}) \log \pi_{\bar{\theta}}(v \mid x, y_{<t}), \tag{4}$$

where $\mathcal{V}$ denotes the vocabulary.

We select the 20 positions with the highest predictive entropy and uniformly sample the intervention position k from this set. This procedure constructs the prior at reasoning states where the current policy has higher uncertainty.

4.2 Reasoning as Posterior Inference

Following the probabilistic modeling view of language model cascades and the inference interpretation of KL-regularized reinforcement learning (Dohan et al., 2022; Korbak et al., 2022), we treat the reasoning continuation $y_{k:T}$ as a latent variable conditioned on the problem x and reasoning prefix $y_{<k}$. The intervention-conditioned distribution $\pi^e_{\bar{\theta}}$ defines a prior over possible continuations, while the task reward induces an optimality likelihood proportional to

$$\exp(R(x, y)/\beta_\tau)\,,$$

where $\beta_\tau > 0$ controls the relative influence of the prior and reward at optimization step τ.

Combining this continuation prior with the reward likelihood yields the reward-tilted posterior

$$\pi^*_{\mathrm{EP}}(y_{k:T} \mid x, y_{<k}) = \frac{\pi^e_\theta(y_{k:T} \mid x, y_{<k}) \exp(R(x,y)/\beta_\tau)}{Z^e_\tau(x, y_{<k})}, \tag{5}$$

where $Z^e_\tau(x, y_{<k})$ denotes the corresponding normalizing constant.

This formulation separates the roles of semantic guidance and task reward. The intervention reshapes the distribution over reasoning continuations, while the task reward provides the learning signal for reinforcing or suppressing the sampled continuations.

4.3 Exploration-Prior KL-Regularized RL

We approximate the reward-tilted posterior in equation 5 by minimizing the reverse KL from the continuation policy to π^*_{EP}. Treating the exploration prior as fixed within each policy update, this is equivalent to maximizing

$$\begin{aligned}\mathcal{J}_{\mathrm{EP}}(\theta; x, y_{<k}) =& \mathbb{E}_{y_{k:T}\sim\pi_\theta(\cdot|x,y_{<k})}\left[R(x,y)\right] \\ &- \beta_\tau D_{\mathrm{KL}}\left(\pi_\theta(\cdot \mid x, y_{<k}) \,\|\, \pi^e_{\bar\theta}(\cdot \mid x, y_{<k})\right).\end{aligned} \tag{6}$$

The reward is evaluated on the full trajectory, whereas the exploration-prior KL is applied only to the suffix following the intervention position. The posterior construction and equivalence are derived in Lemma H.2.

Relation to Entropy Regularization. Entropy regularization can also be interpreted through the lens of prior regularization. On a fixed finite space of reasoning continuations $\mathcal{Y}_{k:T}$, let U denote the uniform distribution over this space. Then

$$\mathbb{E}_{\pi_\theta}[R] + \beta_\tau H(\pi_\theta) = \mathbb{E}_{\pi_\theta}[R] - \beta_\tau D_{\mathrm{KL}}(\pi_\theta \| U) + \beta_\tau \log |\mathcal{Y}_{k:T}|, \tag{7}$$

where the conditioning on $(x, y_{<k})$ is omitted for brevity. Thus, up to a constant independent of θ, entropy regularization corresponds to reverse-KL regularization toward a uniform prior. In contrast, EP regularizes the policy toward a context-dependent prior induced by semantic reasoning guidance, thereby introducing structured preferences over candidate continuations rather than encouraging all continuations uniformly.

Sampled Gradient of the Exploration KL. For a trajectory y_i with intervention position k_i, define the sampled prior-induced likelihood shift for $t \geq k_i$ as

$$\delta^e_{i,t}(\theta) = \log \frac{\pi^e_{\bar\theta}\left(y_{i,t} \mid x, y_{i,<t}\right)}{\pi_\theta\left(y_{i,t} \mid x, y_{i,<t}\right)}. \tag{8}$$

A positive $\delta^e_{i,t}$ indicates that the exploration prior assigns greater likelihood to the sampled token than the current policy.

Following prior work on token-level on-policy distillation (Fu et al., 2026), we use an immediate-term Monte Carlo approximation to the sequence-level KL gradient:

$$\nabla_\theta D_{\mathrm{KL}}\left(\pi_\theta \,\|\, \pi^e_{\bar\theta}\right) \approx -\mathbb{E}_{y_{i,k_i:T}\sim\pi_{\theta_{\mathrm{old}}}}\left[\sum_{t=k_i}^{T} r_{i,t}(\theta)\, \mathrm{sg}\left[\delta^e_{i,t}(\theta)\right] \nabla_\theta \log \pi_\theta\left(y_{i,t} \mid x, y_{i,<t}\right)\right], \tag{9}$$

where the common conditioning on $(x, y_{i,<k_i})$ in the suffix distributions is omitted for brevity. This approximation discards future-token dependencies in the exact sequence-level KL gradient and retains only the local prior-induced likelihood shift associated with each sampled token. We provide the derivation in Lemma H.3.

Stabilizing the Prior Signal. To prevent a small number of large prior-induced shifts from dominating the update, we clip the signal as

$$\bar\delta^e_{i,t} = \mathrm{clip}(\delta^e_{i,t}, -c, c). \tag{10}$$

Combining this term with the GRPO reward advantage yields the token-level effective advantage

$$A_{i,t}^{\mathrm{EP}} = A_i + \beta_\tau \mathbf{1}\{t \geq k_i\} \, \mathrm{sg}[\bar{\delta}_{i,t}^{e}]. \tag{11}$$

Thus, sampled tokens assigned higher likelihood by the exploration prior receive a positive adjustment, whereas those downweighted by the prior receive a negative adjustment. We substitute $A_{i,t}^{\mathrm{EP}}$ into the standard GRPO clipped surrogate. The complete surrogate objective and optimization algorithm are provided in Appendix C.

Motivated by the classic exploration–exploitation trade-off in reinforcement learning (Sutton & Barto, 2018), we decay the exploration-prior coefficient β_τ over training. This places greater emphasis on semantically guided exploration early in training and gradually shifts optimization toward the task reward.

5 Experiments

We conduct experiments to answer four questions:
(1) Does EP improve mathematical reasoning performance over standard RLVR baselines, exploration-enhancing methods, and alternative policy priors? **Yes.** (Section 5.2)
(2) Do these gains generalize across model scales and families? **Yes.** (Section 5.2)
(3) Does EP enable broader and more effective exploration of successful reasoning trajectories? **Yes.** (Section 5.3)
(4) Which components and design choices contribute to EP's performance? (Section 5.4)

5.1 Experimental Settings

Datasets and Evaluation. We use DAPO-MATH-17K (Yu et al., 2025) as the training set for all RL experiments and evaluate on AIME 2024, AIME 2025, AMC23, Minerva Math, and HMMT25. For each problem, Avg@16 is the mean correctness over 16 sampled responses, whereas Pass@16 indicates whether at least one of the 16 responses is correct. We aggregate each metric across the problems in a benchmark. We report both metrics as percentages. All methods use the same decoding configuration and sampling budget.

Baselines and Training. We compare EP with three categories of baselines. **1) Standard RLVR baselines: GRPO** denotes GRPO (Shao et al., 2024) without reference-policy KL regularization, whereas **GRPO w/ Ref-KL** additionally regularizes the policy toward a fixed reference model. **2) Exploration-enhancing methods: GRPO+Entropy** selectively constrains updates at tokens associated with substantial entropy reduction to mitigate policy entropy collapse and has been reported to provide more stable entropy control than directly optimizing an entropy bonus (Cui et al., 2025). **3) Alternative prior construction: GRPO+SDPO** constructs a context-conditioned prior from self-generated successful trajectories, following the use of successful rollouts as implicit feedback in SDPO (Hübotter et al., 2026). We incorporate this prior into the same KL-regularized GRPO framework as EP to isolate the effect of prior construction.

Training Protocol. At the beginning of each policy update, the prior policy $\pi_{\bar{\theta}}$ is initialized from the current policy and then held fixed throughout the update. We use Qwen3-4B as the primary model for comparison. We further evaluate Qwen3-8B to assess transfer across model scales and OLMo3-7B-Instruct to assess transfer across model families. Across all methods, we keep the training data, rollout and evaluation configurations, verifier rewards for RLVR-based methods, and sampling budget fixed. Detailed hyperparameters are provided in Appendix E.

5.2 Main Results

Overall Performance. Table 1 compares EP with all baselines on Qwen3-4B. EP achieves the best mean performance on both metrics, reaching 51.9 Avg@16 and 71.5 Pass@16. Compared with GRPO, this corresponds to improvements of 5.3 and 3.7 percentage points, respectively. EP also outperforms GRPO+Entropy, the strongest non-EP baseline in mean Avg@16, by 3.4 points in Avg@16 and 6.5 points in Pass@16.

Table 1: Main results on Qwen3-4B across mathematical reasoning benchmarks. We report Avg@16 and Pass@16 for each benchmark. Best results in each column are shown in **bold**.

Method	AIME24		AIME25		AMC23		Minerva		HMMT25		AVG	
	Avg@16	Pass@16	Avg@16	Pass@16	Avg@16	Pass@16	Avg@16	Pass@16	Avg@16	Pass@16	Avg@16	Pass@16
Qwen3-4B	21.0	46.7	20.2	43.3	67.8	90.0	40.3	53.3	12.3	23.3	32.3	51.3
GRPO	45.2	**76.7**	36.7	66.7	86.6	**97.5**	42.9	54.8	21.7	43.3	46.6	67.8
GRPO w/ Ref-KL	44.8	70.0	38.8	63.3	86.0	95.0	44.2	55.1	20.0	40.0	46.8	64.7
GRPO+Entropy	51.7	73.3	39.8	66.7	88.0	**97.5**	43.0	54.4	20.2	33.3	48.5	65.0
GRPO+SDPO	17.1	43.3	17.5	33.3	60.8	87.5	38.6	51.5	7.3	20.0	28.3	47.1
EP (Ours)	**54.4**	**76.7**	**45.4**	**73.3**	**89.8**	**97.5**	**44.4**	**56.6**	**25.6**	**53.3**	**51.9**	**71.5**

The choice of prior is also important. GRPO w/ Ref-KL improves mean Avg@16 over GRPO by only 0.2 points while reducing mean Pass@16 by 3.1 points, whereas GRPO+SDPO performs below both GRPO and Qwen3-4B on both metrics. While SDPO reports improvements on scientific reasoning, tool use, and competitive programming (Hübotter et al., 2026), recent work has also shown that self-distillation can degrade performance on challenging mathematical reasoning tasks (Kim et al., 2026). Our results exhibit a similar degradation in this setting, suggesting that the effectiveness of self-generated priors can be task-dependent. These comparisons indicate that the gains of EP are not explained by KL regularization alone and motivate examining how the prior is constructed, which we study further in Section 5.4.

Generalization across Models. EP generalizes across both model scale and family (Table F in Appendix D). Compared with GRPO, EP improves mean Avg@16 and Pass@16 by 4.5 and 3.6 points, respectively, on Qwen3-8B, and by 1.6 and 3.7 points on OLMo3-7B-Instruct. These results indicate that the gains are not specific to Qwen3-4B.

5.3 Analyses

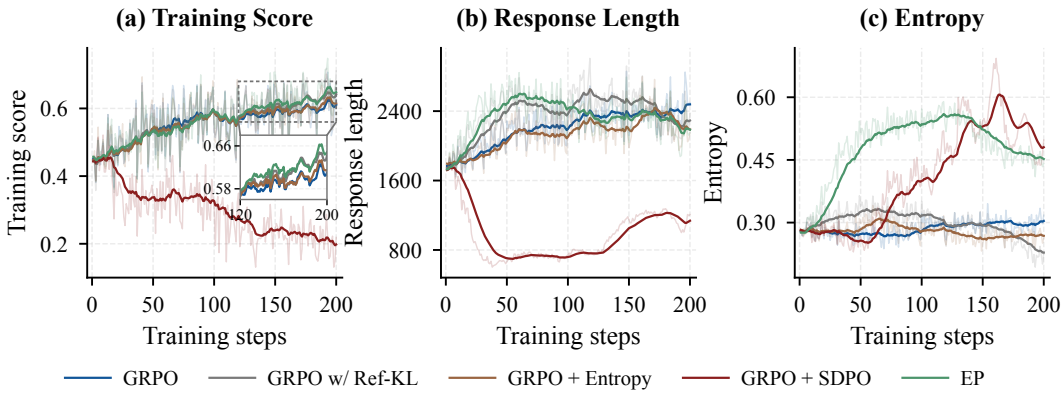

Figure 2: Training dynamics on Qwen3-4B. We report (a) training score, (b) average response length, and (c) policy entropy throughout training. Faint lines show the raw measurements, while solid lines show the smoothed trends. The inset in (a) provides a zoomed-in view of the later training stage for a clearer comparison among methods.

Training Dynamics. Figure 2 compares the training dynamics of EP and the baselines on Qwen3-4B. EP maintains higher policy entropy and longer responses during the early and middle stages of training, while both gradually decrease as the exploration-prior coefficient is annealed. Meanwhile, its training score continues to improve, with the advantage over GRPO becoming more pronounced during the middle and later stages of training. This joint evolution is consistent with the intended role of the annealed prior: encouraging broader exploration early in training and gradually shifting emphasis toward reward optimization.

GRPO+Entropy exhibits higher entropy early in training but achieves lower training scores and downstream performance than EP, suggesting that increased entropy alone is insufficient for effective exploration. GRPO w/ Ref-KL shows a different pattern: despite relatively high training scores, it yields little improvement in Avg@16 and lower Pass@16 than GRPO. One possible explanation is that regularization toward the fixed pre-RL reference restricts exploration. Consistent with this interpretation, its entropy eventually falls below that of GRPO, while Figure 3 shows lower late-stage textual diversity and consistently lower Pass@K, suggesting greater reliance on a narrower set of successful trajectories. GRPO+SDPO exhibits another failure mode: its entropy increases during the middle and later stages of training, while both the training score and response length deteriorate, a pattern also reported by Kim et al. (2026). One possible explanation is that a prior derived from self-generated successful trajectories may reinforce already-discovered reasoning modes rather than provide useful signals for exploring alternative ones.

Exploration Analysis. Figure 3 shows that EP improves both successful-solution coverage and rollout diversity. As shown in Figure 3(a), EP achieves the highest Pass@K across the entire sampling range, with particularly clear gains over GRPO at moderate-to-large sampling budgets. This indicates that increasing the inference-time sampling budget under EP more effectively uncovers successful reasoning trajectories.

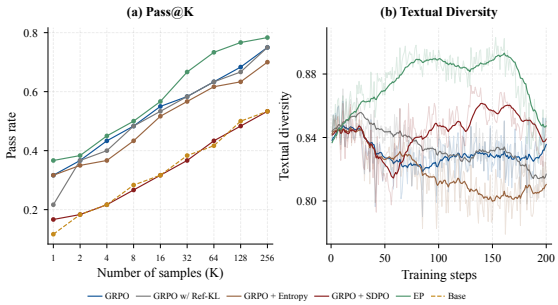

Figure 3: Exploration analysis on Qwen3-4B. (a) Pass@K comparison across all methods on a combined evaluation set consisting of AIME 2025 and HMMT25. (b) BLEU-based textual diversity (TD) (Hu et al., 2026) among rollouts within each group generated from the same training prompt.

Following prior work (Hu et al., 2026), we further measure textual diversity using the average pairwise BLEU-based dissimilarity among rollouts generated for the same prompt. Figure 3(b) shows that EP maintains substantially higher textual diversity than most baselines throughout the majority of training, indicating broader coverage of distinct reasoning trajectories. The decline in textual diversity toward the end of training is also qualitatively consistent with the corresponding reduction in policy entropy in Figure 2(c).

Notably, GRPO+SDPO also exhibits relatively high textual diversity during the middle and later stages of training, yet achieves substantially lower Pass@K; meanwhile, its training score deteriorates markedly in Figure 2(a). This contrast suggests that diversity alone is insufficient for effective exploration: the explored trajectories must also provide useful coverage of successful solutions. Taken together, these results show that EP increases rollout diversity while simultaneously improving successful-solution coverage.

5.4 Ablation Studies

We ablate four key design choices: the semantics of the path-switching intervention, entropy-based selection of the intervention position, the annealing schedule for the exploration prior, and the clipping of the prior signal. Table 2 summarizes the results.

Path-Switching Semantics. We replace the path-switching instruction with an unrelated random intervention (e.g., "The room stays quiet."), while keeping all other components of EP unchanged. This variant achieves a mean Avg@16 of 44.8 and a mean Pass@16 of 65.6, underperforming both GRPO and the full EP method. These results suggest that generic contextual perturbations alone are insufficient to reproduce the gains of EP, highlighting the importance of reasoning-relevant intervention semantics.

Intervention-Position Selection. We replace entropy-guided position selection with uniform sampling over the same candidate region while retaining the path-switching intervention. This variant achieves a mean Avg@16 of 50.7 and a mean Pass@16 of 68.4, preserving most of the Avg@16 improvement over GRPO but underperforming the full EP method, particularly on Pass@16. These

Table 2: Ablation study of EP. GRPO serves as the baseline and EP denotes the complete method. We ablate the path-switching semantics, entropy-based intervention-position selection, the annealing schedule for the exploration prior and the clipping of prior signal. We report Avg@16 and Pass@16 on each benchmark. Best results in each column are shown in **bold**.

Method / Variant	AIME24		AIME25		AMC23		Minerva		HMMT25		AVG	
	Avg@16	Pass@16	Avg@16	Pass@16	Avg@16	Pass@16	Avg@16	Pass@16	Avg@16	Pass@16	Avg@16	Pass@16
GRPO	45.2	**76.7**	36.7	66.7	86.6	**97.5**	42.9	54.8	21.7	43.3	46.6	67.8
Random Intervention	42.5	70.0	35.8	60.0	84.1	95.0	43.4	56.3	18.3	46.7	44.8	65.6
Uniform Placement	51.3	70.0	44.0	70.0	88.1	97.5	44.8	54.4	**25.6**	50.0	50.7	68.4
Fixed β_τ	33.5	56.7	29.6	56.7	81.3	95.0	41.9	54.0	14.2	30.0	40.1	58.5
w/o Prior-Signal Clipping	47.9	73.3	39.0	66.7	89.1	95.0	**46.0**	**58.5**	19.8	40.0	48.3	66.7
EP (Ours)	**54.4**	**76.7**	**45.4**	**73.3**	**89.8**	**97.5**	44.4	56.6	**25.6**	**53.3**	**51.9**	**71.5**

results suggest that entropy-guided placement provides an additional benefit by applying interventions to more uncertain reasoning states.

Exploration-Prior Schedule. We replace the annealed coefficient with a fixed $\beta_\tau = 0.1$, equal to the initial value used by EP. This variant achieves a mean Avg@16 of 40.1 and a mean Pass@16 of 58.5, substantially underperforming GRPO on both metrics. These results suggest that retaining a fixed exploration-prior strength throughout training can interfere with later-stage reward optimization, supporting the annealing of β_τ to encourage exploration early in training while shifting emphasis toward reward optimization later.

Prior-Signal Clipping. We remove clipping of the token-level prior-induced likelihood shift while keeping the remaining training procedure unchanged. Without clipping, the mean Avg@16 decreases from 51.9 to 48.3, while the mean Pass@16 drops from 71.5 to 66.7. These results suggest that clipping prevents excessively large prior-induced shifts from dominating policy updates, thereby improving optimization stability.

Overall, these ablations show complementary roles for the four components: semantic path switching provides the main exploration signal, while entropy-guided placement, coefficient annealing, and prior-signal clipping improve where, when, and how strongly this signal influences policy optimization. Additional sensitivity analyses of the prior-signal clipping threshold and the strength of β_0 are provided in Appendix G.

6 Conclusion

We introduced Exploration Prior (EP), a semantically guided exploration framework that constructs a context-dependent policy prior from answer-free path-switching interventions. By conditioning a frozen policy snapshot on lightweight reasoning guidance, EP shifts the continuation distribution toward alternative reasoning directions that may otherwise receive little probability under the current policy. In this way, the model's natural-language conditioning capability provides the exploration signal, while task rewards favor high-reward continuations under the resulting prior. Annealing the prior coefficient then gradually shifts optimization from guided exploration toward reward maximization. Extensive experiments across three models and five mathematical reasoning benchmarks show that EP consistently outperforms standard RLVR and representative exploration- and prior-based baselines. Further analyses show that EP increases rollout diversity while improving the coverage of successful solutions. These results suggest that semantic reasoning guidance offers a simple and effective way to improve exploration in RLVR without additional solution supervision.

AI Use Statement

In this work, we used generative AI tools to assist with writing and language polishing, literature discovery for related work, and code generation. We did not use generative AI tools to generate synthetic datasets, develop theoretical models or conceptual frameworks, formulate or prove math-

ematical claims, propose or refine hypotheses, design research methodology or experiments, clean or reformat datasets, conduct qualitative or thematic data analysis, or interpret experimental results. Translation assistance was also not used. We reviewed all AI-assisted work. AI-assisted writing was manually checked and revised by the authors before inclusion in the manuscript. For literature discovery, we independently read and verified the suggested papers, removed irrelevant references, and supplemented the related-work coverage with relevant literature not identified by the AI tools. AI-generated code was manually reviewed and checked for correctness before use. We take full responsibility for the final content of this work, including all text, claims, code, and other artifacts produced with the aid of generative AI.

REFERENCES

Zhipeng Chen, Xiaobo Qin, Youbin Wu, Yue Ling, Qinghao Ye, Wayne Xin Zhao, and Guang Shi. Pass@k training for adaptively balancing exploration and exploitation of large reasoning models. *ArXiv*, abs/2508.10751, 2025. URL https://api.semanticscholar.org/CorpusID:280649795.

Ganqu Cui, Yuchen Zhang, Jiacheng Chen, Lifan Yuan, Zhi Wang, Yuxin Zuo, Haozhan Li, Yuchen Fan, Huayu Chen, Weize Chen, Zhiyuan Liu, Hao Peng, Lei Bai, Wanli Ouyang, Yu Cheng, Bowen Zhou, and Ning Ding. The entropy mechanism of reinforcement learning for reasoning language models. *ArXiv*, abs/2505.22617, 2025. URL https://api.semanticscholar.org/CorpusID:278959427.

David Dohan, Winnie Xu, Aitor Lewkowycz, Jacob Austin, David Bieber, Raphael Gontijo Lopes, Yuhuai Wu, Henryk Michalewski, Rif A. Saurous, Jascha Sohl-Dickstein, Kevin Murphy, and Charles Sutton. Language model cascades, 2022. URL https://arxiv.org/abs/2207.10342.

Yuqian Fu, Haohuan Huang, Kaiwen Jiang, Jiacai Liu, Zhuo Jiang, Yuanheng Zhu, and Dongbin Zhao. Revisiting on-policy distillation: Empirical failure modes and simple fixes, 2026. URL https://arxiv.org/abs/2603.25562.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Hanwei Xu, Honghui Ding, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jingchang Chen, Jingyang Yuan, Jinhao Tu, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaichao You, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingxu Zhou, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645

(8081):633–638, September 2025. ISSN 1476-4687. doi: 10.1038/s41586-025-09422-z. URL http://dx.doi.org/10.1038/s41586-025-09422-z.

Jingcheng Hu, Yinmin Zhang, Qi Han, Daxin Jiang, Xiangyu Zhang, and Heung-Yeung Shum. Open-reasoner-zero: An open source approach to scaling up reinforcement learning on the base model, 2025. URL https://arxiv.org/abs/2503.24290.

Zican Hu, Shilin Zhang, Yafu Li, Jianhao Yan, Xuyang Hu, Leyang Cui, Xiaoye Qu, Chunlin Chen, Yu Cheng, and Zhi Wang. Diversity-incentivized exploration for versatile reasoning. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=9G7AbBrd27.

Jonas Hübotter, Frederike Lübeck, Lejs Behric, Anton Baumann, Marco Bagatella, Daniel Marta, Ido Hakimi, Idan Shenfeld, Thomas Kleine Buening, Carlos Guestrin, and Andreas Krause. Reinforcement learning via self-distillation, 2026. URL https://arxiv.org/abs/2601.20802.

Jeonghye Kim, Xufang Luo, Minbeom Kim, Sangmook Lee, Dohyung Kim, Jiwon Jeon, Dongsheng Li, and Yuqing Yang. Why does self-distillation (sometimes) degrade the reasoning capability of llms?, 2026. URL https://arxiv.org/abs/2603.24472.

Tomasz Korbak, Ethan Perez, and Christopher Buckley. RL with KL penalties is better viewed as Bayesian inference. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 1083–1091. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.findings-emnlp.77. URL https://aclanthology.org/2022.findings-emnlp.77/.

Chanuk Lee, Minki Kang, and Sung Ju Hwang. Sage: Shaping anchors for guided exploration in rlvr of llms. *ArXiv*, abs/2605.18864, 2026. URL https://api.semanticscholar.org/CorpusID:288655908.

Sergey Levine. Reinforcement learning and control as probabilistic inference: Tutorial and review. *ArXiv*, abs/1805.00909, 2018. URL https://api.semanticscholar.org/CorpusID:19077536.

Kevin Lu and Thinking Machines Lab. On-policy distillation. *Thinking Machines Lab: Connectionism*, 2025. doi: 10.64434/tml.20251026. https://thinkingmachines.ai/blog/on-policy-distillation.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088.

Amrith Setlur, Matthew Y. R. Yang, Charlie Victor Snell, Jeremiah Greer, Ian Wu, Virginia Smith, Max Simchowitz, and Aviral Kumar. e3: Learning to explore enables extrapolation of test-time compute for LLMs. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=aID0dZmMmM.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. URL https://arxiv.org/abs/2402.03300.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient rlhf framework. In *Proceedings of the Twentieth European Conference on Computer Systems*, pp. 1279–1297, 2025.

Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan J. Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. Learning to summarize from human feedback. *ArXiv*, abs/2009.01325, 2020. URL https://api.semanticscholar.org/CorpusID:221665105.

Richard S. Sutton and Andrew G. Barto. *Reinforcement learning: an introduction*. Adaptive computation and machine learning series. The MIT Press, Cambridge, Massachusetts, second edition edition, 2018. ISBN 978-0-262-03924-6.

Core Team, Bangjun Xiao, Bingquan Xia, Bo Yang, Bofei Gao, Bowen Shen, Chen Zhang, Chenhong He, Chiheng Lou, Fuli Luo, Gang Wang, Gang Xie, Hailin Zhang, Hanglong Lv, Hanyu Li, Heyu Chen, Hongshen Xu, Houbin Zhang, Huaqiu Liu, Jiangshan Duo, Jianyu Wei, Jiebao Xiao, Jinhao Dong, Jun Shi, Junhao Hu, Kainan Bao, Kang Zhou, Lei Li, Liang Zhao, Linghao Zhang, Peidian Li, Qianli Chen, Shaohui Liu, Shihua Yu, Shijie Cao, Shimao Chen, Shouqiu Yu, Shuo Liu, Tianling Zhou, Weijiang Su, Weikun Wang, Wenhan Ma, Xiangwei Deng, Bohan Mao, Bowen Ye, Can Cai, Chenghua Wang, Chengxuan Zhu, Chong Ma, Chun Chen, Chunan Li, Dawei Zhu, Deshan Xiao, Dong Zhang, Duo Zhang, Fangyue Liu, Feiyu Yang, Fengyuan Shi, Guoan Wang, Hao Tian, Hao Wu, Heng Qu, Hongfei Yi, Hongxu An, Hongyi Guan, Xing Zhang, Yifan Song, Yihan Yan, Yihao Zhao, Yingchun Lai, Yizhao Gao, Yu Cheng, Yuanyuan Tian, Yudong Wang, Zhen Tang, Zhengju Tang, Zhengtao Wen, Zhichao Song, Zhixian Zheng, Zihan Jiang, Jian Wen, Jiarui Sun, Jiawei Li, Jinlong Xue, Jun Xia, Kai Fang, Menghang Zhu, Nuo Chen, Qian Tu, Qihao Zhang, Qiying Wang, Rang Li, Rui Ma, Shaolei Zhang, Shengfan Wang, Shicheng Li, Shuhao Gu, Shuhuai Ren, Sirui Deng, Tao Guo, Tianyang Lu, Weiji Zhuang, Weikang Zhang, Weimin Xiong, Wenshan Huang, Wenyu Yang, Xin Zhang, Xing Yong, Xu Wang, Xueyang Xie, Yilin Jiang, Yixin Yang, Yongzhe He, Yu Tu, Yuanliang Dong, Yuchen Liu, Yue Ma, Yue Yu, Yuxing Xiang, Zhaojun Huang, Zhenru Lin, Zhipeng Xu, Zhiyang Chen, Zhonghua Deng, Zihan Zhang, and Zihao Yue. Mimo-v2-flash technical report, 2026. URL https://arxiv.org/abs/2601.02780.

Shenzhi Wang, Le Yu, Chang Gao, Chujie Zheng, Shixuan Liu, Rui Lu, Kai Dang, Xiong hui Chen, Jianxin Yang, Zhenru Zhang, Yuqiong Liu, An Yang, Andrew Zhao, Yang Yue, Shiji Song, Bowen Yu, Gao Huang, and Junyang Lin. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for llm reasoning. *ArXiv*, abs/2506.01939, 2025. URL https://api.semanticscholar.org/CorpusID:279119146.

Ronald J. Williams and Jing Peng. Function optimization using connectionist reinforcement learning algorithms. *Connection Science*, 3:241–268, 1991. URL https://api.semanticscholar.org/CorpusID:18972907.

Fang Wu, Weihao Xuan, Ximing Lu, Mingjie Liu, Yi Dong, Zaid Harchaoui, and Yejin Choi. The invisible leash: Why rlvr may or may not escape its origin, 2026. URL https://arxiv.org/abs/2507.14843.

Chenxu Yang, Chuanyu Qin, Qingyi Si, Minghui Chen, Naibin Gu, Dingyu Yao, Zheng Lin, Weiping Wang, Jiaqi Wang, and Nan Duan. Self-distilled rlvr, 2026. URL https://arxiv.org/abs/2604.03128.

Tianzhu Ye, Li Dong, Xun Wu, Shaohan Huang, and Furu Wei. On-policy context distillation for language models, 2026. URL https://arxiv.org/abs/2602.12275.

Qiying Yu, Zheng Zhang, et al. DAPO: An Open-Source LLM Reinforcement Learning System at Scale, 2025.

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. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=Jpxfof0EaS.

Daniel M. Ziegler, Nisan Stiennon, Jeff Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *ArXiv*, abs/1909.08593, 2019. URL https://api.semanticscholar.org/CorpusID:202660943.

A INTERVENTIONS

Path Switching Interventions

Alternatively, perhaps we can try another approach.
Alternatively, perhaps I can find the answer in another way.
Wait, let's try a different approach.
Let's try another approach.
Wait, let's switch to a better approach.
Let's switch viewpoint and try a different strategy.
Let's double check the initial reasoning. Perhaps a different approach would work better.
Wait, let me cross check with another approach.
Maybe there's another way to solve this.
Perhaps there's a better approach.
Maybe there's a mistake. Alternative approach:
Maybe I made a mistake. Let me try a different approach.

Random Interventions

The window remains quietly open.
A shadow appears nearby.
The room stays quiet.
The surface looks smooth.
A small object rests near the window.
The light falls softly across the surface.
Several ordinary objects remain in the background.
A faint sound comes from a distance.
The nearby area appears mostly unchanged.
A small detail is visible near the edge.
The background remains quiet and relatively plain.

B ON-POLICY DISTILLATION

On-policy distillation (OPD) (Lu & Lab, 2025) mitigates the distribution mismatch of conventional distillation by evaluating the teacher distribution on trajectories sampled from the student policy. Given a student policy π_θ and a teacher distribution π_{teacher}, OPD minimizes the reverse KL divergence from the student distribution to the teacher distribution over prefixes visited by the student:

$$\mathcal{L}_{\text{OPD}} = \mathbb{E}_{x\sim\mathcal{D}, y\sim\pi_\theta(\cdot|x)}\left[\sum_t D_{\text{KL}}\left(\pi_\theta(\cdot \mid x, y_{<t}) \| \,\text{sg}[\pi_{\text{teacher}}(\cdot \mid x, y_{<t})]\right)\right], \tag{12}$$

where $\text{sg}[\cdot]$ denotes the stop-gradient operation. The teacher distribution can be provided by a separate teacher model (Lu & Lab, 2025) or induced by the same underlying model under a different conditioning context (Zhao et al., 2026; Hübotter et al., 2026; Ye et al., 2026).

C PRACTICAL CLIPPED SURROGATE OBJECTIVE.

We instantiate the reward-maximization term using a GRPO-style clipped surrogate, with the group-relative advantage estimator in equation 1 and the token-level importance ratio in equation 2. Combining the resulting reward gradient with the gradient induced by the exploration-KL term yields the following token-level effective advantage:

$$A^{\text{EP}}_{i,t} = A_i + \beta_\tau \mathbf{1}_{\{t\geq k_i\}}\ \text{sg}\left[\bar{\delta}^e_{i,t}\right]. \tag{13}$$

The positive sign follows from the definition $\delta^e_{i,t} = \log\pi^e_{\bar{\theta}} - \log\pi_\theta$: tokens assigned greater relative likelihood by the exploration prior receive a positive adjustment, whereas tokens downweighted by the prior receive a negative adjustment.

Given a group of trajectories $\{y_i\}_{i=1}^{G}$ sampled from the rollout policy $\pi_{\theta_{\text{old}}}$, we optimize the following REINFORCE-style clipped surrogate:

$$\widehat{\mathcal{J}}_{\text{EP}}(\theta) = \mathbb{E}_{x\sim\mathcal{D},\,\{y_i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}}\left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|y_i|}\sum_{t=1}^{|y_i|}\min\Big(r_{i,t}(\theta)A_{i,t}^{\text{EP}},\ \text{clip}\left(r_{i,t}(\theta),1-\epsilon,1+\epsilon\right)A_{i,t}^{\text{EP}}\Big)\right], \tag{14}$$

where ϵ denotes importance ratio clipping threshold. For tokens preceding the intervention position, $t < k_i$, the exploration term vanishes.

D Exploration Algorithm

Algorithm 1 summarizes the optimization procedure for the Exploration Prior (EP) method.

Algorithm 1 Training with Exploration Prior

Require: Training distribution $\mathcal{D}$, policy π_θ, reward function R, intervention set $\mathcal{E}$
Require: Group size G, number of high-entropy candidates N, prior schedule $\{\beta_\tau\}$
Require: Prior-signal clipping threshold c, policy-ratio clipping threshold ϵ

for policy update $\tau = 1, 2, \ldots$ **do**
  Set rollout policy $\pi_{\theta_{\text{old}}} \leftarrow \pi_\theta$ and frozen prior policy $\pi_{\bar{\theta}} \leftarrow \pi_\theta$
  Sample a batch of problems from $\mathcal{D}$
  for each problem x in the batch **do**
    Sample G trajectories $\{y_i\}_{i=1}^{G} \sim \pi_{\theta_{\text{old}}}(\cdot \mid x)$
    Compute $R_i \leftarrow R(x, y_i)$ for $i = 1, \ldots, G$
    Compute group-relative advantages $\{A_i\}_{i=1}^{G}$ using equation 1
    for each trajectory y_i **do**
      Define candidate positions $\mathcal{C}_i = \{t : \lceil 0.2|y_i| \rceil \le t \le \lfloor 0.6|y_i| \rfloor\}$
      Compute token-level predictive entropy under $\pi_{\bar{\theta}}$ for every $t \in \mathcal{C}_i$
      Select the top-N highest-entropy positions $\mathcal{K}_i \subseteq \mathcal{C}_i$ and sample $k_i \sim \text{Uniform}(\mathcal{K}_i)$
      Sample an intervention $e_i \sim \text{Uniform}(\mathcal{E})$
      Construct the exploration prior

$$\pi_{\bar{\theta}}^{e_i}(\cdot \mid x, y_{i,<k_i}) \leftarrow \pi_{\bar{\theta}}(\cdot \mid x, y_{i,<k_i}, e_i)$$

      for $t = k_i, \ldots, |y_i|$ **do**
        Compute prior-induced likelihood shift

$$\delta_{i,t}^{e_i} \leftarrow \log \frac{\pi_{\bar{\theta}}^{e_i}(y_{i,t} \mid x, y_{i,<t})}{\pi_\theta(y_{i,t} \mid x, y_{i,<t})}$$

        $\bar{\delta}_{i,t}^{e_i} \leftarrow \text{clip}(\delta_{i,t}^{e_i}, -c, c)$
        $A_{i,t}^{\text{EP}} \leftarrow A_i + \beta_\tau \,\text{sg}[\bar{\delta}_{i,t}^{e_i}]$
      end for
      $A_{i,t}^{\text{EP}} \leftarrow A_i$ for $t < k_i$
    end for
  end for
  Update θ using the clipped surrogate in equation 14 with clipping threshold ϵ
end for

E Hyperparameters

All models are trained using the VERL framework. During training, we use a maximum rollout length of 4K tokens for Qwen3-4B and 8K tokens for Qwen3-8B and OLMo3-7B-Instruct. For GRPO w/ Ref-KL, the KL coefficient is set to 0.001. For EP, the initial exploration-prior coefficient is set to $\beta_0 = 0.1$; it remains fixed for the first 60 steps and is then annealed to 0 over the next

10 steps. All other training and evaluation configurations are kept identical across methods. For the Qwen3 models, we disable the thinking mode during both training and evaluation. Detailed hyperparameters are provided in Table 3.

Table 3: Training and evaluation hyperparameters.

Hyperparameter	**Value**
Training steps	200
Optimizer	AdamW
Learning rate	1×10^{-6}
Batch size	64
Mini Batch size	64
Number of rollouts	8
Max sequence length (evaluation)	32K
Training rollout sampling	temperature 1.0, top-$p = 1.0$
Evaluation rollout sampling	temperature 0.7, top-$p = 0.8$, top-$k = 20$
Framework	verl (Sheng et al., 2025)

F Generalization across Model Scales and Families

Table 4: Generalization across model scales and families. We report Avg@16 and Pass@16 on Qwen3-8B and OLMo3-7B-Instruct. In addition to GRPO, we include GRPO+Entropy because it achieves the highest mean Avg@16 among the non-EP methods in our primary Qwen3-4B experiments. Best results in each column within each backbone are shown in **bold**.

Backbone / Method	**AIME24**		**AIME25**		**AMC23**		**Minerva**		**HMMT25**		**AVG**	
	Avg@16	Pass@16	Avg@16	Pass@16	Avg@16	Pass@16	Avg@16	Pass@16	Avg@16	Pass@16	Avg@16	Pass@16
Qwen3-8B	39.8	70.0	19.2	40.0	86.7	97.5	43.2	52.9	19.2	40.0	41.6	60.1
GRPO	63.8	**83.3**	49.2	76.7	91.3	97.5	44.0	**55.9**	28.1	50.0	55.3	72.7
GRPO+Entropy	59.4	80.0	43.3	76.7	89.8	97.5	44.0	55.1	28.3	53.3	53.0	72.5
EP (Ours)	**70.6**	80.0	**58.8**	**80.0**	**93.6**	**100.0**	**45.0**	54.8	**31.0**	**66.7**	**59.8**	**76.3**
OLMo3-7B-Instruct	46.3	83.3	36.3	66.7	79.8	97.5	35.5	56.3	19.8	50.0	43.5	70.8
GRPO	54.4	76.7	41.8	70.0	**90.6**	**100.0**	**38.8**	57.7	25.6	**53.3**	50.3	71.5
GRPO+Entropy	56.5	83.3	41.5	73.3	89.5	97.5	**38.8**	57.0	27.1	**53.3**	50.7	72.9
EP (Ours)	**59.8**	**90.0**	**44.0**	**76.7**	**90.6**	**100.0**	37.9	**59.2**	**27.5**	50.0	**51.9**	**75.2**

G Additional Ablations and Sensitivity Analyses

G.1 Sensitivity to Prior-Signal Clipping

We study the sensitivity of EP to the clipping threshold c applied to the token-level prior-induced likelihood shift $\delta^e_{i,t}$. We compare the default $c = 0.3$ with tighter ($c = 0.1$) and looser ($c = 0.6$) clipping while keeping all other optimization settings fixed. The results are reported in Table 5.

Table 5: Sensitivity of EP to the prior-signal clipping threshold c. We report Avg@16 and Pass@16 on each benchmark. The default setting is shaded. Best results in each column are shown in **bold**.

Clipping Threshold	**AIME24**		**AIME25**		**AMC23**		**Minerva**		**HMMT25**		**AVG**	
	Avg@16	Pass@16	Avg@16	Pass@16	Avg@16	Pass@16	Avg@16	Pass@16	Avg@16	Pass@16	Avg@16	Pass@16
$c = 0.1$	44.6	**83.3**	40.4	70.0	85.5	**97.5**	43.5	54.4	21.7	46.7	47.1	70.4
$c = 0.3$ (default)	**54.4**	76.7	**45.4**	**73.3**	**89.8**	**97.5**	44.4	**56.6**	**25.6**	**53.3**	**51.9**	**71.5**
$c = 0.6$	47.7	73.3	41.7	66.7	87.5	**97.5**	**44.6**	55.9	18.3	33.3	50.0	65.3

Table 5 shows that EP performs robustly under moderate clipping thresholds. The default setting $c = 0.3$ achieves the highest mean Avg@16, while the tighter $c = 0.1$ yields a comparable mean

Pass@16. In contrast, loosening the threshold to $c = 0.6$ reduces overall performance, particularly Pass@16. These results suggest that clipping is useful for limiting excessively large token-level prior signals, while overly permissive clipping can make optimization less stable.

G.2 Sensitivity to the Exploration-Prior Strength

The coefficient β_τ controls the strength of the exploration prior relative to reward optimization at step τ. We vary its initial value β_0 over $\{0.05, 0.1, 0.2\}$ while keeping the same annealing schedule. We use $\beta_0 = 0.1$ as the default setting. The results are reported in Table 6. Table 6 shows that EP performs similarly with initial prior strengths $\beta_0 = 0.05$ and 0.1, with $\beta_0 = 0.05$ achieving slightly higher aggregate performance. However, increasing the coefficient to 0.2 substantially degrades both Avg@16 and Pass@16. This suggests that EP is relatively insensitive within a moderate range of prior strengths, whereas overly strong prior regularization can interfere with reward optimization even when the coefficient is annealed over training.

Table 6: Sensitivity of EP to the initial exploration-prior strength β_0. We report Avg@16 and Pass@16 on each benchmark. The default setting is shaded. Best results in each column are shown in **bold**.

Initial Prior Strength	AIME24		AIME25		AMC23		Minerva		HMMT25		AVG	
	Avg@16	Pass@16	Avg@16	Pass@16	Avg@16	Pass@16	Avg@16	Pass@16	Avg@16	Pass@16	Avg@16	Pass@16
$\beta_0 = 0.05$	**54.4**	**80.0**	**46.9**	**83.3**	89.2	95.0	**45.6**	55.5	**29.0**	50.0	**53.0**	**72.8**
$\beta_0 = 0.1$ (default)	**54.4**	76.7	45.4	73.3	**89.8**	**97.5**	44.4	**56.6**	25.6	**53.3**	51.9	71.5
$\beta_0 = 0.2$	36.5	66.7	33.5	60.0	82.8	**97.5**	42.9	55.1	15.2	36.7	42.2	63.2

H Proofs

Lemma H.1 (Entropy regularization as uniform-prior KL regularization)**.** *Fix x, $y_{<k}$, and $\beta_\tau > 0$ at optimization step τ, and let $\mathcal{Y}_{k:T}$ denote a fixed finite space of reasoning suffixes $y_{k:T}$. Define the uniform distribution over this space as*

$$U(y_{k:T} \mid x, y_{<k}) = \frac{1}{|\mathcal{Y}_{k:T}|}, \qquad y_{k:T} \in \mathcal{Y}_{k:T}. \tag{15}$$

Then the entropy-regularized objective

$$\begin{aligned} \mathcal{J}_{\mathrm{Ent}}(\theta) = \mathbb{E}_{y_{k:T} \sim \pi_\theta(\cdot \mid x, y_{<k})} \left[R(x, y) \right] \\ + \beta_\tau H(\pi_\theta(\cdot \mid x, y_{<k})) \end{aligned} \tag{16}$$

is equivalent, up to a constant independent of θ, to

$$\begin{aligned} \mathcal{J}_{\mathrm{Uni}}(\theta) = \mathbb{E}_{y_{k:T} \sim \pi_\theta(\cdot \mid x, y_{<k})} \left[R(x, y) \right] \\ - \beta_\tau D_{\mathrm{KL}}(\pi_\theta(\cdot \mid x, y_{<k}) \,\|\, U(\cdot \mid x, y_{<k})) \,. \end{aligned} \tag{17}$$

Specifically,

$$\mathcal{J}_{\mathrm{Uni}}(\theta) = \mathcal{J}_{\mathrm{Ent}}(\theta) - \beta_\tau \log |\mathcal{Y}_{k:T}|. \tag{18}$$

Hence,

$$\arg\max_\theta \mathcal{J}_{\mathrm{Ent}}(\theta) = \arg\max_\theta \mathcal{J}_{\mathrm{Uni}}(\theta). \tag{19}$$

Proof. For brevity, condition all distributions on the fixed prefix $(x, y_{<k})$. By definition,

$$\begin{aligned} & D_{\mathrm{KL}}(\pi_\theta(\cdot \mid x, y_{<k}) \,\|\, U(\cdot \mid x, y_{<k})) \\ & = \sum_{y_{k:T} \in \mathcal{Y}_{k:T}} \pi_\theta(y_{k:T} \mid x, y_{<k}) \log \frac{\pi_\theta(y_{k:T} \mid x, y_{<k})}{U(y_{k:T} \mid x, y_{<k})}. \end{aligned} \tag{20}$$

Since

$$U(y_{k:T} \mid x, y_{<k}) = \frac{1}{|\mathcal{Y}_{k:T}|}, \tag{21}$$

we obtain

$$D_{\mathrm{KL}}(\pi_\theta \| U) = \sum_{y_{k:T}} \pi_\theta(y_{k:T} \mid x, y_{<k}) \log \pi_\theta(y_{k:T} \mid x, y_{<k}) + \log |\mathcal{Y}_{k:T}| \sum_{y_{k:T}} \pi_\theta(y_{k:T} \mid x, y_{<k}). \tag{22}$$

Because $\pi_\theta(\cdot \mid x, y_{<k})$ is normalized,

$$\sum_{y_{k:T}} \pi_\theta(y_{k:T} \mid x, y_{<k}) = 1. \tag{23}$$

Using

$$H(\pi_\theta(\cdot \mid x, y_{<k})) = -\sum_{y_{k:T}} \pi_\theta(y_{k:T} \mid x, y_{<k}) \log \pi_\theta(y_{k:T} \mid x, y_{<k}), \tag{24}$$

it follows that

$$D_{\mathrm{KL}}(\pi_\theta \| U) = \log |\mathcal{Y}_{k:T}| - H(\pi_\theta(\cdot \mid x, y_{<k})) \,. \tag{25}$$

Substituting equation 25 into equation 17 gives

$$\begin{aligned} \mathcal{J}_{\mathrm{Uni}}(\theta) &= \mathbb{E}_{y_{k:T} \sim \pi_\theta(\cdot \mid x, y_{<k})} [R(x, y)] \\ &\quad + \beta_\tau H(\pi_\theta(\cdot \mid x, y_{<k})) - \beta_\tau \log |\mathcal{Y}_{k:T}| \\ &= \mathcal{J}_{\mathrm{Ent}}(\theta) - \beta_\tau \log |\mathcal{Y}_{k:T}|. \end{aligned} \tag{26}$$

Since $|\mathcal{Y}_{k:T}|$ is fixed and therefore independent of θ, the two objectives have identical gradients with respect to θ and the same maximizers. □

The lemma clarifies the prior interpretation of entropy regularization. Under a fixed finite suffix space, entropy regularization is equivalent to reverse-KL regularization toward a uniform prior over reasoning suffixes. This prior assigns equal probability to every suffix in $\mathcal{Y}_{k:T}$ and therefore imposes no structured preference over different reasoning regions. In contrast, equation 6 uses the intervention-conditioned distribution $\pi^e_{\bar{\theta}}(\cdot \mid x, y_{<k})$ as a structured prior. By leveraging the semantic response of the model to natural-language interventions, this prior reallocates probability mass toward alternative reasoning regions, thereby increasing their relative support before reward-based reweighting.

The equivalence above assumes a fixed finite suffix space $\mathcal{Y}_{k:T}$. For practical LLMs with variable-length generation, token-level entropy regularization instead corresponds to a token-wise uniform autoregressive prior, which is generally not uniform over complete suffixes. In particular, a suffix of length L receives prior probability proportional to $|\mathcal{V}|^{-L}$, so suffixes of different lengths are assigned different probabilities. Although the fixed-suffix formulation abstracts away this distinction, we adopt it for simplicity because it is sufficient to expose the prior interpretation relevant to our comparison: entropy regularization induces an unstructured uniform prior, whereas EP leverages the natural-language understanding of the model to construct a structured prior from answer-free interventions that increases support for alternative reasoning regions.

Lemma H.2 (Exploration-prior regularization as posterior inference)**.** *Fix a prompt x, a prefix $y_{<k}$, an intervention e, and $\beta_\tau > 0$ at optimization step τ. Let $y = (y_{<k}, y_{k:T})$ denote the complete reasoning trajectory. We treat y as a latent variable in the probabilistic model; conditioned on x and the observed prefix $y_{<k}$, its latent component is the continuation $y_{k:T}$. Assume that $R(x, y)$ is bounded and that the normalized prior $\pi^e_{\bar{\theta}}(\cdot \mid x, y_{<k})$ is held fixed during optimization.*

Introduce a binary optimality variable O, where $O = 1$ represents task-reward-based optimality evidence, with likelihood

$$p(O = 1 \mid x, y) \propto \exp\left(\frac{R(x, y)}{\beta_\tau}\right).$$

The proportionality factor is independent of y and θ and is chosen so that the likelihood is a valid probability. Then the posterior over latent continuations is

$$\begin{aligned} \pi^*_{\mathrm{EP}}(y_{k:T} \mid x, y_{<k}) &:= p(y_{k:T} \mid x, y_{<k}, O = 1) \\ &= \frac{\pi^e_{\bar{\theta}}(y_{k:T} \mid x, y_{<k}) \exp(R(x, y)/\beta_\tau)}{Z^e_\tau(x, y_{<k})}, \end{aligned} \tag{27}$$

where the fixed intervention e is implicit in the posterior notation and

$$Z_\tau^e(x, y_{<k}) = \sum_{y_{k:T}} \pi_{\bar{\theta}}^e(y_{k:T} \mid x, y_{<k}) \exp\left(\frac{R(x,y)}{\beta_\tau}\right).$$

For policies with finite reverse KL to the prior, the EP objective satisfies

$$\begin{aligned}\mathcal{J}_{\mathrm{EP}}(\theta) = \beta_\tau \log Z_\tau^e(x, y_{<k}) \\ - \beta_\tau D_{\mathrm{KL}}(\pi_\theta(\cdot \mid x, y_{<k}) \,\|\, \pi_{\mathrm{EP}}^*(\cdot \mid x, y_{<k}))\,.\end{aligned} \tag{28}$$

Consequently, maximizing $\mathcal{J}_{\mathrm{EP}}(\theta)$ is equivalent to minimizing the reverse KL divergence to this posterior.

Proof. Taking $\pi_{\bar{\theta}}^e(\cdot \mid x, y_{<k})$ as the prior, Bayes' rule gives

$$\begin{aligned}&p(y_{k:T} \mid x, y_{<k}, O = 1) \\ &\quad \propto \pi_{\bar{\theta}}^e(y_{k:T} \mid x, y_{<k}) p(O = 1 \mid x, y) \\ &\quad \propto \pi_{\bar{\theta}}^e(y_{k:T} \mid x, y_{<k}) \exp\left(\frac{R(x,y)}{\beta_\tau}\right).\end{aligned}$$

Normalizing yields equation 27. The bounded reward and normalized prior ensure that $0 < Z_\tau^e(x, y_{<k}) < \infty$.

For the remainder of the proof, all expectations are over $y_{k:T} \sim \pi_\theta(\cdot \mid x, y_{<k})$. Substituting the posterior into the reverse KL divergence gives

$$\begin{aligned}&D_{\mathrm{KL}}(\pi_\theta(\cdot \mid x, y_{<k}) \,\|\, \pi_{\mathrm{EP}}^*(\cdot \mid x, y_{<k})) \\ &= \mathbb{E}_{\pi_\theta}\left[\log \frac{\pi_\theta(y_{k:T} \mid x, y_{<k})}{\pi_{\mathrm{EP}}^*(y_{k:T} \mid x, y_{<k})}\right] \\ &= \mathbb{E}_{\pi_\theta}\left[\log \frac{\pi_\theta(y_{k:T} \mid x, y_{<k})}{\pi_{\bar{\theta}}^e(y_{k:T} \mid x, y_{<k})} - \frac{R(x,y)}{\beta_\tau}\right] + \log Z_\tau^e(x, y_{<k}) \\ &= D_{\mathrm{KL}}\big(\pi_\theta(\cdot \mid x, y_{<k}) \,\|\, \pi_{\bar{\theta}}^e(\cdot \mid x, y_{<k})\big) \\ &\qquad - \frac{1}{\beta_\tau}\mathbb{E}_{\pi_\theta}[R(x,y)] + \log Z_\tau^e(x, y_{<k}).\end{aligned}$$

Rearranging and using the definition of the EP objective yields

$$\begin{aligned}\mathcal{J}_{\mathrm{EP}}(\theta) &= \mathbb{E}_{\pi_\theta}[R(x,y)] \\ &\qquad - \beta_\tau D_{\mathrm{KL}}\big(\pi_\theta(\cdot \mid x, y_{<k}) \,\|\, \pi_{\bar{\theta}}^e(\cdot \mid x, y_{<k})\big) \\ &= \beta_\tau \log Z_\tau^e(x, y_{<k}) \\ &\qquad - \beta_\tau D_{\mathrm{KL}}(\pi_\theta(\cdot \mid x, y_{<k}) \,\|\, \pi_{\mathrm{EP}}^*(\cdot \mid x, y_{<k}))\,.\end{aligned}$$

Since the prior and β_τ are fixed within optimization step τ, $\beta_\tau \log Z_\tau^e(x, y_{<k})$ is independent of θ. Therefore,

$$\arg\max_\theta \mathcal{J}_{\mathrm{EP}}(\theta) = \arg\min_\theta D_{\mathrm{KL}}(\pi_\theta(\cdot \mid x, y_{<k}) \,\|\, \pi_{\mathrm{EP}}^*(\cdot \mid x, y_{<k}))\,.$$

□

Lemma H.3 (Immediate-term approximation to the exploration-prior KL gradient)**.** *Assume that the exploration prior $\pi_{\bar{\theta}}^e$ is fixed with respect to θ. For a fixed trajectory index i, and let $\delta_{i,t}^e(\theta)$ be defined as in equation 8. Then*

$$\begin{aligned}&\nabla_\theta D_{\mathrm{KL}}\left(\pi_\theta(\cdot \mid x, y_{i,<k_i}) \,\|\, \pi_{\bar{\theta}}^e(\cdot \mid x, y_{i,<k_i})\right) \\ &\quad = -\mathbb{E}_{y_{i,k_i:T} \sim \pi_\theta(\cdot \mid x, y_{i,<k_i})}\left[\sum_{t=k_i}^{T}\left(\sum_{s=t}^{T} \delta_{i,s}^e(\theta)\right) \nabla_\theta \log \pi_\theta(y_{i,t} \mid x, y_{i,<t})\right].\end{aligned} \tag{29}$$

Retaining only the immediate likelihood-shift term at each position yields the approximation:

$$\nabla_\theta D_{\mathrm{KL}}\left(\pi_\theta(\cdot \mid x, y_{i,<k_i}) \,\|\, \pi_{\bar{\theta}}^e(\cdot \mid x, y_{i,<k_i})\right) \approx -\mathbb{E}_{y_{i,k_i:T}\sim\pi_\theta(\cdot|x,y_{i,<k_i})}\left[\sum_{t=k_i}^{T} \delta_{i,t}^e(\theta)\nabla_\theta \log \pi_\theta(y_{i,t} \mid x, y_{i,<t})\right]. \tag{30}$$

For rollouts sampled from $\pi_{\theta_{\mathrm{old}}}$, we use the corresponding token-local GRPO surrogate

$$\nabla_\theta D_{\mathrm{KL}}\left(\pi_\theta(\cdot \mid x, y_{i,<k_i}) \,\|\, \pi_{\bar{\theta}}^e(\cdot \mid x, y_{i,<k_i})\right) \approx -\mathbb{E}_{y_{i,k_i:T}\sim\pi_{\theta_{\mathrm{old}}}(\cdot|x,y_{i,<k_i})}\left[\sum_{t=k_i}^{T} r_{i,t}(\theta)\ \mathrm{sg}\big[\delta_{i,t}^e(\theta)\big]\ \nabla_\theta \log \pi_\theta(y_{i,t} \mid x, y_{i,<t})\right], \tag{31}$$

where $r_{i,t}(\theta)$ is the importance ratio defined in equation 2.

Proof. By autoregressive factorization and the definition of $\delta_{i,t}^e(\theta)$,

$$D_{\mathrm{KL}}\left(\pi_\theta \,\|\, \pi_{\bar{\theta}}^e\right) = -\mathbb{E}_{y_{i,k_i:T}\sim\pi_\theta}\left[\sum_{s=k_i}^{T} \delta_{i,s}^e(\theta)\right],$$

where all distributions are conditioned on the fixed prefix $(x, y_{i,<k_i})$.

Since $\pi_{\bar{\theta}}^e$ is fixed with respect to θ, applying the log-derivative identity gives

$$\nabla_\theta D_{\mathrm{KL}} = -\mathbb{E}_{\pi_\theta}\left[\sum_{t=k_i}^{T}\sum_{s=k_i}^{T} \delta_{i,s}^e(\theta)\nabla_\theta \log \pi_\theta(y_{i,t} \mid x, y_{i,<t})\right],$$

where the additional score term vanishes because $\mathbb{E}_{\pi_\theta}[\nabla_\theta \log \pi_\theta(y_{i,k_i:T} \mid x, y_{i,<k_i})] = 0$.

For $s < t$, $\delta_{i,s}^e(\theta)$ depends only on tokens preceding $y_{i,t}$, while

$$\mathbb{E}\left[\nabla_\theta \log \pi_\theta(y_{i,t} \mid x, y_{i,<t}) \mid x, y_{i,<t}\right] = 0.$$

Hence all terms with $s < t$ vanish in expectation, leaving

$$\nabla_\theta D_{\mathrm{KL}} = -\mathbb{E}_{\pi_\theta}\left[\sum_{t=k_i}^{T}\left(\sum_{s=t}^{T} \delta_{i,s}^e(\theta)\right)\nabla_\theta \log \pi_\theta(y_{i,t} \mid x, y_{i,<t})\right],$$

which proves equation 29.

equation 30 follows from Fu et al. (2026) by retaining only the $s = t$ term. When trajectories are generated by $\pi_{\theta_{\mathrm{old}}}$, we evaluate this immediate contribution using the token-local importance ratio $r_{i,t}(\theta)$, yielding equation 31. □