JUST REUSE THE ROLLOUT LOG-PROBS: STABLE MIXTURE-OF-EXPERTS REINFORCEMENT LEARNING WITH A VOCABULARY-KL CLIP

Anonymous authors
Paper under double-blind review

ABSTRACT

In reinforcement learning of large language models, responses are typically generated by a fast inference engine while gradients are computed by a different training backend, and the two disagree on the same tokens because of numerical and, for mixture-of-experts models, expert-routing differences. The PPO/GRPO ratio clip that stabilizes these updates acts on the sampled token's probability ratio: a one-dimensional proxy that neither measures how far the full next-token distribution moved nor, under trainer recompute, reflects which policy generated the data. We find that a double clip stabilizes these updates, and that with it the cheapest correction suffices. VoKL-Clip clips each token by a top-k bucket KL between the rollout and current policies (computed for free from the top-k log-probs the inference engine returns on request), removing drifted tokens from the gradient by stop-gradient, and keeps a wide ratio clip alongside it. The bucket coarsening yields the sharpest full-vocabulary total-variation bound available from top-k information, yet the vocabulary-KL clip provably cannot control importance-weight moments, which the ratio clip bounds; dropping either clip costs 7–11 AIME24 points. With both, VoKL-Clip can just reuse the rollout's own log-probs as the behavior policy: the importance ratio is faithful to the sampler, with no recompute and no routing replay. On Qwen3-30B-A3B trained on DAPO-Math-17k, VoKL-Clip reaches 64.5% AIME24 (ave@32)—versus 58.1% for DAPO and 57.2% for GSPO, each at its overall best rollout↔training mismatch-correction level—leads across the reasoning-benchmark suite, and on AIME24 exceeds every baseline at every correction level at lower correction cost, with stable training dynamics.

1 INTRODUCTION

Reinforcement learning is now a standard way to train large language models, and at scale responses are generated by a fast inference engine (e.g. vLLM, Kwon et al., 2023) while gradients are computed by a different training backend. The two evaluate the *same* tokens under subtly different numerics and, for mixture-of-experts models, different expert routing, so the rollout policy that actually produced the data disagrees with the policy the trainer optimizes. This rollout↔training mismatch destabilizes the standard PPO/GRPO update (Schulman et al., 2017; Shao et al., 2024), whose main stabilizer is a *clip on the sampled token's probability ratio*—a one-dimensional proxy that neither measures how much the full next-token distribution moved nor, under trainer recompute, reflects which distribution generated the data (Figure 1).

We introduce VOKL-CLIP,[1] a token-level vocabulary-KL distance, a *distributional divergence*. At each token we compare the rollout and current policies with a top-k bucket KL (the KL over the top-k next-token probabilities plus a lumped tail), computed for free from the top-k log-probs the inference engine returns on request. Tokens whose distribution has drifted past a threshold are removed from the gradient (stop-gradient) rather than clamped, so VOKL-CLIP *measures* policy

[1]The name abbreviates *Vocabulary-KL Clip*. The "vocabulary KL" is the next-token KL, computed from a top-k bucket coarsening of the full vocabulary distribution and characterized information-theoretically in Appendix A; "clip" denotes the distributional generalization of PPO's scalar ratio clip: each token is gated by this KL distance via stop-gradient (§2.2), while the ordinary ratio clip is kept, widened, as the second clip.

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.

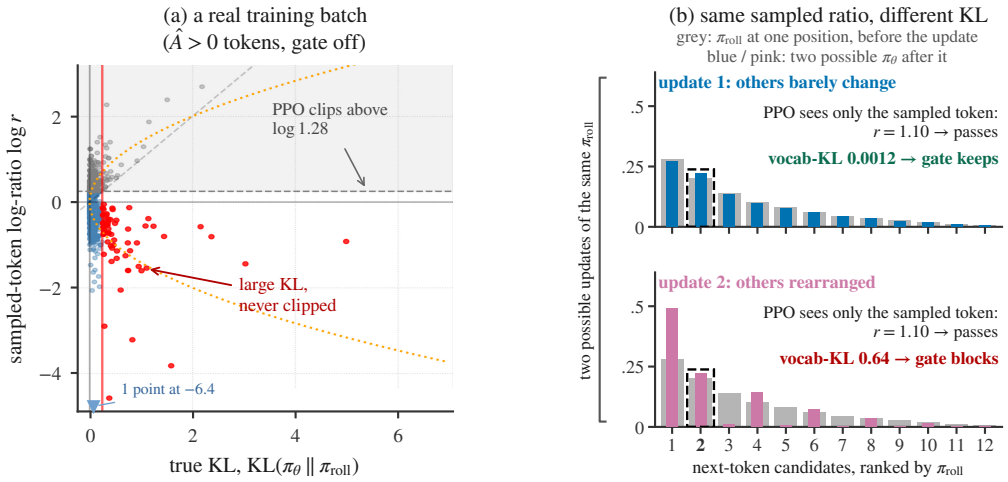


Figure 1: **Ratio clipping is blind to distributional movement.** **(a)** Positive-advantage response tokens from one training batch of a standard-clip run (gate off). The x-axis is the KL between the current and rollout next-token distributions (plotted as $\mathrm{KL}(\pi_\theta \| \pi_{\text{roll}})$; the gate uses the reverse direction), and the y-axis is the log-ratio of the *sampled* token, the one quantity PPO clips. Grey tokens lie above the clip's upper bound $\log 1.28$ and are clipped; blue tokens have KL below $\log 1.28$ and lie below that upper bound, so the clip leaves them unclipped; red tokens have a KL above $\log 1.28 \approx 0.25$ yet are never clipped, because with $\widehat{A} > 0$ the clip acts only on the upper side. They are a small fraction of the batch, but ratio clipping cannot see them at all, and the tokens scatter far from the single-token approximation $\mathrm{KL} \approx \frac{1}{2}(\log r)^2$ (orange dotted). **(b)** Why: grey is the rollout's next-token distribution at one position before the update, and blue and pink are two possible distributions after it. Both give the sampled token (dashed box) the same ratio 1.10, so PPO, which sees only that token, treats them identically; yet the vocabulary-KL distance is 0.0012 when the other candidates barely change and 0.64 when they are rearranged, so the gate keeps the first and blocks the second (§2.2).

movement faithfully instead of conflating it with a ratio clamp. Similar divergence masks have been explored before (§5); the key here is the second clip of the double clip, a widened ratio clip that bounds the heavy-tailed importance weights the vocabulary-KL clip provably cannot control. As in recent work (§5), VOKL-CLIP takes the old log-probs *directly from the rollout*, so the importance ratio is faithful to the true sampler: this removes the bias of trainer recomputation and needs only the lightest mismatch correction—no extra forward pass and no routing replay.

Our contributions are as follows.

(1) Vocabulary-KL double clipping for MoE RL. Targeting the training instability of mixture-of-experts RL, we find that pairing a token-level top-k vocabulary-KL clip, applied by stop-gradient, with a wide ratio clip stabilizes training at the cheapest correction level. VOKL-CLIP is a plug-and-play `policy_loss` that leaves the advantage estimator, rollout and optimizer untouched; on Qwen3-30B-A3B it leads DAPO (Yu et al., 2025) and GSPO (Zheng et al., 2025b), each at its overall best correction level, on all six benchmarks (§2, §4.2).

(2) Why both clips are needed. We prove that the vocabulary-KL clip certifies full-vocabulary total variation—its bucket coarsening gives the sharpest such bound available from top-k information—but cannot bound importance-weight moments, and that even the full KL cannot (Theorem 3.3); the ratio clip supplies this control, consistent with ablations in which removing either clip costs 7–11 AIME24 points (§3, §4.5–4.6).

(3) A four-level mismatch-correction comparison. Extending the routing-replay levels R2 (Zheng et al., 2025b) and R3 (Ma et al., 2026) of Zheng et al. (2025a) with **R0** (Reuse the rollout log-probs) and **R1** (Recompute), we sweep DAPO and GSPO across all four levels; VOKL-CLIP at the cheapest level R0 beats every baseline at every level on AIME24 (§4.4).

2 METHOD

2.1 PRELIMINARIES: GRPO AND THE ROLLOUT–TRAINING MISMATCH

We fine-tune a policy π_θ on prompts $x \sim \mathcal{D}$. For each prompt we sample a group of G responses from the *rollout* (behavior) policy π_{roll} served by the inference engine (vLLM), score them with a rule-based verifier R, and form group-standardized advantages $\widehat{A}_{i,t} = \big(R(x,y_i) - \text{mean}_j R(x,y_j)\big)/\text{std}_j R(x,y_j)$, shared across all tokens t of response y_i (GRPO; Shao et al., 2024). A policy-gradient step reweights each sampled token by the importance ratio

$$r_{i,t}(\theta) = \pi_\theta(y_{i,t} \mid x, y_{i,<t}) \big/ \pi_{\text{old}}(y_{i,t} \mid x, y_{i,<t}), \tag{2.1}$$

and GRPO optimizes the PPO-style clipped surrogate over the sampled response tokens as in Shao et al. (2024), the objective used by DeepSeek-R1 (Guo et al., 2025), here without its reference-KL term:

$$\mathcal{L}_{\text{GRPO}}(\theta) = -\,\mathbb{E}\Big[\tfrac{1}{G}\textstyle\sum_i \frac{1}{|y_i|}\sum_t \min\big[r_{i,t}\widehat{A}_{i,t},\ \text{clip}(r_{i,t}, 1{-}\epsilon, 1{+}\epsilon)\,\widehat{A}_{i,t}\big]\Big], \tag{2.2}$$

the expectation being over $x \sim \mathcal{D}$ and $\{y_i\}_{i=1}^G \sim \pi_{\text{roll}}(\cdot \mid x)$. Here ϵ is the clip half-width, so $[1{-}\epsilon, 1{+}\epsilon]$ is GRPO's ratio-clip band on the *sampled token's* ratio. Following DAPO (Yu et al., 2025) we use *no* reference-KL penalty in the reward or the loss, and aggregate with a token-mean over the response mask. VOKL-CLIP adds a token-level vocabulary-KL distance on top of this objective (§2.2).

Where π_{old} comes from matters. In an LLM-RL system the tokens are drawn by the rollout engine π_{roll}, but the trainer computes gradients under π_θ recomputed on a different backend; the two disagree because of kernel/precision and, for MoE models, expert-routing differences. One option (**R1**) is to *recompute* π_{old} in the trainer; but then $\pi_{\text{old}} \neq \pi_{\text{roll}}$, so equation 2.1 is a biased importance ratio against the true sampler, at the cost of an extra forward pass. VOKL-CLIP instead adopts the *rollout log-probs* directly, $\pi_{\text{old}} \equiv \pi_{\text{roll}}$ (the **R0** setting): the ratio $r_{i,t} = \pi_\theta/\pi_{\text{roll}}$ is then *faithful* to the distribution that actually generated the data, which removes the recompute bias and is the cheapest correction level—no trainer old-log-prob pass and no routing replay (R2/R3). In this setting the ratio itself carries the whole importance correction, with no auxiliary reweighting; the residual mismatch is handled by the double clip below (the vocabulary-KL clip and the wide ratio clip) rather than by heavier correction machinery.

2.2 GATING BY A TOKEN-LEVEL VOCABULARY-KL DISTANCE

On top of GRPO we add *our clip*: a token-level vocabulary-KL distance enforced by a multiplicative *indicator*. Instead of clamping the sampled token's ratio—a one-dimensional proxy blind to how the rest of the next-token distribution shifts (Figure 1)—we keep a token in the update only while an observable coarsening of its next-token distribution stays close to the rollout policy in a genuine KL divergence. The ideal but expensive quantity would be the full next-token KL,

$$D_{\text{KL}}\big(\pi_{\text{roll}}(\cdot \mid \cdot) \,\|\, \pi_\theta(\cdot \mid \cdot)\big) \;=\; \sum_{v \in \mathcal{V}} q_v \,\log \frac{q_v}{p_v}, \qquad q_v = \pi_{\text{roll}}(v \mid \cdot),\ p_v = \pi_\theta(v \mid \cdot), \tag{2.3}$$

over the full vocabulary $\mathcal{V}$. In practice the rollout emits only the top-k next-token log-probs (indices $\mathcal{V}^k_{i,t}$, the k most likely tokens under π_{roll}) and the trainer recomputes π_θ on the *same* indices. We therefore apply a deterministic coarsening, as in prior divergence masks (§5), that leaves those k tokens as singleton buckets and lumps the remaining vocabulary into one tail bucket, with exact masses $q_\perp = 1 - \sum_{v \in \mathcal{V}^k} q_v$ and $p_\perp = 1 - \sum_{v \in \mathcal{V}^k} p_v$, and use the resulting *top-k bucket KL*

$$D^{\text{tkl}}_{i,t} \;=\; \underbrace{\sum_{v \in \mathcal{V}^k_{i,t}} q_v \,\log \frac{q_v}{p_v}}_{\text{top-}k\text{ buckets}} \;+\; \underbrace{q_\perp \,\log \frac{q_\perp}{p_\perp}}_{\text{tail bucket}}. \tag{2.4}$$

This uses only the top-k log-probs the rollout already produces, so it is essentially free to compute. The implementation applies the floor $\delta = 10^{-8}$ only as a floating-point guard when evaluating the two tail logs; the mathematical quantity in equation 2.4 uses the exact masses above. It is a genuine KL divergence between the observable $(k{+}1)$-bucket distributions and, unlike the sampled-token ratio equation 2.1, measures movement across every observed bucket; §3 states what it certifies.

Our clip: a KL-gated surrogate. We turn this distance into *our clip* by gating GRPO's clipped surrogate equation 2.2 with a stop-gradient indicator: a token contributes its gradient only while its top-k vocabulary-KL distance to the rollout policy stays within a threshold τ_i. This gives the complete VOKL-CLIP objective

$$\mathcal{L}(\theta) \;=\; -\,\mathbb{E}_{\substack{x\sim\mathcal{D}\\ \{y_i\}_{i=1}^{G}\sim\pi_{\mathrm{roll}}(\cdot|x)}}\Bigg[\frac{1}{\sum_{j=1}^{G}|y_j|}\sum_{i=1}^{G}\sum_{t=1}^{|y_i|} \mathbf{1}\big[\,D_{i,t}^{\mathrm{tkl}}\le\tau_i\,\big]\;\max\Big(\min\big[r_{i,t}\widehat{A}_{i,t},\;\mathrm{clip}(r_{i,t},\rho_{\mathrm{lo}},\rho_{\mathrm{hi}})\widehat{A}_{i,t}\big],\;c\min(\widehat{A}_{i,t},0)\Big)\Bigg]. \tag{2.5}$$

This indicator is *our clip* and *is* what constrains the update: token (i,t) updates the policy iff $D_{i,t}^{\mathrm{tkl}} \le \tau_i$, a *fixed* threshold $\tau_i \in \{\tau^-,\tau^+\}=\{0.05, 0.10\}$ chosen by the sign of the response advantage $\widehat{A}_i$. The ratio clip inside the $\min$ is kept deliberately *wide*, $[\rho_{\mathrm{lo}},\rho_{\mathrm{hi}}]=[0.5, 4.5]$ (with a dual-clip floor c=10 for negative advantages (Ye et al., 2020)): it does not govern the update but is the second clip, capping the few extreme rollout↔train ratios that the gate provably cannot bound (§3); §4.5 shows both are needed.

Applying the indicator (stop-gradient). We realize the indicator in equation 2.5 by *stop-gradient* ($\mathrm{sg}[\cdot]$) rather than by literally zeroing the loss: a token that has drifted past its threshold keeps its value in the token-mean (so the denominator still counts all response tokens and the logged loss stays comparable) but sends *no* gradient—so it acts as an indicator on the gradient, switching a token's update off exactly when, and only when, the policy has moved too far. The two thresholds are asymmetric, $\tau_i=\tau^+$ if $\widehat{A}_i>0$ else τ^- (with $\tau^+>\tau^-$; since the advantage $\widehat{A}_{i,t}=\widehat{A}_i$ is shared across a response, the gate applies one threshold per response by its sign): we allow more drift when reinforcing good tokens than when suppressing bad ones, since negative-advantage updates are the main source of instability (Qi et al., 2026; Gao et al., 2025). This decouples *measuring* the move (in distribution) from *penalizing* the ratio, which PPO conflates into a single clamp.

Configuration. Unless noted otherwise we use k=16 top-k buckets, vocabulary-KL thresholds τ^-=0.05 (negative advantage) and τ^+=0.10 (positive advantage), extreme-ratio bounds $[\rho_{\mathrm{lo}},\rho_{\mathrm{hi}}]=[0.5, 4.5]$, dual-clip constant c=10, tail floor $\delta=10^{-8}$, and token-mean aggregation; §4.1 gives the shared training setup and §4.5 varies each of these axes.

3 WHAT VOCABULARY-KL CLIPPING CERTIFIES

Fix a decoding context, write $q=\pi_{\mathrm{roll}}(\cdot\mid\cdot)$ and $p=\pi_\theta(\cdot\mid\cdot)$, let $\mathcal{V}^k$ be the rollout top-k set with complement $\mathcal{T}$, and let $\bar q=(q_v: v\in\mathcal{V}^k,\, q_\perp)$ and $\bar p=(p_v: v\in\mathcal{V}^k,\, p_\perp)$ be the coarsened distributions, so $D^{\mathrm{tkl}}=D_{\mathrm{KL}}(\bar q\|\bar p)$; write $\mathrm{TV}(q,p)=\frac12\|q-p\|_1$ and $\chi^2(p\|q)=\sum_v(p_v-q_v)^2/q_v$ (Appendix A.1). The coarsening bounds the full KL only from below: D^{tkl} is its smallest value over *completions* of $(\bar q,\bar p)$ (full-vocabulary pairs that induce them), and its supremum is $+\infty$ once $q_\perp,p_\perp>0$ and $|\mathcal{T}|\ge 2$ (Lemma A.1). We therefore certify total variation.

Theorem 3.1 (Sharp total-variation envelope)**.**

$$\mathrm{TV}(q,p)\;\le\;U_k(q,p):=\mathrm{TV}(\bar q,\bar p)+\min(q_\perp,p_\perp)\;\le\;\sqrt{D^{\mathrm{tkl}}/2}+q_\perp, \tag{3.1}$$

and when $|\mathcal{T}|\ge 2$, U_k is the supremum of $\mathrm{TV}(q,p)$ over all completions of $(\bar q,\bar p)$.

Every token admitted by equation 2.5 thus has $\mathrm{TV}(q,p)\le\sqrt{\tau_i/2}+q_{\perp,i,t}$ (Pinsker terms 0.158 for τ^- and 0.224 for τ^+), so no next-token statistic with range $[m,M]$ moves by more than $M-m$ times this (Corollary A.3), with no tail likelihood-ratio assumption. This holds at the evaluated policy; it does not claim that a later optimizer step keeps every context inside the gate.

Proposition 3.2 (The wide ratio clamp is not redundant)**.** *For every $\eta\in(0,\frac{1}{k+1}]$ there exist q,p with $q_\perp=p_\perp=\eta$ and $D^{\mathrm{tkl}}=0$ but arbitrarily large $\chi^2(p\|q)$. Under R0 the ratio equation 2.1 of a token $Y\sim q$ is p_Y/q_Y, whose second moment is $1+\chi^2(p\|q)$, so the gate alone cannot bound it; the wide ratio and dual clip cap its effective coefficient at $C=\max\{\rho_{\mathrm{hi}},c\}=10$, so each admitted token's direct logit gradient has norm at most $C\sqrt{2(G-1)}$.*

Here p and q agree on every observed bucket—the tail bucket records only the tail totals—while p raises, by an arbitrarily large factor, a tail token that q almost never emits: that token is rarely sampled, but when it is, its ratio is huge. The obstruction is not an artifact of the coarsening. For p, q with equal supports write $r = p/q$, so that $\chi^2(p\|q) = \mathrm{Var}_q(r)$ is the variance of the importance weight, let $d_\infty = \max_{v:q_v>0} |r(v) - 1|$, and let $\kappa(r) = (r - 1 - \log r)/(r - 1)^2$ with $\kappa(1) = \frac{1}{2}$.

Theorem 3.3 (Local equivalence of KL and importance-weight variance)**.** *If $d_\infty \le d < 1$, then*

$$\kappa(1+d)\,\chi^2(p\|q) \;\le\; D_{\mathrm{KL}}(q\|p) \;\le\; \kappa(1-d)\,\chi^2(p\|q),$$

so $D_{\mathrm{KL}}(q\|p) = (\frac{1}{2} + O(d))\,\chi^2(p\|q)$. No such bound holds globally: for every $\varepsilon > 0$ there are q, p with $D_{\mathrm{KL}}(q\|p) < \varepsilon$ and arbitrarily large $\chi^2(p\|q)$ (Appendix A.7).

Even the full KL therefore controls the importance-weight variance only locally, and never a single token's ratio. Conversely, a clamp on the sampled ratio leaves the rest of the distribution free (Figure 1b). The gate certifies distributional change, the clamp bounds the heavy-tailed increments (a ratio bound is likewise what makes χ^2 controllable in Zheng et al., 2026), and neither implies the other; the ablations of §4.5 find each alone insufficient. Appendix A gives the proofs, refinement in k, and martingale concentration results.

4 Experiments

We compare VoKL-Clip with two widely used strong RL baselines for LLM reasoning, DAPO (Yu et al., 2025) and GSPO (Zheng et al., 2025b), under a controlled setting where all methods share the same base model, data, infrastructure, and evaluation protocol, and ask whether it improves over ratio-based clipping (Schulman et al., 2017) (§4.2), how it compares with the heavier corrections R1–R3 (§4.4), and which design choices matter (§4.5).

4.1 Experimental Setup

Model and data. All runs fine-tune **Qwen3-30B-A3B-Base**, a 30B-parameter mixture-of-experts model with 3B active parameters (Yang et al., 2025), on the **DAPO-Math-17k** prompt set (Yu et al., 2025). Training uses the GRPO advantage estimator (Shao et al., 2024) with group-standardized rewards, *no* KL penalty (neither in the reward nor as an auxiliary loss), learning rate 1×10^{-6}, a prompt batch of 512 with 16 sampled responses each, a maximum response length of $20{,}480$ tokens, and a rule-based math verifier (normalized final-answer exact match as in MATH (Hendrycks et al., 2021), with the Minerva answer normalizer (Lewkowycz et al., 2022)) with an overlong-length penalty. Rollouts use vLLM (Kwon et al., 2023) with per-token top-k log-probabilities recorded (k=16 by default), and training runs on the Megatron backend (Shoeybi et al., 2019) through the verl framework (Sheng et al., 2025).

The rollout–training mismatch axis (R0–R3). On MoE models the rollout policy (served by vLLM) and the training policy (recomputed by the trainer) disagree because of expert-routing and numerical differences. We compare four ways of handling this mismatch. **R0** (*Reuse*) takes the old log-probs *directly from the rollout* (the *rollout log-probs*) with no recompute and no routing replay; this is the cheapest level and the one VoKL-Clip is designed for. **R1** (*Recompute*) recomputes the old log-probs in the trainer (an extra forward pass). The remaining two add expert-routing replay: **R2** records the routing during a trainer recompute and replays it in the update to counter routing drift across updates (built on R1; Zheng et al., 2025b), while **R3** recomputes the log-probs and replays the exact experts the rollout used at generation time—Rollout Routing Replay (Ma et al., 2026). The baselines are swept across all four levels in §4.4.

Evaluation. We report accuracy on six benchmarks. For the math competition sets **AIME24/25/26** and **AMC23** we sample n=32 completions at temperature 1.0, top-p 0.7, and report *ave@32* (mean pass@1); AIME24 ave@32 is our primary metric and is logged every 10 steps during training. For **GPQA-Diamond** (Rein et al., 2024) and **MMLU-Pro** (Wang et al., 2024) we use LightEval with @n=8. AIME24 curves are exported on a fixed 0–360-step grid and are stitched from resumed runs; gaps are shown as breaks in the plots. AIME25 is reported only in the tables, and no table value is interpolated.

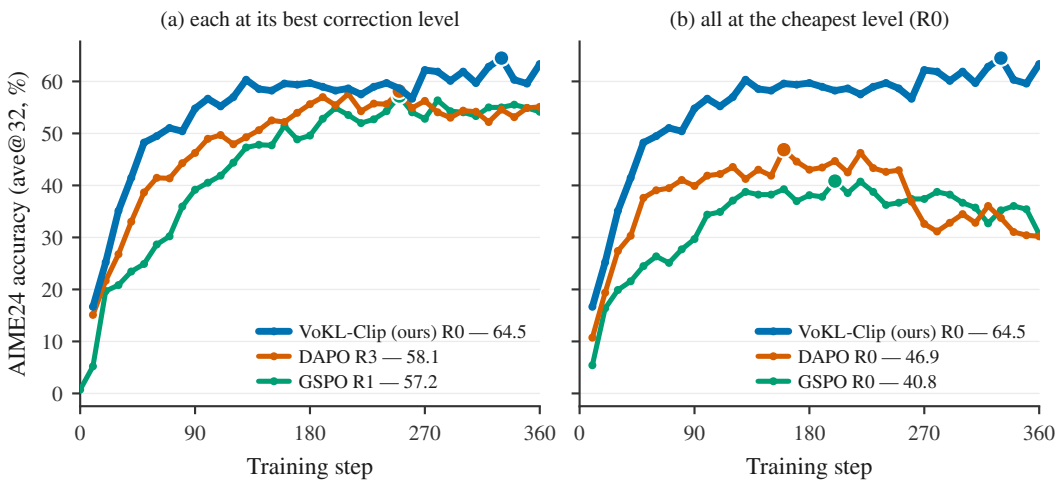


Figure 2: AIME24 accuracy (ave@32) versus training step (0–360). The two panels share a y-axis and VoKL-Clip's curve is the *same* in both—only the baselines move. **(a)** each method at its **overall best** mismatch-correction level: VoKL-Clip at its native R0 (rollout log-probs; Top16, clip $[0.5, 4.5]$, τ^{-}=0.05, τ^{+}=0.10), DAPO at R3, GSPO at R1. Even at their best level the baselines saturate several points below. **(b)** all three at the shared, **cheapest** R0 setting (rollout log-probs, no routing replay): VoKL-Clip keeps climbing to 64.5 while the standard-clip baselines stall and then decline (DAPO 46.9, GSPO 40.8). Dots mark each run's peak; the legend lists each run's level and peak. Curves are W&B AIME24 ave@32 exports on a fixed 10-step grid.

4.2 Main Results

Table 1 compares VoKL-Clip against DAPO and GSPO across six reasoning benchmarks, and Figure 2 shows the AIME24 training curves. To be maximally fair to the baselines, we report *each method at its overall best-performing mismatch-correction level*—VoKL-Clip at its native R0 (rollout log-probs), DAPO at R3, and GSPO at R1 (§4.4, Appendix C.3)—selecting within each method the checkpoint $s^{\star}$ that maximizes AIME24 ave@32 among those evaluated offline and reporting every benchmark at that $s^{\star}$; the full trajectories in Figure 2 make clear that $s^{\star}$ is not a cherry-picked outlier.

Even under this best-vs-best protocol, VoKL-Clip leads on *every* benchmark: 64.5 AIME24 (vs. 58.1 for DAPO and 57.2 for GSPO), 43.2 AIME25 (vs. 39.2 and 41.5), and it is best on AIME26, AMC23, GPQA-Diamond, and MMLU-Pro. Peak-vs-peak on AIME25 the top two are close (44.4 vs. 43.8); with a single seed we do not claim statistical equivalence (Appendix Table 7). VoKL-Clip also dominates the *trajectory* (Figure 2): it reaches 60% AIME24 by step 130 and keeps climbing, while both baselines—even at their best correction level—saturate below 60%.

Table 1: Main results on Qwen3-30B-A3B: each method at its *overall best-performing* mismatch-correction level (§4.4)—VoKL-Clip at its native R0 (rollout log-probs), DAPO at R3, GSPO at R1. AIME24/AIME25/AIME26/AMC23 are ave@32 (n=32, mean pass@1); GPQA-Diamond and MMLU-Pro use LightEval (@n=8). Best per column in **bold**.

Method	AIME24	AIME25	AIME26	AMC23	GPQA-D	MMLU-Pro
VoKL-Clip (ours), R0	**64.5**	**43.2**	**57.6**	**93.2**	**62.7**	**75.5**
DAPO (R3)	58.1	39.2	51.4	89.2	61.2	70.1
GSPO (R1)	57.2	41.5	50.2	89.1	59.8	65.4

4.3 Head-to-head at the cheap R0 setting

Table 2 removes the best-vs-best handicap and places all methods at the shared, cheapest **R0** setting (rollout log-probs, no routing replay), scored on all six benchmarks. Here VoKL-Clip's overall

Table 2: Head-to-head at the shared cheap **R0** setting (rollout log-probs, no routing replay) on Qwen3-30B-A3B: each method at a single checkpoint (Appendix B.2). AIME24/AIME25/AIME26/AMC23 are ave@32 (n=32, mean pass@1); GPQA-Diamond and MMLU-Pro use LightEval (@n=8). DAPO $[0.5, 4.5]$ is a non-standard variant that adopts VoKL-Clip's wide clamp. Best per column in **bold**.

Method	AIME24	AIME25	AIME26	AMC23	GPQA-D	MMLU-Pro
VoKL-Clip (ours)	**64.5**	43.2	**57.6**	**93.2**	**62.7**	**75.5**
DAPO ($[0.8, 1.28]$)	44.7	34.6	37.7	77.4	56.9	65.8
DAPO ($[0.5, 4.5]$, our clamp)	57.1	**43.8**	50.1	92.0	61.6	74.5
GSPO	40.8	29.1	28.8	75.2	55.5	64.8

margin is widest: it leads by a wide margin on AIME24, AIME26, AMC23, GPQA-Diamond and MMLU-Pro, because the standard-clip baselines (DAPO $[0.8, 1.28]$, GSPO) are *weak* at R0—they stall and then decline (Figure 2b), while their rollout↔training KL stays bounded; catastrophic KL divergence appears only in DAPO's R2 run, whose per-step rollout↔training KL climbs from a median of 1.5×10^{-3} (steps $\leq$ 180) to 69.9 at step 197 while its gradient norm spikes to 124.7 (Figure 5, Appendix C.6). The only close call is AIME25, where the wide-clip DAPO variant that borrows VoKL-Clip's clamp is essentially tied (43.8 vs. 43.2; single seed, so we make no significance claim). Holding the mismatch-correction level fixed isolates the faithful distributional vocabulary-KL distance as the source of the gain.

4.4 Correcting the Rollout–Training Mismatch

A natural alternative to a better vocabulary-KL distance is to *eliminate* the rollout↔training mismatch directly. At the cheap R0 setting (rollout log-probs, no replay) the standard-clip baselines are weak: DAPO peaks at only 46.9 and GSPO at 40.8, and both then drift downward (Figure 2b). GSPO's sequence-level clip binds often at this setting: its logged (token-weighted) clip fraction has a median of 0.28, so roughly a quarter of response tokens lie in clipped sequences, against 0.0030 for VoKL-Clip, which is consistent with its accuracy. These baselines must therefore pay for heavier corrections: an extra recompute (R1), or expert-routing replay (R2/R3). Table 8 (Appendix C.3) sweeps both baselines across all four levels. The corrections help unevenly: GSPO is best *overall* with a plain recompute (R1), which beats its own R3 on four of the six benchmarks despite a marginally lower AIME24 peak (57.2 vs. 57.9), and DAPO likewise with rollout-side routing replay (R3, 58.1), while DAPO's R2 run momentarily reaches 61.0 but collapses by step 200 (Appendix C.3). Crucially, on AIME24 VoKL-Clip at its native R0 (64.5) exceeds *every* baseline at *every* correction level while being the cheapest—R0 skips both the trainer's old-log-prob forward pass and the MoE routing replay of R2/R3. A faithful distributional vocabulary-KL distance is thus more effective than progressively heavier mismatch-correction machinery.

4.5 Ablations

Table 3: Ablations under R0. Every column is read at one checkpoint per row (Appendix B.2); the last four columns are its offline evaluation. The default (Top16, clip $[0.5, 4.5]$, tkl τ^-=0.05, τ^+=0.10, full double-clip) gives the best AIME24: a tight clip range collapses training and removing the extreme-ratio clamp (VoKL-only) costs >10 points. Best per column in **bold**.

Configuration	AIME24	AIME25	AIME26	AMC23	GPQA-D	MMLU-Pro
VoKL-Clip (Top16, main)	**64.5**	43.2	**57.6**	93.2	**62.7**	**75.5**
TopK = 8	61.7	44.2	53.8	**93.3**	62.4	75.1
TopK = 32	63.2	**45.8**	53.3	91.3	61.6	74.2
clip $[0.8, 1.28]$ (tight)	47.9	32.9	30.3	72.1	57.3	64.5
VoKL-only (no clamp)	53.3	36.4	47.2	89.5	60.0	73.1

Table 3 varies one design axis at a time around the default configuration (Top16, clip $[0.5, 4.5]$, tkl thresholds τ^-=0.05, τ^+=0.10, full double-clip loss). Three findings stand out. **(i) Bucket size:** k=16 gives the highest AIME24 peak (64.5, vs. 61.7 for k=8 and 63.2 for k=32), but the method

is robust to this choice: the two alternatives trail by only 2.8 and 1.3 points. A plausible cause: the top-k buckets already capture almost the whole distribution; the tail mass outside top-k stays tiny and shrinks with k (mean $1.6/1.2/0.6\%$ for k=$8/16/32$; Appendix A.4), so coarsening loosens the TV certificate by at most $q_\perp$ (Theorem 3.1) and larger k has little left to add. **(ii) The wide clip range is essential:** replacing the extreme-ratio clamp $[0.5, 4.5]$ with a tight DAPO-style $[0.8, 1.28]$ collapses training (AIME24 falls to 47.9 and degrades after $\sim$step 250; Figure 4b, Appendix C.1), confirming that the top-k vocabulary-KL distance, not ratio clipping, should govern the update. **(iii) The two clips are complementary—we tried each alone.** Neither piece reaches the full loss on its own. Keeping only our KL-distance indicator and dropping the wide ratio clamp (VoKL-only) costs over ten AIME24 points (53.3 vs. 64.5); conversely, keeping only the wide clamp *without* the KL indicator (the DAPO $[0.5, 4.5]$ variant that borrows our clamp) peaks at only 57.1 (Table 7). The two pieces are complementary rather than redundant because they act on *opposite sides* of the same quantity: without the wide clamp entropy stays pinned near 0.07 and the policy barely explores, while without the KL indicator entropy inflates late in training to 1.11 (§4.6). Only together do they hold exploration inside a usable band, and each alone gives up 7–11 AIME24 points. This is consistent with the division of labor of §3. The gate caps each admitted token's next-token TV from the rollout at $\sqrt{\tau_i/2} + q_{\perp,i,t}$ (Theorem 3.1) but cannot see a rare tail token whose probability jumps (Proposition 3.2): such a token can cost D^{tkl} nothing and the full KL only on the order of its new mass p_v, yet adds about p_v^2/q_v to the importance-weight variance $\chi^2(p\|q)$ (Theorem 3.3). The wide and dual clips cap any such token's coefficient at C and are inactive inside $[0.5, 4.5]$, where only the gate acts, so each clip covers the other's blind spot.

4.6 Training Dynamics

Figure 3 switches the gate and the clamp on and off under R0, which separates the two clips of Proposition 3.2 empirically.[2] *The clamp is what bounds the mismatch.* Removing it while keeping the gate is the worst run in panel (b): its rollout↔training KL sits at a median of 5.0×10^{-3} against our 2.6×10^{-3} and spikes to 5.2×10^{-2} at step 100, an order of magnitude above anything the clamped runs reach, at $6\times$ our clip fraction (0.018 vs. 0.0030). *The gate is what holds entropy.* Among the clamped runs the KL trajectories are comparable (ours and the no-gate run share the *same* wide clamp $[0.5, 4.5]$, a matched-clamp control), so the gate does not act by suppressing that signal further. What separates them is *entropy*, held from opposite sides; the accuracy costs are from Tables 3 and 2. Without the clamp, entropy stays near 0.07 (median 0.068) for the whole run apart from one excursion to 0.43 at step 180—the policy barely explores—at a cost of over ten AIME24 points. Without the gate it starts from the same level (first-half median 0.38 vs. 0.37) but inflates late to 1.11 by step 360, twice its own median, costing 7.4 points; the tight clamp $[0.8, 1.28]$ instead peaks at 0.63 and collapses to 0.07. With both, entropy neither collapses nor inflates: it rises to a median of 0.45 and ends there (0.449 at step 360). Response length (panel c) grows fastest for ours (4.9k tokens at step 40 vs. at most 2.9k for the other three runs). Neither component is sufficient alone, consistent with §3.

5 Related Work

Trust regions and divergence masks. PPO (Schulman et al., 2017) turns TRPO's KL constraint (Schulman et al., 2015) into a sampled-token ratio clip, kept by GRPO (Shao et al., 2024), decoupled by DAPO (Yu et al., 2025) and made sequence-level by GSPO (Zheng et al., 2025b). This token-level clip neither bounds the ratio nor enforces a trust region (Wang et al., 2020; Engstrom et al., 2020); Truly PPO triggers it on a KL condition instead. For LLMs, TROLL (Becker et al., 2026) projects next-token distributions into a KL trust region, TRM (Li et al., 2026b) masks sequences by their maximum token KL, and DPPO (Qi et al., 2026)—the closest to our work—replaces the ratio clip by a direction-aware mask on a binary or top-k divergence to the rollout policy, with smooth (Yao et al., 2026) and predictive (Zhou et al., 2026) variants. VoKL-Clip shares this rollout-anchored divergence mask but not DPPO's treatment of the ratio: DPPO drops the clip (its large-scale runs only truncate the importance weight), whereas we prove that the top-k coarsening, though it yields the sharpest full-vocabulary TV certificate top-k information allows (derived concurrently by C-DPPO; Li

[2] At R0 the trainer's old policy *is* the rollout policy, $\pi_{\mathrm{old}} \equiv \pi_{\mathrm{roll}}$, so the PPO-KL and the rollout↔training KL are the same quantity by construction; we therefore plot the latter only.

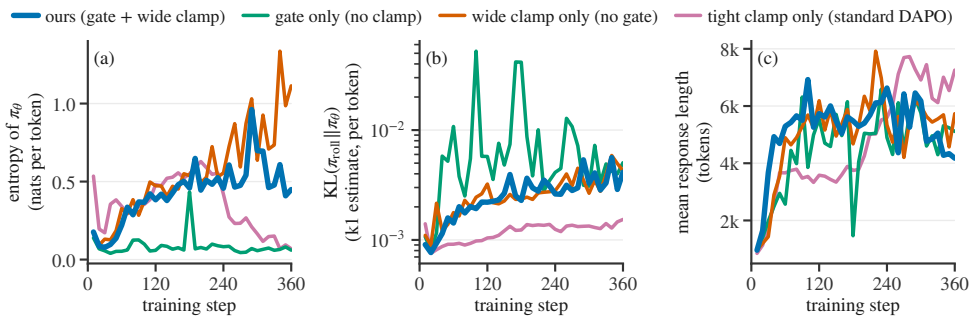


Figure 3: Component dynamics under R0: the gate and the clamp switched on and off. **(a)** Entropy of the current policy π_θ (nats per token). **(b)** Rollout↔training KL, the per-token (k_1) estimate $\mathrm{KL}(\pi_{\text{roll}}\|\pi_\theta) \approx$ mean of $\log \pi_{\text{roll}}(y) - \log \pi_\theta(y)$ over sampled response tokens, measured against the current policy during each update (log scale). **(c)** Mean response length (tokens). Ours and the no-gate run share the *same* wide clamp $[0.5, 4.5]$, so that pair isolates the gate: comparable KL, separating entropy. Dropping the clamp instead (gate only) roughly doubles the median KL and lets it spike to 5.2×10^{-2} at step 100, an order of magnitude above the clamped runs. The tight-clamp run is standard DAPO ($[0.8, 1.28]$). At R0 the PPO-KL coincides with the rollout↔training KL ($\pi_{\text{old}} \equiv \pi_{\text{roll}}$), so it is not shown separately.

et al., 2026a), cannot control the importance weight's second moment, so the gate must be paired with a ratio bound. We keep a wide clamp, and removing it costs over ten AIME24 points (§4.5). Our gate also drops DPPO's direction test: its radius is set by the advantage sign alone.

Rollout–training mismatch and MoE instability. With generation and training on different engines, the data-generating policy differs from the trainer's—through numerics and, on MoE models, expert routing, which makes MoE RL particularly unstable. Remedies align the two engines (FP16, Qi et al., 2025; R3, Ma et al., 2026), reweight or mask by the rollout–training discrepancy (Yao et al., 2025; Liu et al., 2025; Ling Team, 2025; DeepSeek-AI, 2025), or counter routing drift across updates (R2, Zheng et al., 2025b; Zhang et al., 2026); most of them, like the standard pipeline, recompute the old log-probs in the training engine (R1). GSPO allows the rollout engine's likelihoods instead (Zheng et al., 2025b), and DPPO anchors its trust region on them, removing the need for the recompute (Qi et al., 2026); VoKL-Clip, too, runs without it (at its native R0 the old log-probs come straight from the rollout engine), and we compare it with DAPO and GSPO, each run at every mismatch-correction level R0–R3: heavier corrections do not buy stability (DAPO's R2 run even collapses, §4.4), and it is the retained clamp, not the gate, that bounds the heavy-tailed mismatch (§4.6).

6 Conclusion

We presented VoKL-Clip, which stabilizes RL of language models with a double clip: a top-k bucket-KL clip, enforced by stop-gradient, whose coarsening yields the sharpest full-vocabulary total-variation bound available from top-k probabilities, and a wide ratio clip that supplies the moment control the first provably cannot (§3). With both in place it reuses the rollout log-probs directly—the lightest correction—as a plug-and-play `policy_loss`. On the MoE model Qwen3-30B-A3B it improves over the strongest DAPO and GSPO baselines and, at its cheap R0 setting, beats them on AIME24 at every mismatch-correction level, with stable training dynamics. Our study has several limitations: a single MoE model and a single math training set, single-seed runs, no wall-clock measurements, and no clean $\tau \to \infty$ (gate-off) ablation across correction levels to separate the rollout-log-prob ratio from the gate. The bucket KL is only an observable coarsening of the full divergence, so we certify total variation rather than the full KL (Appendix A). Extending VoKL-Clip to long-horizon agentic and model-judged-reward settings, to dense and larger models, and to multi-seed and efficiency studies is the natural next step.

AI USE STATEMENT

LLMs are the object of study in this paper. In preparing it, the authors used LLM-based assistants for language editing and for checking parts of the text, and reviewed all such output. The authors checked all numbers against their experiment records and take full responsibility for the content of the paper.

REFERENCES

Philipp Becker, Niklas Freymuth, Serge Thilges, Fabian Otto, and Gerhard Neumann. TROLL: Trust regions improve reinforcement learning for large language models. In *International Conference on Learning Representations (ICLR)*, 2026.

DeepSeek-AI. DeepSeek-V3.2: Pushing the frontier of open large language models. *arXiv preprint arXiv:2512.02556*, 2025.

Logan Engstrom, Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Firdaus Janoos, Larry Rudolph, and Aleksander Madry. Implementation matters in deep RL: A case study on PPO and TRPO. In *International Conference on Learning Representations (ICLR)*, 2020.

David A. Freedman. On tail probabilities for martingales. *The Annals of Probability*, 3(1):100–118, 1975.

Chang Gao, Chujie Zheng, Xiong-Hui Chen, Kai Dang, Shixuan Liu, Bowen Yu, An Yang, Shuai Bai, Jingren Zhou, and Junyang Lin. Soft adaptive policy optimization. *arXiv preprint arXiv:2511.20347*, 2025.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645:633–638, 2025.

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

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 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 29th Symposium on Operating Systems Principles (SOSP)*, 2023.

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, Yuhuai Wu, Behnam Neyshabur, Guy Gur-Ari, and Vedant Misra. Solving quantitative reasoning problems with language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 3843–3857, 2022.

Cheng Li, Jiexiong Liu, Yixuan Chen, and Chi Hong. Train what you deploy: Token-faithful post-training of a production coding agent. *arXiv preprint arXiv:2609.04678*, 2026a.

Yingru Li, Jiacai Liu, Jiawei Xu, Yuxuan Tong, Ziniu Li, and Baoxiang Wang. Trust region masking for long-horizon LLM reinforcement learning. In *International Conference on Machine Learning (ICML)*, 2026b.

Ling Team. Every step evolves: Scaling reinforcement learning for trillion-scale thinking model. *arXiv preprint arXiv:2510.18855*, 2025.

Jiacai Liu, Yingru Li, Yuqian Fu, Jiawei Wang, Qian Liu, and Yu Shen. When speed kills stability: Demystifying RL collapse from the training-inference mismatch. Blog post, 2025. URL https://richardli.xyz/rl-collapse.

Wenhan Ma et al. Stabilizing MoE reinforcement learning by aligning training and inference routers. In *International Conference on Machine Learning (ICML)*, 2026.

Penghui Qi, Zichen Liu, Xiangxin Zhou, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Defeating the training-inference mismatch via FP16. *arXiv preprint arXiv:2510.26788*, 2025.

Penghui Qi, Xiangxin Zhou, Zichen Liu, Tianyu Pang, Chao Du, Min Lin, and Wee Sun Lee. Rethinking the trust region in LLM reinforcement learning. In *International Conference on Machine Learning (ICML)*, 2026.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A graduate-level Google-proof Q&A benchmark. In *Conference on Language Modeling (COLM)*, 2024.

John Schulman, Sergey Levine, Philipp Moritz, Michael I. Jordan, and Pieter Abbeel. Trust region policy optimization. In *International Conference on Machine Learning (ICML)*, 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, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

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 (EuroSys)*, pp. 1279–1297, 2025.

Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. Megatron-LM: Training multi-billion parameter language models using model parallelism. *arXiv preprint arXiv:1909.08053*, 2019.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. MMLU-Pro: A more robust and challenging multi-task language understanding benchmark. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2024.

Yuhui Wang, Hao He, and Xiaoyang Tan. Truly proximal policy optimization. In *Proceedings of the 35th Conference on Uncertainty in Artificial Intelligence (UAI)*, volume 115 of *Proceedings of Machine Learning Research*, pp. 113–122, 2020.

An Yang et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Feng Yao, Liyuan Liu, Dinghuai Zhang, Chengyu Dong, Jingbo Shang, and Jianfeng Gao. Your efficient RL framework secretly brings you off-policy RL training. Blog post, 2025. URL https://fengyao.notion.site/off-policy-rl.

Jiarui Yao, Xiangxin Zhou, Penghui Qi, Wee Sun Lee, Liefeng Bo, and Tianyu Pang. Rethinking the divergence regularization in LLM RL. *arXiv preprint arXiv:2606.09821*, 2026.

Deheng Ye, Zhao Liu, Mingfei Sun, Bei Shi, Peilin Zhao, Hao Wu, Hongsheng Yu, Shaojie Yang, Xipeng Wu, Qingwei Guo, et al. Mastering complex control in MOBA games with deep reinforcement learning. In *AAAI Conference on Artificial Intelligence*, 2020.

Qiying Yu et al. DAPO: An open-source LLM reinforcement learning system at scale. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Di Zhang et al. Towards stable and effective reinforcement learning for mixture-of-experts. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 25446–25457, 2026.

Chujie Zheng, Kai Dang, Bowen Yu, Mingze Li, Huiqiang Jiang, Junrong Lin, Yuqiong Liu, Hao Lin, Chencan Wu, Feng Hu, An Yang, Jingren Zhou, and Junyang Lin. Stabilizing reinforcement learning with LLMs: Formulation and practices. *arXiv preprint arXiv:2512.01374*, 2025a.

Chujie Zheng et al. Group sequence policy optimization. *arXiv preprint arXiv:2507.18071*, 2025b.

Haizhong Zheng, Jiawei Zhao, and Beidi Chen. Prosperity before collapse: How far can off-policy RL reach with stale data on LLMs? In *International Conference on Learning Representations (ICLR)*, 2026.

Xiangxin Zhou, Jiarui Yao, Penghui Qi, Bowen Ping, Jiaqi Tang, Haonan Wang, and Tianyu Pang. Predictive divergence masks for LLM RL. *arXiv preprint arXiv:2607.10848*, 2026.

A Omitted Proofs

A.1 Notation

Throughout this appendix we fix a decoding context and suppress its index. Write $q = \pi_{\mathrm{roll}}(\cdot \mid \cdot)$ and $p = \pi_\theta(\cdot \mid \cdot)$ for the rollout and current next-token distributions over the vocabulary $\mathcal{V}$. Let $\mathcal{V}^k$ be the rollout top-k set, $\mathcal{T} = \mathcal{V} \setminus \mathcal{V}^k$ its complement, and

$$a = q_\perp = q(\mathcal{T}), \qquad b = p_\perp = p(\mathcal{T}).$$

When $a, b > 0$ let $\widetilde{q} = q(\cdot \mid \mathcal{T})$ and $\widetilde{p} = p(\cdot \mid \mathcal{T})$ be the tail-conditional distributions; zero tail masses are handled by the usual KL conventions. The two *coarsened* $(k{+}1)$-bucket distributions are

$$\bar{q} = (q_v : v \in \mathcal{V}^k,\, a), \qquad \bar{p} = (p_v : v \in \mathcal{V}^k,\, b),$$

so that the quantity the algorithm computes is $D^{\mathrm{tkl}} = D_{\mathrm{KL}}(\bar{q} \| \bar{p})$. We write $\mathrm{TV}(q,p) = \frac{1}{2}\|q - p\|_1$ for total variation, $\chi^2(p\|q) = \sum_v (p_v - q_v)^2/q_v$ for the Pearson divergence, and $U_k(q,p) = \mathrm{TV}(\bar{q},\bar{p}) + \min(a,b)$ for the envelope of Theorem 3.1. A *compatible completion* (*completion* for short) is any pair (q,p) of full-vocabulary distributions inducing the observed $(\bar{q},\bar{p})$ on the fixed partition $(\mathcal{V}^k, \mathcal{T})$; we do not require $\mathcal{V}^k$ to remain the top-k set of q. If this is imposed, the extremal completions below still respect it whenever $|\mathcal{T}| \geq \lceil q_\perp / q_{(k)} \rceil + 1$, where $q_{(k)} = \min_{v \in \mathcal{V}^k} q_v$.

A.2 Identifiability of the full KL

The coarsening determines the full KL only from below.

Lemma A.1 (KL decomposition and non-identifiability)**.** *$D_{\mathrm{KL}}(q\|p) = D^{\mathrm{tkl}} + a\, D_{\mathrm{KL}}(\widetilde{q}\|\widetilde{p})$. Consequently D^{tkl} is the smallest full KL among compatible completions and is attained when $\widetilde{q} = \widetilde{p}$. If $a, b > 0$ and $\mathcal{T}$ contains at least two symbols, the supremum of the full KL over compatible completions is $+\infty$.*

Proof. Splitting the full KL over head and tail,

$$\begin{aligned} D_{\mathrm{KL}}(q\|p) &= \sum_{v \in \mathcal{V}^k} q_v \log \frac{q_v}{p_v} + \sum_{v \in \mathcal{T}} a\widetilde{q}_v \log \frac{a\widetilde{q}_v}{b\widetilde{p}_v} \\ &= \sum_{v \in \mathcal{V}^k} q_v \log \frac{q_v}{p_v} + a \log \frac{a}{b} + a D_{\mathrm{KL}}(\widetilde{q}\|\widetilde{p}) \;=\; D^{\mathrm{tkl}} + a D_{\mathrm{KL}}(\widetilde{q}\|\widetilde{p}). \end{aligned} \tag{A.1}$$

The residual is nonnegative and vanishes when the tail-conditionals agree, giving the minimum. For the supremum, keep any admissible $\widetilde{q}$ (e.g. the true one), pick $u \in \mathcal{T}$ with $\widetilde{q}(u) > 0$, and let $\widetilde{p}(u) = e^{-L}$ with the remaining mass spread over $\mathcal{T} \setminus \{u\}$; by the log-sum inequality $D_{\mathrm{KL}}(\widetilde{q}\|\widetilde{p}) \geq \widetilde{q}(u)\big(L + \log \widetilde{q}(u)\big) + (1 - \widetilde{q}(u)) \log(1 - \widetilde{q}(u)) \to \infty$. The construction respects the top-k order and, with $\widetilde{p}$ of full support, strictly positive softmax probabilities. □

Lemma A.1 is the reason we do not certify the full KL: no assumption-free upper bound on it exists from top-k information, so any such bound must import an unobservable tail likelihood-ratio constant. Total variation, by contrast, admits an exact envelope.

A.3 Proof of Theorem 3.1

Proof. Decompose total variation over head and tail,

$$\mathrm{TV}(q,p) = \frac{1}{2}\sum_{v\in\mathcal{V}^k} |q_v - p_v| + \frac{1}{2}\|a\widetilde{q} - b\widetilde{p}\|_1.$$

If $a \geq b$, the triangle inequality gives

$$\frac{1}{2}\|a\widetilde{q} - b\widetilde{p}\|_1 \leq \frac{a-b}{2} + b\,\mathrm{TV}(\widetilde{q},\widetilde{p}) \leq \frac{|a-b|}{2} + \min(a,b),$$

and the case $b > a$ is symmetric; adding the head terms yields $\mathrm{TV}(q,p) \leq \mathrm{TV}(\bar{q},\bar{p}) + \min(a,b) = U_k(q,p)$. If $|\mathcal{T}| \geq 2$ (or $\min(a,b) = 0$), the bound is attained: choosing disjoint tail-conditionals (e.g. $\widetilde{p}$ a point mass at some $w \in \mathcal{T}$ and $\widetilde{q}$ supported on $\mathcal{T}\setminus\{w\}$) makes $\frac{1}{2}\|a\widetilde{q} - b\widetilde{p}\|_1 = (a+b)/2$, and with a full-support requirement disjointness is approached arbitrarily closely. Hence U_k is the exact worst case over compatible completions, i.e. the smallest uniform upper bound obtainable from $(\bar{q},\bar{p})$. Finally Pinsker's inequality on the $(k{+}1)$-bucket distributions gives $\mathrm{TV}(\bar{q},\bar{p}) \leq \sqrt{D_{\mathrm{KL}}(\bar{q}\|\bar{p})/2} = \sqrt{D^{\mathrm{tkl}}/2}$, and $\min(a,b) \leq a = q_\perp$. □

Proposition A.2 (Monotone refinement in k)**.** *For nested rollout top-k sets, $D_1^{\mathrm{tkl}} \leq D_2^{\mathrm{tkl}} \leq \cdots \leq D_{\mathrm{KL}}(q\|p)$ and $U_1 \geq U_2 \geq \cdots \geq \mathrm{TV}(q,p)$.*

Proof. The $(k{+}1)$-bucket partition refines the k-bucket partition, so the KL chain rule—equivalently data processing under the coarsening map—makes the coarsened KL nondecreasing. For TV, $\frac{1}{2}|a-b| + \min(a,b) = \frac{1}{2}(a+b)$ gives $U_k = \frac{1}{2}\sum_{v\in\mathcal{V}^k}|q_v - p_v| + \frac{1}{2}(q_\perp + p_\perp)$, and moving a symbol v from the tail to the head changes it by $\frac{1}{2}(|q_v - p_v| - q_v - p_v) = -\min(q_v, p_v) \leq 0$. Both sequences equal their full-distribution counterparts once every symbol is exposed. □

A.4 The selective-update certificate

Corollary A.3 (Selective-update certificate)**.** *Every token admitted by equation 2.5 satisfies $\mathrm{TV}(q,p) \leq \varepsilon_{i,t} := \sqrt{\tau_i/2} + q_{\perp,i,t}$, hence $|\mathbb{E}_q f - \mathbb{E}_p f| \leq (M-m)\,\varepsilon_{i,t}$ for every token statistic $f : \mathcal{V} \to [m,M]$; a rejected token contributes zero direct gradient.*

Proof. An admitted token satisfies $D_{i,t}^{\mathrm{tkl}} \leq \tau_i$; substituting into equation 3.1 gives $\mathrm{TV}(q,p) \leq \sqrt{\tau_i/2} + q_{\perp,i,t}$. For $f \in [m,M]$ the dual characterization of total variation gives $|\mathbb{E}_q f - \mathbb{E}_p f| \leq (M-m)\,\mathrm{TV}(q,p)$. The stop-gradient implementation makes the direct derivative of a rejected token exactly zero. □

The rollout top-k captures most of the probability mass in our runs: averaged over R0 training steps 0–360 it captures $98.4/98.8/99.4\%$ for k=8/16/32, i.e. mean rollout tails $1.6/1.2/0.6\%$. These averages describe the scale of the correction but are *not* substituted for the tokenwise $q_{\perp,i,t}$ in the certificate. Unlike an unknown tail likelihood-ratio bound, both $q_{\perp,i,t}$ and the sharper $U_k(q,p)$ are computable from the probabilities the loss already uses.

A.5 Proof of Proposition 3.2

Proof. Fix $\eta \in (0, \frac{1}{k+1}]$ and let $p_v = q_v = (1-\eta)/k$ on $\mathcal{V}^k$; since $(1-\eta)/k \geq \eta$, no tail probability below exceeds a head probability, so $\mathcal{V}^k$ is indeed the rollout top-k set. On two tail symbols u, w and for $0 < \gamma < \eta/2$ set

$$q_u = \gamma, \quad q_w = \eta - \gamma, \qquad p_u = p_w = \eta/2.$$

The observed head probabilities and tail masses coincide, so $D^{\mathrm{tkl}} = 0$, yet

$$\chi^2(p\|q) = \sum_v \frac{(p_v - q_v)^2}{q_v} \;\geq\; \frac{(\eta/2-\gamma)^2}{\gamma} \longrightarrow \infty \qquad (\gamma \downarrow 0).$$

Since $\mathbb{E}_q[r(Y)^2] = 1 + \chi^2(p\|q)$ for $r = p/q$, neither a zero coarsened KL nor a small full TV controls the raw importance-weight second moment.

For the gradient bound, let $\zeta_i = (R_i - \bar{R})/\sigma$ be the population-standardized rewards of a nonconstant group, so $\sum_i \zeta_i = 0$ and $\sum_i \zeta_i^2 = G$. Cauchy–Schwarz gives

$$\zeta_i^2 = \Big(\sum_{j \neq i} \zeta_j\Big)^2 \leq (G-1)\sum_{j \neq i} \zeta_j^2 = (G-1)(G-\zeta_i^2),$$

hence $|\zeta_i| \leq \sqrt{G-1}$ (the sample standard deviation only tightens this). For softmax logits z, $\|\nabla_z \log p(Y)\|_2 = \|e_Y - p\|_2 \leq \sqrt{2}$. In equation 2.5, for $\widehat{A} > 0$ and $r > \rho_{\mathrm{hi}}$ the clip branch is constant, and for $\widehat{A} < 0$ and $r > c$ the outer $\max$ selects the constant $c\widehat{A}$; both have zero gradient, so the effective ratio coefficient is at most $C = \max\{\rho_{\mathrm{hi}}, c\} = 10$. Multiplying the three factors gives $\|g_{i,t}^{\mathrm{logit}}\|_2 \leq C\sqrt{2(G-1)}$ on an admitted token, and the detach makes it zero on a rejected one. □

A.6 Concentration from bounded increments

The deterministic bound above is exactly the hypothesis martingale concentration needs.

Proposition A.4 (Gradient concentration in logit space)**.** *Let (g_j) be normalized logit-gradient contributions in a fixed Euclidean space (e.g. the zero-padded logit space $\mathbb{R}^{G \times T_{\max} \times |\mathcal{V}|}$), adapted to a filtration $(\mathcal{F}_j)$ with $\mathcal{F}_j$ containing the first j groups, and with $\|g_j\|_2 \leq L_g := C\sqrt{2(G-1)}$. For any fixed unit vector ν of that space, with probability at least $1-\delta$,*

$$\Big|\frac{1}{n}\sum_{j=1}^{n} \big\langle \nu,\ g_j - \mathbb{E}[g_j \mid \mathcal{F}_{j-1}]\big\rangle\Big| \leq L_g\sqrt{\frac{2\log(2/\delta)}{n}}.$$

Proof. The summands form a bounded martingale difference sequence with increments in an interval of length at most $2L_g$; conditional Hoeffding–Azuma gives the stated deviation. □

This is a logit-space statement; a parameter-space version additionally needs a bound on the network Jacobian, or uses the global gradient-norm clip of our implementation.

Proposition A.5 (Autoregressive concentration of unobserved-tail exposure)**.** *Along a rollout let $\mathcal{F}_{t-1}$ be the history before token t, let $\mathcal{V}_t^k$ be the rollout top-k set, and put $I_t = \mathbf{1}\{Y_t \notin \mathcal{V}_t^k\}$ and $q_{\perp,t} = q_t(\mathcal{V} \setminus \mathcal{V}_t^k)$. If $\sum_{t=1}^{n} q_{\perp,t} \leq Q$ almost surely, then for every $x > 0$,*

$$\mathbb{P}\Big(\sum_{t=1}^{n} I_t > Q + \sqrt{2Qx} + \tfrac{2x}{3}\Big) \leq e^{-x}.$$

Proof. Since $\mathcal{V}_t^k$ and q_t are determined before Y_t is sampled, $\mathbb{E}[I_t \mid \mathcal{F}_{t-1}] = q_{\perp,t}$, so $M_n = \sum_t (I_t - q_{\perp,t})$ is a martingale with increments bounded above by 1 and predictable quadratic variation $V_n = \sum_t q_{\perp,t}(1 - q_{\perp,t}) \leq Q$. Freedman's inequality (Freedman, 1975, Thm. (4.1)), in its Bernstein form, gives $\mathbb{P}(M_n \geq s,\ V_n \leq Q) \leq \exp\{-s^2/[2(Q + s/3)]\}$, and $s = \sqrt{2Qx} + 2x/3$ makes the exponent at least x. □

Proposition A.5 is why martingale concentration, rather than an i.i.d. token bound, is the natural tool here: it converts cumulative rollout tail mass into a high-probability bound on how often the sampled action falls in the unresolved tail. The sampled-ratio clamp acts on every sampled token with an extreme ratio: on these tail tokens, where Proposition 3.2 shows the unclipped weight can be arbitrarily heavy-tailed, and also on low-probability head tokens admitted by the gate (Proposition A.8).

A.7 KL versus importance-weight variance: local equivalence, no global control

Keep $q = \pi_{\mathrm{roll}}(\cdot \mid \cdot)$ and $p = \pi_\theta(\cdot \mid \cdot)$ as above and assume $p_v > 0$ exactly where $q_v > 0$ (equal supports), as for softmax policies. Write $r(v) = p_v/q_v$ for the density ratio (under R0, the ratio equation 2.1 at $y = v$), $\xi(v) = r(v) - 1$, and

$$d_\infty = \max_{v:\, q_v > 0} |r(v) - 1|, \qquad \chi^2(p\|q) = \mathbb{E}_q[\xi^2] = \mathrm{Var}_q(r),$$

the last equality because $\mathbb{E}_q r = 1$. Thus $\chi^2(p\|q)$ is the variance of the importance weight of a token sampled from the rollout policy—the moment of Proposition 3.2—and d_∞ is the largest multiplicative change of any single token.

Theorem A.6 (Quadratic expansion of the KL)**.** $D_{\mathrm{KL}}(q\|p) = \frac{1}{2}\chi^2(p\|q) + \mathcal{E}_2$*, where*

$$\mathcal{E}_2 = \mathbb{E}_q\Big[\xi^2 \int_0^1 (1-t)\Big(\frac{1}{(1+t\xi)^2} - 1\Big)\, dt\Big].$$

If $d_\infty \le d < 1$*, then*

$$|\mathcal{E}_2| \;\le\; \frac{2+d}{6(1-d)^2}\,\mathbb{E}_q|\xi|^3 \;\le\; \frac{d(2+d)}{6(1-d)^2}\,\chi^2(p\|q).$$

Proof. Since $\mathbb{E}_q r = 1$, $D_{\mathrm{KL}}(q\|p) = \mathbb{E}_q[-\log r] = \mathbb{E}_q[g(r)]$ with $g(r) = r - 1 - \log r$. As $g(1) = g'(1) = 0$ and $g''(r) = r^{-2}$, Taylor's theorem with integral remainder gives $g(1+\xi) = \xi^2 \int_0^1 (1-t)(1+t\xi)^{-2}\, dt$ for $\xi > -1$; subtracting $\frac{1}{2}\xi^2 = \xi^2 \int_0^1 (1-t)\, dt$ and taking $\mathbb{E}_q$ yields $\mathcal{E}_2$. If $|\xi| \le d < 1$, then $\big|(1+t\xi)^{-2} - 1\big| = t|\xi|\,|2+t\xi|/(1+t\xi)^2 \le t|\xi|(2+d)/(1-d)^2$, and $\int_0^1 t(1-t)\, dt = \frac{1}{6}$ gives the first bound; $|\xi|^3 \le d\,\xi^2$ gives the second. □

Proposition A.7 (Sharp local constants)**.** *Let* $\kappa(r) = (r - 1 - \log r)/(r-1)^2$ *for* $r \neq 1$ *and* $\kappa(1) = \frac{1}{2}$*, so that* $D_{\mathrm{KL}}(q\|p) = \mathbb{E}_q[\xi^2 \kappa(r)]$ *exactly. Then* κ *is strictly decreasing on* $(0, \infty)$*, and if* $d_\infty \le d < 1$*,*

$$\kappa(1+d)\,\chi^2(p\|q) \;\le\; D_{\mathrm{KL}}(q\|p) \;\le\; \kappa(1-d)\,\chi^2(p\|q).$$

Proof. For $r \neq 1$, $\kappa'(r) = h(r)/(r-1)^3$ with $h(r) = 2\log r - r + 1/r$. Since $h(1) = 0$ and $h'(r) = -(1 - 1/r)^2 \le 0$, h is positive on $(0,1)$ and negative on $(1,\infty)$, so $\kappa' < 0$ on both sides. The bounds follow because $r(v) \in [1-d, 1+d]$ for every v with $q_v > 0$. □

For $d = 0.2$ the constants are $\kappa(1.2) = 0.442$ and $\kappa(0.8) = 0.579$, so $D_{\mathrm{KL}}(q\|p)$ equals $\frac{1}{2}\chi^2(p\|q)$ to within $0.079\,\chi^2(p\|q)$; the cruder bound of Theorem A.6 gives 0.115. In this local regime the KL and the importance-weight variance are interchangeable up to these constants. Because both results hold for any pair with equal supports, they apply in particular to the $(k{+}1)$-bucket pair $(\bar{q}, \bar{p})$ that the gate evaluates, whose masses are all positive. Since κ is decreasing, the lower bound of Proposition A.7 needs only $r \le \lambda$: an admitted token, $D^{\mathrm{tkl}} \le \tau_i$, whose bucket-level ratios are all at most λ has a coarsened importance weight of variance at most $\tau_i/\kappa(\lambda)$—for $\tau^+{=}0.10$ and $\lambda = 2$ ($\kappa(2) = 1 - \log 2 = 0.307$), at most 0.326, a standard deviation of 0.57, whereas $\bar{r} \le 2$ alone permits variance up to $\lambda - 1 = 1$. By data processing the variance of the full weight can only be larger, and Proposition 3.2 shows it can be arbitrarily larger. The local condition is a statement about the logit update: if $\pi_{\mathrm{roll}} = \mathrm{softmax}(z)$ and $\pi_\theta = \mathrm{softmax}(z + \Delta)$, then $|\log r(v)| \le \max_w \Delta_w - \min_w \Delta_w$, so d_∞ is small exactly when the update shifts all logits by nearly the same amount.

Proposition A.8 (No global control)**.** *For every* $\varepsilon > 0$*: (i) for every* $\lambda > 1$ *there exist* q, p *with* $D_{\mathrm{KL}}(q\|p) < \varepsilon$ *and* $d_\infty = \lambda - 1$*; (ii) there exist* q, p *with* $D_{\mathrm{KL}}(q\|p) < \varepsilon$ *and arbitrarily large* $\chi^2(p\|q)$*.*

Proof. Take two symbols and $q = (s, 1-s)$. For (i) let $p = (\lambda s, 1 - \lambda s)$ with $0 < s < \min(\frac{1}{2}, \frac{1}{\lambda})$: then $r = \lambda$ on the first symbol, $d_\infty = \lambda - 1$, and $D_{\mathrm{KL}}(q\|p) = s\log\frac{1}{\lambda} + (1-s)\log\frac{1-s}{1-\lambda s} = s(\lambda - 1 - \log\lambda) + O(s^2) \to 0$ as $s \downarrow 0$. For (ii) let $p = (s^{1/3}, 1 - s^{1/3})$: then $\chi^2(p\|q) \ge (s^{1/3} - s)^2/s \to \infty$ while $D_{\mathrm{KL}}(q\|p) = s\log s^{2/3} + (1-s)\log\frac{1-s}{1-s^{1/3}} = O(s^{1/3}) \to 0$. □

The local equivalence therefore cannot be made global. Proposition 3.2 shows that the *coarsened* KL cannot bound the importance-weight variance; Proposition A.8 shows that even the *full* KL cannot, nor can it bound the ratio of any single token. The obstruction is a rare token whose probability increases by a large factor: it contributes little to the KL, but when it is sampled it carries a large importance weight. That is the event the sampled-ratio clamp is built for: whatever its raw weight, an admitted token's gradient coefficient is at most $C = \max\{\rho_{\mathrm{hi}}, c\} = 10$ (Proposition 3.2). Conversely, if $d_\infty \le \frac{1}{2}$ every ratio lies in $[\rho_{\mathrm{lo}}, \rho_{\mathrm{hi}}] = [0.5, 4.5]$ and the clamp is inactive; it binds only in the regime of Propositions 3.2 and A.8.

Table 4: Backbone and RL algorithm.

Setting	Value
Base model	Qwen3-30B-A3B-Base (Yang et al., 2025)
Architecture	Mixture-of-experts, 30B total / 3B active
Advantage estimator	GRPO, group-standardized rewards (Shao et al., 2024)
KL penalty	none (β=0: no reward KL, no auxiliary KL loss)
Entropy bonus	0
Training framework	verl (Sheng et al., 2025)
Training backend	Megatron (Shoeybi et al., 2019)
Rollout engine	vLLM (Kwon et al., 2023)

Table 5: VOKL-CLIP gate and clipping. The wide ratio clip bounds the importance weight; the gate radii $\tau^{\pm}$ select which tokens update (§2.2, Appendix A).

Setting	Value
Top-k bucket size k	16
Gate radius τ^{+} (advantage > 0)	0.10
Gate radius τ^{-} (advantage < 0)	0.05
Ratio clip band $[\rho_{\text{lo}}, \rho_{\text{hi}}]$	$[0.5,\ 4.5]$ (ϵ_{lo}=0.5, ϵ_{hi}=3.5 about 1)
Dual-clip constant c	10 (dual clip of Ye et al., 2020)
Tail floor δ	10^{-8}
Gate mechanism	stop-gradient indicator (`dapo_tkl_double_clip_ppo`)
Loss aggregation	token-mean

B EXPERIMENTAL AND IMPLEMENTATION DETAILS

This appendix records the exact configuration behind §4.1: the hyperparameters (Tables 4–6), the evaluation protocol, and the data and verifier pipeline. Every value is the one used for the VOKL-CLIP runs; the DAPO and GSPO baselines share all of these settings and differ only in the clip family under study and, where indicated, the mismatch-correction level R0–R3.

B.1 HYPERPARAMETERS

The top-k bucket is anchored to the rollout policy. The k token identities that define each bucket $\mathcal{V}^k_{i,t}$ are the top-k ids of the *rollout* (old) policy at that position—exactly the ids the inference engine returns alongside its log-probs. The current policy is then read on that *same* id set, so the top-k bucket KL $D^{\text{tkl}}_{i,t}$ compares both policies over a fixed, rollout-chosen support rather than over each policy's own shifting top-k. This is what makes the distance free (§2): no extra forward pass and no re-selection of indices on the training side.

B.2 EVALUATION PROTOCOL AND METRICS

Our headline metric is *ave@32*, the mean pass@1 over 32 independent samples. For a benchmark of N problems we draw n=32 completions per problem and report

$$\text{ave@32} = \frac{1}{N}\sum_{q=1}^{N}\frac{1}{32}\sum_{j=1}^{32}\mathbf{1}\left[\widehat{a}_{q,j} = a^{\star}_{q}\right],$$

the unbiased pass@1 estimator, rather than the sample-maximum ceiling best@32 (the fraction of problems solved by at least one of the 32 samples). We headline ave@32 because it separates methods more sharply and has lower run-to-run variance than the ceiling metric (Appendix C).

Evaluation decoding differs from training: at evaluation we sample at temperature 1.0 and top-p 0.7 (top-k disabled), whereas training rollouts use an untruncated top-p 1.0. We report six benchmarks: the competition-math sets AIME24, AIME25, AIME26, and AMC23 at n=32 (ave@32), and GPQA-Diamond (Rein et al., 2024) and MMLU-Pro (Wang et al., 2024) with LightEval (Habib et al., 2023)

Table 6: Optimization, batching, and training-time rollout sampling.

Setting	Value
Learning rate	1×10^{-6}
Warmup	10 steps
Weight decay	0.1
Optimizer	AdamW (default β_1=0.9, β_2=0.999)
Gradient-norm clip	1.0
Prompt batch size	512
Mini-batch (prompts)	32
Responses per prompt G	16
Max prompt length	2,048 tokens
Max response length	20,480 tokens
Dynamic batching	enabled
Rollout sampling (train)	temperature 1.0, top-p 1.0, top-k off
Total training steps	360
Eval / checkpoint cadence	every 10 steps
Overlong buffer / penalty factor	4,096 tokens / 1.0

at n=8. AIME24 ave@32 is the primary metric, logged every 10 steps on a fixed 0–360-step grid; curves are stitched across resumed runs with gaps shown as breaks, and no table value is interpolated.

Checkpoint selection. Each table row reports a single checkpoint: the step that maximizes AIME24 ave@32 within 0–360, with every benchmark in the row read at that step. For tables with offline-evaluated columns (Tables 1, 2 and 3) the maximum is taken over the checkpoints evaluated offline; Table 8, which reports only the two AIME curves, takes it over all logged steps. Table 7 instead reports each metric at its own peak step.

B.3 Training data, prompt, and verifier

Training uses **DAPO-Math-17k** (Yu et al., 2025) (data source `math_dapo`, ∼17k competition-math problems, ability MATH); validation during training is on the AIME2024 and AIME2025 problem sets. Each problem is presented as a single user turn with no system prompt, using the template below (`{question}` is replaced by the problem text; the model's own chat template is applied on top at generation time).

```
Solve the following math problem step by step.  The last line
of your response should be of the form Answer:  $Answer (without
quotes) where $Answer is the answer to the problem.
{question}
Remember to put your answer on its own line after "Answer:".
```

Rule-based verifier. Rewards come from the `dapo` reward manager, which matches normalized final answers as in MATH (Hendrycks et al., 2021). For a completion we (i) keep its last 300 characters, (ii) extract the answer with the case-insensitive pattern `Answer\s*:\s*([^\n]+)`, taking the last match, (iii) canonicalize it with a LaTeX/unit normalizer, and (iv) test it for exact string equality against the normalized ground truth (Minerva's answer normalizer (Lewkowycz et al., 2022) without its SymPy-equivalence check; no strict-box requirement). The reward is $+1$ for a correct answer and -1 otherwise.

Overlong soft penalty. On top of the ± 1 correctness reward we add the DAPO length penalty (Yu et al., 2025): with maximum response length L=20,480 and buffer B=4,096, a completion of length ℓ receives an additive penalty

$$r_{\text{len}}(\ell) = \min\Big(-\tfrac{\ell-(L-B)}{B},\ 0\Big),$$

i.e. zero until ℓ exceeds $L - B = 16{,}384$ tokens, then decreasing linearly to -1 at the full length L.

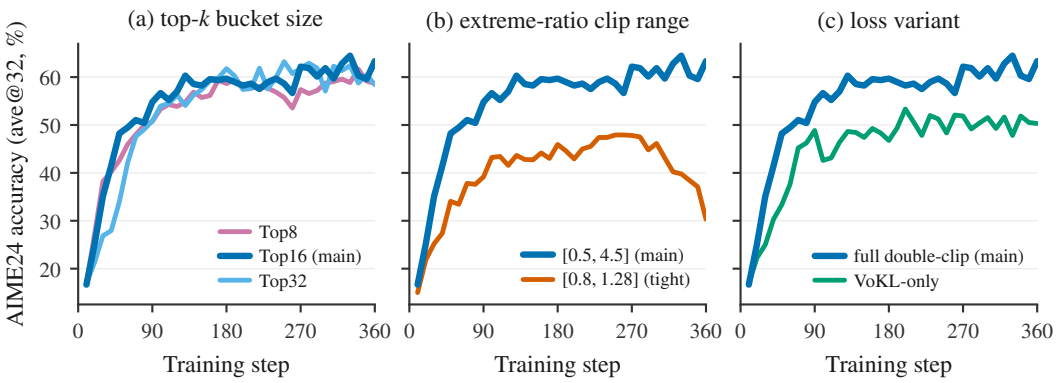


Figure 4: Ablation training curves (AIME24 ave@32, R0). (a) top-k bucket size, (b) extreme-ratio clip range (the tight $[0.8, 1.28]$ variant destabilizes and collapses), (c) loss variant (removing the extreme-ratio clamp hurts throughout).

Table 7: Peak AIME24/AIME25 over 0–360 for the R0 methods, reported as ave@32 (the mean pass@1 over the 32 samples) at the step where each metric peaks. We report ave@32 rather than the sample-maximum best@32 because the mean pass@1 separates methods more sharply and has lower run-to-run variance than a ceiling metric, which compresses differences once accuracy approaches saturation. DAPO here is its strongest R0 run, the wide-clip $[0.5, 4.5]$ variant that adopts VoKL-Clip's clamp; the standard-clip baseline sweep is in Table 8.

	ave@32	
Method	AIME24	AIME25
VoKL-Clip (ours)	64.5	44.4
DAPO ($[0.5, 4.5]$)	57.1	43.8
GSPO	40.8	30.1

C Additional Experiments and Explorations

This appendix collects results that were moved out of the main text for space: peak metrics across the full trajectory, the exploratory baseline configurations we ran while calibrating the comparison, the sensitivity of the ablation grid, and the provenance of the reported numbers. None of these change the conclusions of the main text; they document the surrounding evidence.

C.1 Ablation training curves

Figure 4 plots the AIME24 trajectories behind Table 3. The table reports each row at its designated checkpoint; the curves show that those readings are not isolated peaks—the top-k variants track one another after the first $\sim$70 steps (Top32 starts slower), while the tight-clamp and no-clamp variants sit below VoKL-Clip for the whole run.

C.2 Peak ave@32 across steps

Table 7 reports each R0 method's peak AIME24/AIME25 over the full 0–360-step trajectory as ave@32, the mean pass@1 over the 32 samples, taken at the step where each metric peaks. We report the mean pass@1 rather than the sample-maximum best@32 because averaging over the 32 samples separates methods more sharply and has lower run-to-run variance than a ceiling metric, which compresses differences once accuracy approaches saturation. VoKL-Clip attains the highest peak on both benchmarks.

Table 8: Effect of the rollout↔training mismatch correction level on Qwen3-30B-A3B, for the standard-clip baselines (DAPO $[0.8, 1.28]$, GSPO). Both columns at each level are ave@32 read at the single step that maximizes AIME24 within 0–360, so each pair describes one checkpoint. At the cheap R0 (rollout log-prob) setting both baselines are weak; they need the expensive corrections—recompute (R1) or router replay (R2/R3)—to recover, yet no entry reaches VOKL-CLIP's 64.5 AIME24 / 43.2 AIME25 at its native R0 (Table 1); on AIME25 the closest entry, DAPO at R2, is within run-to-run variation of it. †Read before that run diverges (Appendix C.6). DAPO with VOKL-CLIP's wide clip $[0.5, 4.5]$ reaches 57.1 (AIME24) at R0 (a non-standard variant that adopts our clamp), still below VOKL-CLIP's 64.5.

	R0		R1		R2		R3	
Method	AIME24	AIME25	AIME24	AIME25	AIME24	AIME25	AIME24	AIME25
DAPO	46.9	32.4	54.4	40.3	61.0†	42.5	58.1	39.2
GSPO	40.8	29.1	57.2	41.5	50.0	34.8	57.9	38.6

Table 9: Exploratory baseline configurations that did not enter the main tables, reported as the peak AIME24 ave@32 over 0–360. These are *additional* knobs we tried on the standard-clip baselines, beyond the per-level sweep in Table 8 (which lists the strongest configuration we found at each correction level: DAPO 58.1 and GSPO 57.9 at R3). None of the extra configurations closes the gap to VOKL-CLIP, whose native-R0 result is repeated here for reference.

Configuration	Level	AIME24 ave@32
VOKL-CLIP (ours)	R0	**64.5**
GSPO, default sequence clip	R3	52.1

C.3 BASELINE MISMATCH-CORRECTION SWEEP (R0–R3)

Table 8 gives the full per-level sweep behind the summary in §4.4: both standard-clip baselines (DAPO $[0.8, 1.28]$, GSPO) at every mismatch-correction level R0–R3, each entry read at the step that maximizes AIME24. VOKL-CLIP at its native R0 reaches 64.5 AIME24 and 43.2 AIME25 (Table 1), above every entry in the sweep. The strongest AIME24 entry is DAPO's R2 at 61.0, from the run that diverges by step 200 (§4.4); the strongest that trains stably is DAPO's R3 at 58.1.

How each baseline's level is chosen. The *overall best-performing* level used in Table 1 is selected in two stages: we first drop any level whose run fails to train stably through the reporting window, then among the rest take the level that wins the most of the six benchmarks at its own $s^\star$. For DAPO this drops R2—it leads on four of the six benchmarks but diverges at step 200—and R3 then wins four of six against R1 and all six against R0. For GSPO no level is dropped, and R1 wins four of six against R3, five of six against R2 and all six against R0; R3's slightly higher AIME24 peak (57.9 vs. 57.2) is offset by AIME25, AIME26, GPQA-Diamond and MMLU-Pro. Selecting on AIME24 alone would instead pick DAPO's R2 and GSPO's R3, both of which score lower across the suite as a whole.

C.4 EXPLORATORY BASELINE CONFIGURATIONS

The per-level comparison in Table 8 reports, for each baseline, the strongest configuration we found at each correction level. Along the way we tried a further knob on the standard-clip baselines that did not make the main tables; we report it in Table 9 for completeness. Moving DAPO's tight clip $[0.8, 1.28]$ to the strongest correction level R3—the configuration reported in Table 8—lifts it from its weak R0 result to 58.1. For GSPO at R3, keeping the default sequence-clip boundaries reaches only 52.1, below the 57.9 obtained with the boundaries retuned (Table 8). Across all of these, the best releasable baseline configuration reaches 58.1, still 6.4 points below VOKL-CLIP's 64.5 at the cheap R0 setting: the extra corrections and tuning narrow the gap but do not close it.

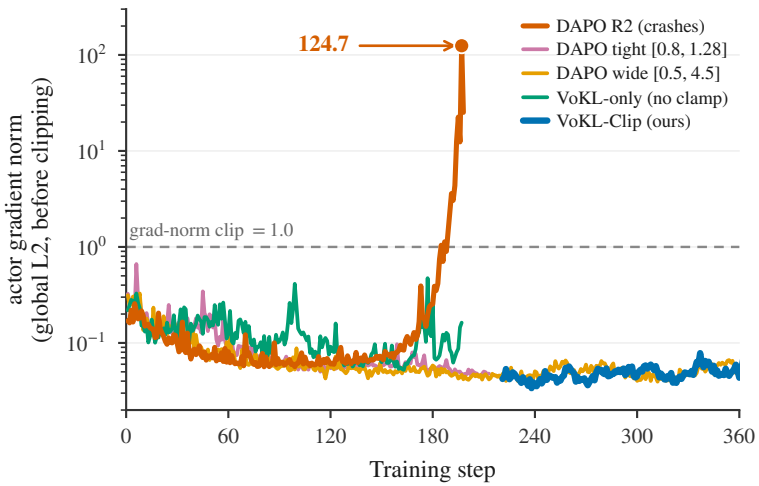


Figure 5: Actor gradient norm (log scale) over training, from the per-step optimizer traces. The DAPO run at the R2 correction level spikes to 124.7 as its rollout↔training KL (not plotted) rises to 69.9 and then dies—exceeding the 1.0 gradient-norm clip on 12 steps. Every run that does not diverge stays under that clip at every logged step, the largest excursion being 0.664 (tight-clamp DAPO); VoKL-Clip's own maximum over its resumed s=221–360 segment is 0.078, comparable to the wide-clamp DAPO run over the same steps (0.066). Each curve is drawn over its own observed steps.

C.5 Sensitivity of the ablation grid

The ablations in Table 3 sweep only a handful of values per axis, and VoKL-Clip is not finely tuned to any of them. The bucket width is swept over $k \in \{8, 16, 32\}$ with AIME24 peaks of 61.7/64.5/63.2: the default k=16 is a round power-of-two chosen a priori, and the spread across the three settings is under three points, so the method is insensitive to the exact bucket size rather than tuned to it. The gate radii, the only hyperparameters the gate adds besides k, are not finely tuned either: doubling both radii to τ^-=0.10, τ^+=0.20 moves the AIME24 peak from 64.5 to 60.9, a 3.6-point drop across a 2× change in the gate width. The default τ^-=0.05, τ^+=0.10 was fixed a priori from the desired asymmetric gate scale rather than selected on this comparison; §3 reports its assumption-free total-variation interpretation. The only setting that materially hurts is the tight ratio clip $[0.8, 1.28]$, which collapses training, and removing the extreme-ratio clamp entirely (VoKL-only, 53.3), which costs more than ten points—both of which we highlight in the main text. We did not tune beyond this minimal grid.

C.6 Gradient-norm stability

The rollout↔training mismatch that collapses DAPO's accuracy at the R2 correction level is visible in the optimizer before it ever reaches the metrics. Figure 5 plots the actor gradient norm across training for the DAPO-R2 run and for VoKL-Clip and the ablation variants. The DAPO-R2 run sits in the same stable ∼0.1 band as the others for most of training and then, as its rollout↔training KL (not plotted) rises to 69.9, its gradient norm spikes to 124.7 within a few steps and the run dies. Every run that does not diverge—including the no-clamp VoKL-only variant—instead keeps its gradient norm below 0.7 at every logged step, and VoKL-Clip's own stays below 0.08 over s=221–360, so the updates never leave the region the gradient-norm clip (1.0) is meant to protect. This is the optimizer-level counterpart of Corollary A.3 and Proposition 3.2: the gate certifies local distributional change while the wide clamp bounds the effective importance coefficient. These results alone do not bound the parameter-space gradient without a network-Jacobian or global gradient-clip condition; the observed traces show that the combined mechanism remains inside that global clip in practice. In the figure VoKL-Clip's own trace is the resumed s=221–360 segment; over the earlier steps the other non-diverging runs stay well below the clip (maximum 0.664).

C.7 PROVENANCE OF REPORTED NUMBERS

Every number we report traces to the exported W&B training curves or to the offline-evaluation logs of the corresponding checkpoint. Figure 1a is the one panel not drawn from those aggregate curves: it is rendered from a saved per-token snapshot of a single training batch, cropped to the region that contains the data (the single point below the crop, at $\log r = -6.4$, is marked at the lower edge). Figure 1b is a constructed two-distribution example whose values are computed in closed form.