Generation Entropy as a Multi-Granular Training Signal for Mathematical Reasoning

Anonymous Author(s)

Abstract

Group Relative Policy Optimization (GRPO) collapses two qualitatively different reasoning behaviors—confident correct answers and lucky guesses—into identical training signals. We identify per-token generation entropy as an effective proxy for this missing confidence dimension: low entropy reflects committed decisions while high entropy signals active exploration. We introduce MGE-PO, an entropy-centric training framework that uses this signal across three orthogonal yet synergistic dimensions: (1) modulating policy advantages to reinforce confident correct reasoning and penalize uncertain errors, (2) adapting per-token KL regularization to anchor low-entropy tokens while relaxing exploration constraints, and (3) prioritizing training samples within a dynamic entropy-based curriculum focused on the model's zone of proximal development. On Qwen2.5-7B-Base, MGE-PO improves over standard GRPO by $+3.9$pp in average $\mathrm{avg@32}$ and $+4.2$pp in average $\mathrm{pass@32}$ across four math benchmarks, with consistent $\mathrm{pass@32}$ transfer to Qwen2.5-Math-1.5B and Qwen2.5-Math-7B. Analyses confirm that these mechanisms are orthogonal and complementary, jointly refining partially correct reasoning while mitigating entropy collapse and length inflation.

1 Introduction

Reinforcement learning from verifiable rewards has become a central paradigm for improving the reasoning capabilities of large language models. Group Relative Policy Optimization (GRPO) (DeepSeek-AI, 2025) and its successor DAPO (Yu et al., 2025) achieve strong mathematical reasoning by computing group-relative advantages without a value network. However, these methods treat completions with identical rewards uniformly, collapsing confident correct reasoning and lucky guesses into identical training signals. This is a particularly costly omission in mathematical reasoning, where incorrect trajectories are often coherent and differ from correct ones mainly at uncertainty points.

We argue that per-token generation entropy provides an effective online proxy for this missing confidence dimension. Language models are meaningfully calibrated (Kadavath et al., 2022), and recent entropy-minimization RL (Prabhudesai et al., 2025) confirms that internal probability distributions serve as reliable training signals: low entropy reflects familiar reasoning regimes while high entropy signals difficult intermediate states where multiple continuations compete.

Motivated by these observations, we propose MGE-PO, a plug-in framework that injects an entropy-derived signal across three complementary components:

1. **EA-Adv: Entropy-aware advantage injection.** Per-token entropy is aggregated and EMA-normalized into a confidence signal; confident correct completions receive a reinforcement bonus while uncertain errors are penalized. High-entropy regions revert to standard GRPO, yielding a compact four-quadrant rule.
2. **EA-KL: Entropy-aware KL regularization.** The per-token KL penalty is reweighted by relative entropy—low-entropy tokens are anchored more strongly to the reference policy, while high-entropy tokens are relaxed, decoupling consolidation from exploration.
3. **DynZPD: Dynamic curriculum sampling.**

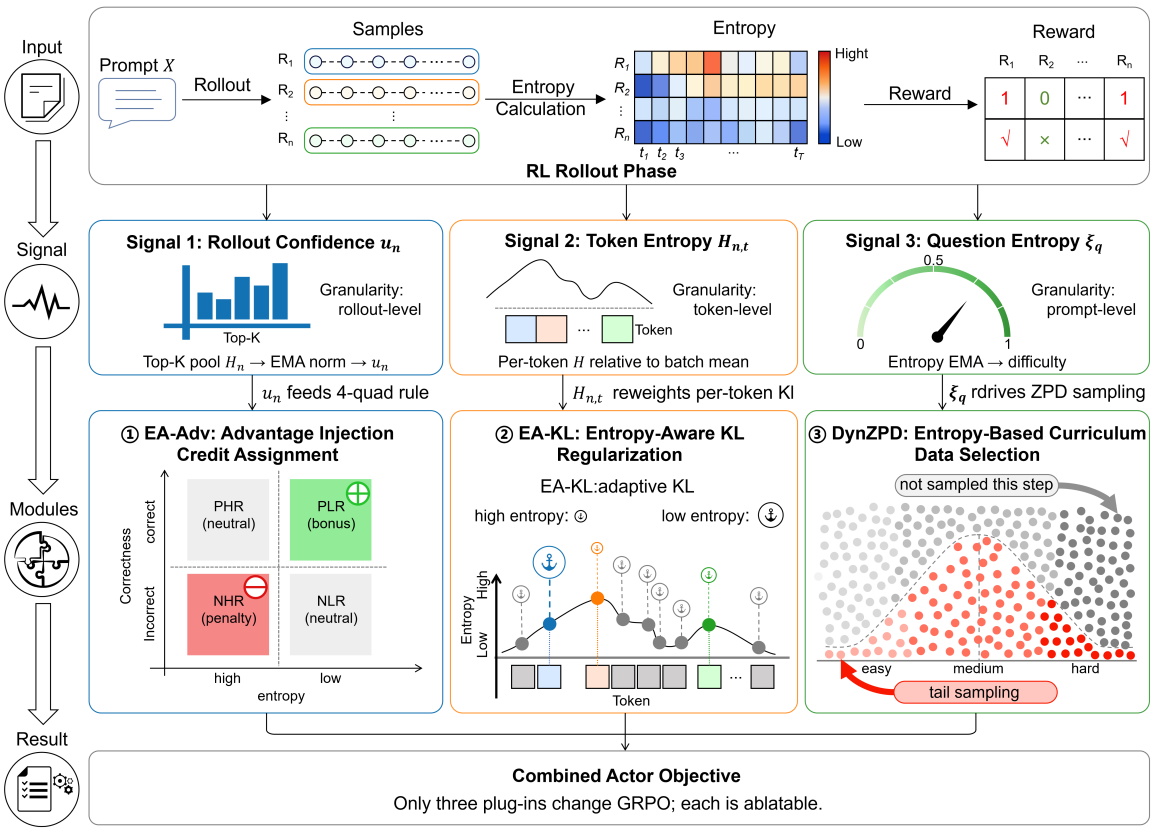

Figure 1: **MGE-PO framework.** MGE-PO reuses cached token-level entropy as a unified online signal for three plug-ins: EA-Adv adjusts rollout-level credit by activating confident-correct bonuses and uncertain-incorrect penalties; EA-KL converts uniform KL regularization into entropy-adaptive token anchoring; and DynZPD samples questions from a bell-shaped difficulty distribution that emphasizes medium-difficulty examples while preserving tail exploration. Each plug-in modifies a distinct stage of GRPO and can be ablated independently.

We replace uniform question selection with a Gaussian sampling weight over EMA-tracked question difficulty ξ_q, defined as the per-question mean token entropy. Trivially easy questions (low entropy, $\xi_q \to 0$) and intractable questions (high entropy, $\xi_q \gg 0$) decay toward a uniform sampling floor, concentrating training on the model's zone of proximal development.

A diagnostic of model responses on mathematical datasets reveals that low-entropy incorrect rollouts (NLR quadrant) account for 45.5% of base-model outputs (Table 5), confirming confident errors as a dominant failure mode. We therefore identify generation entropy as an online proxy for reasoning difficulty and integrate it into a three-component GRPO framework spanning advantage injection, KL regularization, and dynamic data selection. Empirically, on Qwen2.5-7B base trained on DAPO-Math, MGE-PO improves pass@32 over standard GRPO by +9.7pp on AIME 2024 (30.3 → 40.0) and +7.0pp on AIME 2025 (23.3 → 30.3); ablations show the full stack is complementary and non-redundant, improving by +3.9pp versus +1.8pp for EA-Adv alone in average avg@32.

2 Related Work

GRPO and its variants. RLHF with PPO (Schulman et al., 2017; Ouyang et al., 2022) requires a separate value network, which GRPO (DeepSeek-AI, 2025) removes by using group-relative rewards. Subsequent variants further refine a *single* pipeline stage: DAPO (Yu et al., 2025) adds clip-higher and dynamic sampling; XRPO (Bamba et al., 2025), SEED-GRPO (Chen et al., 2025), and MAPO (Huang et al., 2025) modify advantage shaping; ADRPO (Fan et al., 2025), AEPO (Wang et al., 2025a), and KL-Cov (Cui et al., 2025) adapt the KL coefficient; and DUMP (Wang et al., 2025c) and CDAS (Kong et al., 2026) adjust the data curriculum. The closest comparison to EA-KL is KL-Cov (Cui et al., 2025), which uses an entropy–

covariance criterion to *single out* high-covariance tokens for the KL penalty so they cannot collapse policy entropy. By contrast, EA-KL keeps every token in the KL term but *reweights* them by relative entropy (Eq. 8): sharp (low-entropy) tokens are anchored *more* strongly to π_{ref} to prevent confident-error collapse, with diffuse tokens left relaxed. MGE-PO additionally reuses the same online signal—token-level generation entropy—across all three pipeline stages simultaneously.

Entropy as a training signal. LLMs are well calibrated (Kadavath et al., 2022; Xiong et al., 2024), making token-level entropy a reliable confidence proxy. Traditional RL adds a uniform entropy bonus to encourage exploration (Schulman et al., 2017; Williams, 1992; Mnih et al., 2016), but Wang et al. (Wang et al., 2025b) show that this is suboptimal for LLM CoTs: only a high-entropy minority of tokens act as decision "forks", and training on just the top-20% can match or surpass full-token DAPO. Related work identifies "critical tokens" for KL relaxation or contrastive estimation (Vassoyan et al., 2025; Lin et al., 2025), minimizes entropy to improve reasoning (Prabhudesai et al., 2025), or aggregates semantic entropy at the question level (Chen et al., 2025). Our work is closest to Wang et al. (Wang et al., 2025b) but differs in two ways: (i) we do *not* discard low-entropy tokens, but exploit the asymmetry between confident-correct and confident-incorrect rollouts at the *rollout* level (versus EAPO's (He et al., 2026) concurrent *token*-level analysis); (ii) we treat entropy as a unified signal across rollout, token, and question granularities.

3 Preliminaries

Given questions $\{x^{(g)}\}_{g=1}^{G}$, GRPO (DeepSeek-AI, 2025) samples n completions per question from policy π_θ, receives a binary rule-based reward $r_i \in \{0, 1\}$, and estimates advantage without a critic via group-relative normalization:

$$A_i^{\text{GRPO}} = \frac{r_i - \mu_{g(i)}}{\sigma_{g(i)} + \varepsilon}, \tag{1}$$

where μ_g and σ_g are the group mean and standard deviation. The GRPO objective combines a clipped surrogate with a KL penalty from reference policy π_{ref}:

$$\mathcal{L}_{\text{GRPO}} = \frac{1}{B} \sum_{i,t} \Big[\mathcal{L}_{i,t}^{\text{clip}}(A_i^{\text{GRPO}}) - \beta\, D_{\text{KL}}(\pi_\theta \| \pi_{\text{ref}})_{i,t} \Big]. \tag{2}$$

Token-level generation entropy. At each position t during rollout, the policy defines a distribution over the vocabulary. We record the Shannon entropy of this distribution:

$$H_{i,t} = -\sum_{v \in \mathcal{V}} \pi_\theta(v \mid c_{i,t}) \log \pi_\theta(v \mid c_{i,t}), \tag{3}$$

where $c_{i,t} = (x^{(g(i))}, a_{i,<t})$ is the prefix context. Low $H_{i,t}$ means the model is confident about its next token; high $H_{i,t}$ indicates active exploration among competing continuations. These values are computed once during rollout generation and cached, with zero recomputation cost during the policy update. We use this entropy at three granularities in Section 4: rollout-level for advantage modulation, token-level for KL reweighting, and question-level for curriculum sampling.

4 Method

Motivation. Standard GRPO is purely outcome-driven and collapses two qualitatively different policy states, confident commitment and active exploration, into identical training signals. Token-level generation entropy $H_{i,t}$ (Eq. 3) provides an online measurement of this missing dimension: low H reflects a committed decision while high H reflects active exploration among competing continuations.

Our paradigm: entropy as a unified signal. Rather than stacking three independent plug-ins, MGE-PO proposes a single *entropy-as-unified-signal* paradigm that threads $H_{i,t}$ through the three orthogonal axes on which any policy-gradient algorithm operates: *advantage*—what to reinforce (Section 4.2); *KL*—how far to drift (Section 4.3); and *data*—what to learn from next (Section 4.4). All $H_{i,t}$ are computed once during rollout generation and cached, never recomputed during the policy update. The three mechanisms act on distinct target objects (rollout, token, question) and are separately ablatable (Section 5.4); their synergy emerges precisely because they share the same online signal (Table 1).

4.1 Token-Entropy Signal

For rollout-level decisions we pool token entropies by *value*—the K_i tokens with the largest $H_{i,\cdot}$, not the last K_i positions—because the most-uncertain steps in a long CoT trajectory cluster around rea-

soning branch points, not the final answer:

$$\bar{H}_i = \frac{1}{K_i} \sum_{t \in \text{Top-}K_i(H_{i,\cdot})} H_{i,t}, \qquad K_i = \text{clamp}\big(\lfloor r_K L_i \rfloor,\ 1,\ K_{\max}\big), \tag{4}$$

with $r_K = 0.1$ and $K_{\max} = 4096$ (see Appendix D). Section 5.4 compares Top-K, last-K, and a global mean. The correctness indicator q_i is defined on the raw grader output and the cap-aware gate of Eq. (21) handles length-induced false negatives (Appendix B).

4.2 EA-Adv: Entropy-Aware Advantage Injection

To exploit the online token-entropy signal for fine-grained credit assignment, we augment the base GRPO advantage A_i^{GRPO} by classifying each rollout into four quadrants based on its correctness $q_i \in \{0, 1\}$ and a logistic confidence score $u_i \in [0, 1]$. Under our default asymmetric configuration, the entropy-aware adjusted advantage A_i^{EA} is compactly defined as:

$$A_i^{\text{EA}} = A_i^{\text{GRPO}} + g_i \big[q_i \gamma_{\text{PLR}}\, u_i - (1 - q_i) \gamma_{\text{NHR}} (1 - u_i) \big], \tag{5}$$

where $g_i = 1 - \mathbf{1}[o_i \text{ truncated}]$ acts as a safeguard gate that zeroes out modifications on truncated rollouts, and $(\gamma_{\text{PLR}}, \gamma_{\text{NHR}}) = (0.5, 0.3)$ are the default hyperparameters. The confidence score $u_i \in [0, 1]$ is obtained by standardizing $\bar{H}_i$ against running batch statistics (μ_H, τ_H) and applying a logistic mapping $u_i = 1 - \sigma(z_i)$, so low entropy yields $u_i \to 1$ and high entropy yields $u_i \to 0$; full EMA tracking and batch re-centering are in Appendix B (Eq. 11).

Design Rationale. The asymmetric activation in Eq. (5)—visualized in Figure 2—encodes a directional bias on the confidence–correctness plane: only confidence-aligned outcomes participate in learning, while ambiguous quadrants are deliberately silenced.

- **PLR ($q_i = 1$, high confidence): consolidating correct basins.** GRPO's binary reward treats every correct rollout identically, regardless of whether the model committed to its reasoning or guessed. Under the basin-deepening view (Section 5.2), only confidence-aligned correct trajectories should be reinforced: amplifying lucky-correct (high-entropy) rollouts broadens the basin in the *wrong* direction by teaching the policy to treat unstable paths as reinforcement targets. The bonus $+\gamma_{\text{PLR}}\, u_i$ therefore scales *proportionally* with confidence, so the strongest reinforcement signal lands on the rollouts the model itself was committed to—exactly the trajectories that densify the existing correct-reasoning basin.

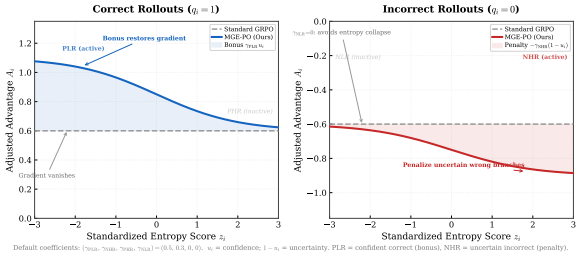

Figure 2: **Entropy-aware advantage injection.** MGE-PO augments GRPO with confidence-dependent shaping: confident correct rollouts receive a PLR bonus $\gamma_{\text{PLR}}\, u_i$ that scales *proportionally* with confidence, while uncertain incorrect rollouts receive an NHR penalty $-\gamma_{\text{NHR}}(1 - u_i)$ that scales *proportionally* with uncertainty. Here, u_i denotes confidence and $1 - u_i$ denotes uncertainty. Lucky-correct (PHR) and confident-incorrect (NLR) quadrants are inactive by default to avoid reinforcing noisy successes or triggering distribution collapse.

- **NLR ($\pi_\theta(a_{\text{wrong}}) \to 1$, $q_i = 0$): distribution collapse.** The GRPO advantage is already maximally negative on a confident-incorrect token; pushing π_θ further toward the corner of the simplex incurs an antagonistic pair of limits:

$$\underbrace{\mathcal{H}(\pi_\theta)}_{\text{entropy reservoir}} \xrightarrow{\pi_\theta(a_{\text{wrong}}) \to 1} 0^+, \qquad \underbrace{\left.\frac{\partial\, \text{KL}(\pi_\theta \,\|\, \pi_{\text{ref}})}{\partial\, \pi_\theta(a_{\text{wrong}})}\right|}_{\text{marginal deviation cost}} \xrightarrow{\pi_\theta(a_{\text{wrong}}) \to 1} \infty. \tag{6}$$

A vanishing entropy reservoir under a divergent *marginal* deviation cost tears the local semantic manifold: an infinitesimal further push toward the simplex corner demands an unbounded KL budget, formalizing the trust-region collapse argued in Appendix B.3. Recent evidence supports this stance: Wang et al. (Wang et al., 2025b) show that in RLVR the ∼20% high-entropy "forking tokens" carry essentially all of the reasoning signal, while the ∼80% low-entropy tokens remain stable throughout training, so disturbing the latter yields little informational gain at high collapse cost. We accordingly

set $\gamma_{\text{NLR}} = 0$ and delegate gentle restoration of exploration to EA-KL (Section 4.3).

4.3 EA-KL: Entropy-Aware KL Regularization

The unified entropy signal next acts on the *KL axis* at *token granularity*: where EA-Adv tells the policy what to reinforce, EA-KL controls how far each token may drift from the reference. Standard token-level KL regularization uniformly penalizes deviations from π_{ref}, ignoring that low-entropy (sharp) and high-entropy (diffuse) tokens are in fundamentally different optimization regimes. We therefore replace the uniform KL with an entropy-adaptive one that scales the penalty by local policy sharpness:

$$\mathcal{L}^{\text{EA-KL}} = \beta_{\text{KL}} \frac{\sum_{i,t} \text{mask}_{i,t} w_{i,t} \text{KL}_{i,t}}{\sum_{i,t} \text{mask}_{i,t}}, \quad (7)$$

where $\text{KL}_{i,t}$ is the low-variance token KL estimator, and $w_{i,t}$ is a dynamic reweighting factor computed as:

$$w_{i,t} = \text{clip}\left(1 + \alpha\left(1 - \frac{H_{i,t}}{\bar{H}}\right),\ 0,\ w_{\max}\right). \quad (8)$$

Here, $\bar{H}$ denotes the mask-aware mean token entropy within the current batch, and $\alpha = 0.2$ governs the scaling intensity. Sharper-than-average tokens ($H_{i,t} < \bar{H}$) receive an amplified anchor ($w_{i,t} > 1$) to suppress further over-compression and logit explosion; diffuse tokens ($H_{i,t} > \bar{H}$) enjoy a relaxed penalty ($w_{i,t} < 1$) to preserve the necessary exploratory budget. The detailed bounded constraints and low-variance formulations are deferred to Appendix C.1.

Division of labor with EA-Adv on the NLR quadrant. EA-KL completes the asymmetric design from Section 4.2: instead of suppressing confident errors with extra advantage penalties, which would trigger distribution collapse, we anchor them more strongly to π_{ref} through $w_{i,t} > 1$, letting the reference policy gently restore exploration on already-sharp distributions. EA-Adv and EA-KL are thus not redundant penalties but two complementary outlets of the same entropy signal: one decides *which direction* the policy should move on a rollout (advantage), the other decides *how much each token is allowed to move* (KL).

4.4 DynZPD: Dynamic Curriculum Sampling

The same token-level entropy signal, when aggregated to the question level, becomes a label-free difficulty indicator ξ_q that requires no external estimator: easy questions are solved with confident, low-entropy completions; intractable questions persist in chaotic, high-entropy generations; questions on the model's learning frontier sit at intermediate entropy. The unified entropy signal therefore opens its third outlet on the *data axis* at *question granularity*: gradients are most informative on questions inside the model's "Zone of Proximal Development" (ZPD)—neither already mastered nor entirely intractable. We model the sampling weight $w_q^{(t)}$ for question q at step t as a Gaussian centered at a target difficulty μ_{ZPD}:

$$w_q^{(t)} = \eta + (1-\eta) \exp\left(-\frac{(\xi_q^{(t)} - \mu_{\text{ZPD}})^2}{2\sigma_{\text{ZPD}}^2}\right), \quad (9)$$

where $\xi_q^{(t)}$ is the EMA of per-question mean token entropy, $\sigma_{\text{ZPD}} = 0.15$ controls the bandwidth, and $\eta = 0.10$ is a uniform exploration floor. Easy and intractable questions decay to η, concentrating updates on the learning frontier (tracking and initialization in Appendix C.2).

4.5 Training Objective

The actor objective combines the clipped policy term with EA-KL, sampled from the curriculum question distribution:

$$\mathcal{L}_{\text{actor}} = \mathbb{E}_{q \sim \{w_q^{(t)}\}}\Big[\mathcal{L}^{\text{clip}}(\widetilde{A}) - \mathcal{L}^{\text{EA-KL}}\Big], \quad (10)$$

where $\mathcal{L}^{\text{clip}}(\widetilde{A})$ denotes the clipped policy-gradient surrogate. The three entropy-aware mechanisms—advantage injection, KL reweighting, and curriculum sampling—are the only changes relative to GRPO. Each is separately ablatable: $\gamma_* = 0$ recovers standard advantages, $\alpha = 0$ recovers uniform KL weight, $\sigma_{\text{ZPD}} \to \infty$ recovers uniform question sampling. All three share the *same cached online signal* $H_{i,t}$, so their complementarity in Section 5.4 arises not from stacked independent components but from a single confidence dimension acting on three orthogonal axes of policy optimization: advantage, KL, and data.

5 Experiments

Our experiments are organized around four questions: (Q1) Does the full MGE-PO framework outperform GRPO and DAPO under matched conditions? (Q2) Does each mechanism contribute a non-redundant gain? (Q3) Are the predicted coefficient

Table 1: **Where each training component acts.** The three entropy-aware components consume a single low-level signal (token entropy $H_{i,t}$, or its question-level entropy aggregation ξ_q in DynZPD) at three different aggregation granularities.

Component	Stage	Modifies	Signal
EA-adv. (Eq. (5))	Credit	Advantage	$\bar{H}_i$
EA-KL (Eq. (8))	Reg.	KL weight	$H_{i,t}$
DynZPD (Eq. (9))	Data	Question prob.	ξ_q

sign patterns empirically necessary? (Q4) Is generation entropy a valid training signal, and where do the gains concentrate?

5.1 Experimental Setup

We evaluate **MGE-PO** on three Qwen2.5 models—**Qwen2.5-7B-Base** and **Qwen2.5-Math-1.5B/7B-Base** (Yang et al., 2024)—trained on the **DAPO-Math** dataset (Yu et al., 2025) (17,398 questions) without external filtering. Evaluation covers four benchmarks of varying difficulty: in-distribution **MATH-500** (Hendrycks et al., 2021), medium-difficulty **AMC 2023**, and high-difficulty OOD **AIME 2024 / 2025**. We report avg@32 and pass@32 with $T = 1.0$, $\text{top}_p = 0.7$, max length 12k, and 32 samples per problem.

Training follows the standard GRPO protocol for 8 epochs with a global batch size of 128 (full implementation details in Appendix D). Token-level entropy is computed once during rollout and cached. We compare against **Base** (zero-shot CoT), **GRPO** ($\beta_{\text{KL}} = 0.04$), **DAPO**, and **KL-Cov** (Cui et al., 2025) (entropy-covariance-aware token-level KL adjustment), and ablate three components: **EA-Adv**, **EA-KL**, and **DynZPD**.

5.2 Main Results

Principal evaluation (Q1). On **Qwen2.5-7B-Base** (Table 2), MGE-PO improves AIME 2024 pass@32 by +9.7pp and AIME 2025 by +7.0pp over GRPO while matching MATH-500 in distribution. On **Qwen2.5-Math-7B-Base** the gains widen further (+13.3pp / +7.9pp on AIME 2024 / 2025 pass@32, and a striking +2.5pp / +13.3pp on AIME 2024 avg@32 / pass@32), and DAPO underperforms on AMC 2023 (87.5% pass@32 vs. GRPO's 92.5%), suggesting that removing the KL penalty entirely is less stable on out-of-distribution evaluation than the entropy-adaptive KL anchoring used by MGE-PO.

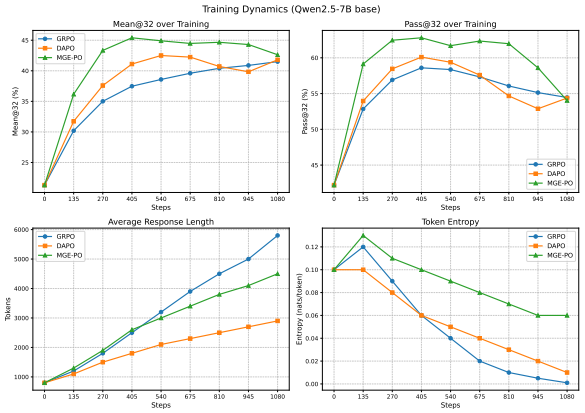

Figure 3: **Training and validation dynamics** on Qwen2.5-7B base over 8 epochs. MGE-PO achieves higher peak accuracy while mitigating the severe entropy collapse and length inflation of standard GRPO.

From lucky hits to dense correctness: the AMC 2023 signature. On AMC 2023 the two metrics appear to move in opposite directions: avg@32 rises from 58.7% to **65.5%** (+6.8pp) while pass@32 slightly retreats (92.5% → 90.0%). Both numbers are consistent with a single underlying policy transformation. GRPO already finds a correct trajectory in 92.5% of 32-sample bundles, so the *breadth* of the correct-reasoning basin is near-saturated; what entropy-guided credit assignment further changes is its *depth*: confident-correct (PLR) rollouts are reinforced and uncertain-incorrect (NHR) rollouts are pruned, so any individual sample, not just the best-of-32 champion, becomes more likely to land on a correct path, lifting avg@32. Reading the negative side similarly, the −2.5pp pass@32 movement (equivalent to one problem on the 40-problem AMC 2023 set) matches filtering out a small number of high-entropy lucky-correct rollouts together with their unstable neighborhood. We therefore report the +6.8pp / −2.5pp pattern as a *directional* signature of a shift from *chaotic coverage* toward *committed correctness*, not as an isolated significance claim, and corroborate it via the NLR→PLR quadrant migration in Table 5 and the AIME pass@32 gains in Table 2.

5.3 Training and Validation Dynamics

Figure 3 shows three projections of the same basin-deepening process from Section 5.2.

Smoother accuracy curves = basin consolidation. MGE-PO's avg@32 and pass@32 curves oscillate less than GRPO/DAPO with a monotonically widening gap. EA-Adv refuses to amplify high-entropy lucky-correct rollouts (Section 4.2),

Table 2: **Main results.** Qwen2.5 base models trained on DAPO-Math. We report avg@32 and pass@32 on four benchmarks; "Avg." denotes the unweighted average across all four datasets. Best per model–metric is in **bold**.

	AIME 2025		AIME 2024		AMC 2023		MATH-500		Avg.	
	avg@32	pass@32	avg@32	pass@32	avg@32	pass@32	avg@32	pass@32	avg@32	pass@32
Qwen2.5-Math-1.5B base										
Base	*3.3*	*28.7*	*6.9*	*10.3*	*32.5*	*72.5*	*32.8*	*77.8*	*18.9*	*47.3*
GRPO	6.3	30.7	9.5	30.3	52.7	90.0	66.5	89.4	33.7	60.1
DAPO	7.1	26.7	**11.0**	32.7	54.5	87.5	68.7	90.0	35.3	59.2
KL-Cov (Cui et al., 2025)	8.3	35.7	10.9	31.8	55.1	87.5	68.8	**90.8**	35.8	61.5
MGE-PO	**8.9**	**38.3**	10.2	**33.3**	**55.3**	**92.5**	**69.4**	89.2	**36.0**	**63.3**
Qwen2.5-Math-7B base										
Base	3.5	26.7	9.0	30.0	38.3	85.3	47.1	86.4	24.5	57.1
GRPO	10.8	35.4	20.6	36.7	61.9	90.0	72.4	90.8	41.4	63.2
DAPO	11.8	36.7	21.3	43.3	63.9	90.0	73.7	91.0	42.7	65.3
KL-Cov (Cui et al., 2025)	13.1	38.3	22.9	44.4	62.4	91.3	**75.8**	90.6	**43.6**	66.2
MGE-PO	**13.6**	**43.3**	**23.1**	**50.0**	**64.3**	**95.0**	73.1	**91.6**	43.5	**70.0**
Qwen2.5-7B base										
Base	1.3	6.7	2.3	16.7	20.0	70.0	61.8	75.4	21.3	42.2
GRPO	9.6	23.3	20.8	30.3	58.7	**92.5**	76.7	88.2	41.5	58.6
DAPO	10.1	25.7	22.1	36.7	59.5	87.5	78.3	90.4	42.5	60.1
KL-Cov (Cui et al., 2025)	12.7	28.3	22.6	39.5	61.4	91.2	**80.8**	89.8	44.4	62.2
MGE-PO	**12.9**	**30.3**	**23.2**	**40.0**	**65.5**	90.0	79.8	**90.8**	**45.4**	**62.8**

so consolidating *already-stable* correct paths is the only positive-advantage direction.

Surviving entropy = basin width, not noise. GRPO collapses to ~ 0.001 nats/token; MGE-PO stabilizes near 0.06. EA-KL/EA-Adv anchor the sharp regions predicted by Eq. (6) and leave forking tokens free to vary, so residual entropy is *structured* on tokens that steer reasoning.

Shorter responses = signal purification. GRPO inflates to $\sim$6,000 tokens by epoch 8; MGE-PO settles at $\sim$4,500. Penalizing NHR rollouts removes the incentive to pad chaotic CoTs, and EA-KL prevents drift into long-but-low-information continuations.

5.4 Ablation Studies

Component ablation (Q2). Table 3 reports the ablation on Qwen2.5-7B-Base using average avg@32. The full stack (45.4, +3.9pp) outperforms all partial configurations. EA-Adv is the largest single-component driver (+1.8pp); EA-KL alone yields +0.7pp, consistent with its stabilization role; DynZPD alone is flat but becomes useful once combined with EA-Adv or EA-KL, indicating that changing the question distribution requires a coupled learning-signal modification.

Quadrant coefficient ablation (Q3). Table 4 tests whether the asymmetric quadrant design in Section 4.2 is necessary. We evaluate coefficient variants on Qwen2.5-Math-7B-Base using AIME 2024. The default configuration, denoted as **Ours**, activates only the PLR bonus and NHR penalty.

Table 3: **Component ablation analysis.** Evaluated on Qwen2.5-7B-Base. "Avg" denotes the unweighted average of avg@32 across AIME 2025, AIME 2024, AMC 2023, and MATH-500. Δ tracks the absolute margin over standard GRPO.

EA-Adv	EA-KL	DynZPD	Avg (%)	Δ (pp)
✓	✓	✓	**45.4**	**+3.9**
✓	✓		44.3	+2.8
✓		✓	43.9	+2.4
	✓	✓	42.5	+1.0
✓			43.3	+1.8
	✓		42.2	+0.7
		✓	41.6	+0.1
Standard GRPO (No Entropy Signal)			41.5	0.0

It achieves the strongest performance under both avg@32 and pass@32, while alternative sign patterns or coefficient placements consistently degrade performance.

Magnitude ordering. Removing either PLR or NHR drops pass@32 to 36.7, confirming both active quadrants are necessary. Swapping the two magnitudes partially recovers avg@32 but still trails Ours on pass@32, indicating overly strong penalties shrink best-of-32 coverage.

Stability constraint — empirical confirmation of distribution collapse. Activating NLR is harmful despite its intuitive appeal: the NHR$\leftrightarrow$NLR swap yields the lowest pass@32 (33.3), and even $\gamma_{\text{NLR}} = 0.1$ underperforms the default. We read this as the empirical realization of Eq. (6): the regions where an extra NLR penalty looks most warranted ($\pi_\theta \to 1$) are the same regions where

Table 4: **Sensitivity ablation of quadrant coefficients.** Each configuration toggles one part of the default asymmetric injection rule. Evaluated on Qwen2.5-Math-7B-Base using AIME 2024. Both avg@32 and pass@32 are reported because coefficient changes can affect per-sample accuracy and best-of-32 coverage differently.

Config.	PLR	NHR	PHR	NLR	avg	pass
Ours	0.5	0.3	0	0	**23.1**	**50.0**
PLR Removed	0	0.3	0	0	21.3	36.7
NHR Removed	0.5	0	0	0	22.2	36.7
Magnitudes Swapped	0.3	0.5	0	0	22.6	40.0
PHR Activated	0.5	0.3	0.3	0	21.8	36.7
NHR/NLR Swapped	0.5	0	0	0.3	21.6	33.3
NLR Activated	0.5	0.3	0	0.1	22.4	36.7

its compressive force diverges while the entropy budget vanishes, producing a measurable basin contraction (16.7pp pass@32 drop) rather than the intended correction. EA-Adv's asymmetry is thus a structural prerequisite for stable optimization, not a hyperparameter trade-off.

Sign pattern — correct-but-uncertain $\neq$ correct. Activating PHR also degrades performance, mirroring NLR on the positive side: amplifying outcomes the policy was not committed to teaches it to reinforce unstable paths and undoes the purification driving the AMC 2023 gain. The default asymmetric pattern thus encodes a directional bias—only confidence-aligned outcomes (PLR) deepen the basin, only confidence-misaligned errors (NHR) prune it, ambiguous quadrants are silenced.

5.5 Entropy Validity and Source of Gains

Quadrant migration: microscopic basin deepening. Table 5 shows that NLR dominates the base model (45.5%): low entropy alone does not imply correctness. After GRPO training, PLR nearly triples ($4.5\% \rightarrow 12.8\%$) while high-entropy quadrants barely move. Mechanically, *the dominant learning operation is conversion of NLR mass into PLR mass*, exactly what basin deepening predicts at the token-distribution level.

Entropy-stratified gains: where basins can be deepened. Figure 4 stratifies problems by base-model entropy. Gains concentrate on the low-to-intermediate buckets where a correct basin exists but is not yet sharp. The highest-entropy bucket (Q5) shows negative transfer: with no consolidating basin to deepen, diffuse exploration leaves no positive signal to amplify. Bucket-wise avg@32 gains are at least as strong as pass@32 gains, contrary to the usual RLHF picture and consistent with

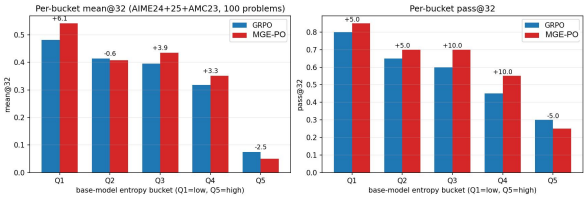

Figure 4: **Entropy-stratified gains.** The pooled evaluation set combines AIME 2024, AIME 2025, and AMC 2023 (100 problems total: $30+30+40$). Each problem is scored by $\bar{H}_i$ (Eq. (4), top-K pooled token entropy with $r_K = 0.1$) averaged over 64 base-model rollouts, and the pool is partitioned into five equal-frequency quintiles (Q1–Q5; $n = 20$ problems per bucket). Bars: MGE-PO minus GRPO within each quintile. Left: avg@32; right: pass@32.

Table 5: **Quadrant distribution** on 3,840 randomly sampled rollouts from the evaluation suite, computed on the *same* set before ("Base") and after standard GRPO ("GRPO"). Quadrants combine grader q_i with confidence u_i (Eq. (11), threshold 0.5). NLR dominates in the base model; after GRPO, PLR nearly triples.

	Base		GRPO	
Quadrant	Count	%	Count	%
PLR: Low ent. + Correct	173	4.5	492	12.8
PHR: High ent. + Correct	74	1.9	96	2.5
NLR: Low ent. + Incorrect	1747	45.5	1428	37.2
NHR: High ent. + Incorrect	1846	48.1	1824	47.5

the AMC 2023 signature.

6 Conclusion

MGE-PO injects per-token entropy at three granularities—rollout advantage, token KL, and question curriculum. On Qwen2.5-7B and Qwen2.5-Math-7B trained on DAPO-Math, it improves over GRPO by +3.9pp average avg@32 and +13.3pp AIME 2024 pass@32; ablations show the three are non-redundant and the asymmetric sign pattern is necessary. The +6.8/−2.5pp AMC 2023 shift with NLR→PLR migration indicates MGE-PO *densifies* the correct-reasoning basin rather than widening it: confidence at the right granularities turns outcome-driven RL into basin-deepening.

Limitations

Five caveats bound the present scope. (i) Generation entropy is a proxy for *internal confidence*, not objective difficulty; the two decouple at the extremes, and Q5 (Section 5.5) is the operational manifestation: when no recoverable basin exists, no entropy-aware mechanism can synthesize one. (ii) Principal results cover 1.5B/7B scales; transfer to 14B/32B and to non-mathematical domains, to-

gether with length-invariant entropy aggregation, remains to be verified. (iii) The added hyperparameters are each non-redundant (Table 3) and show a broad plateau in the sweep (Appendix E); what matters is the asymmetric *sign pattern*, not magnitudes. (iv) The smallest evaluation sets (AIME 2024 and AIME 2025, 30 problems each) have a single-problem resolution of ≈ 3.3pp, so individual pp-level differences on those two benchmarks need to be read alongside the larger MATH-500 and AMC 2023 evidence, together with the consistent direction of effect across all four datasets in Table 2. (v) The quadrant-coefficient ablation (Table 4) is run on AIME 2024 only; the ordering $\gamma_{\text{PLR}} > \gamma_{\text{NHR}} > 0$ with PHR/NLR silent follows from Eq. 6, and we reproduced it on AIME 2025 in our internal logs, though exact pp gaps could shift across datasets.

Ethical Considerations

MGE-PO is a post-training algorithm evaluated on publicly released mathematical reasoning benchmarks (DAPO-Math (Yu et al., 2025) for training; AIME 2024/2025, AMC 2023, and MATH-500 (Hendrycks et al., 2021) for evaluation), with no human subjects or newly collected data. Like any method that improves LLM reasoning, the resulting model carries dual-use risk in scenarios such as automated assistance with assessments; our changes are restricted to the credit-assignment and KL-regularization terms of GRPO and do not alter the model's deployment surface relative to standard GRPO baselines. Computationally, MGE-PO runs at $\sim 1.0\times$ the wall-clock of GRPO under the same compute budget (Appendix G), so it does not materially raise the entry barrier for follow-up research. We note an entropy-specific failure mode: a lower-entropy policy is not the same as a more correct policy. Table 5 shows that low-entropy incorrect rollouts (NLR) still account for 37.2% of rollouts after standard GRPO, so methods that sharpen committed trajectories can also sharpen confident errors; in high-stakes deployment (e.g., automated tutoring or assessment), users should not interpret lower generation entropy as calibrated epistemic uncertainty.

References

Udbhav Bamba, Minghao Fang, Yifan Yu, Haizhong Zheng, and Fan Lai. 2025. XRPO: Pushing the limits of GRPO with targeted exploration and exploitation. *arXiv preprint arXiv:2510.06672*.

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

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. 2025. The entropy mechanism of reinforcement learning for reasoning language models. *arXiv preprint arXiv:2505.22617*.

DeepSeek-AI. 2025. DeepSeek-R1: Incentivizing reasoning in LLMs through reinforcement learning. *Nature*, 645(8081):633–638.

Jiajun Fan, Tong Wei, Chaoran Cheng, Yuxin Chen, and Ge Liu. 2025. Adaptive divergence regularized policy optimization for fine-tuning generative models. *arXiv preprint arXiv:2510.18053*.

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

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the MATH dataset. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*.

Wenke Huang, Quan Zhang, Yiyang Fang, Jian Liang, Xuankun Rong, Huanjin Yao, Guancheng Wan, Ke Liang, Wenwen He, Mingjun Li, Leszek Rutkowski, Mang Ye, Bo Du, and Dacheng Tao. 2025. MAPO: Mixed advantage policy optimization. *arXiv preprint arXiv:2509.18849*.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan

Hume, Anna Chen, Yuntao Bai, Samuel Bowman, Stanislav Fort, and 17 others. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Deyang Kong, Qi Guo, Xiangyu Xi, Wei Wang, Jingang Wang, Xunliang Cai, Shikun Zhang, and Wei Ye. 2026. Rethinking the sampling criteria in reinforcement learning for LLM reasoning: A competence-difficulty alignment perspective. In *Proceedings of the 40th AAAI Conference on Artificial Intelligence*, pages 31438–31446.

Zicheng Lin, Tian Liang, Jiahao Xu, Qiuzhi Liu, Xing Wang, Ruilin Luo, Chufan Shi, Siheng Li, Yujiu Yang, and Zhaopeng Tu. 2025. Critical tokens matter: Token-level contrastive estimation enhances LLM's reasoning capability. In *Proceedings of the 42nd International Conference on Machine Learning*, pages 37906–37918.

Volodymyr Mnih, Adrià Puigdomènech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. 2016. Asynchronous methods for deep reinforcement learning. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1928–1937. PMLR.

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. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744.

Mihir Prabhudesai, Lili Chen, Alex Ippoliti, Katerina Fragkiadaki, Hao Liu, and Deepak Pathak. 2025. Maximizing confidence alone improves reasoning. *arXiv preprint arXiv:2505.22660*.

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

Jean Vassoyan, Nathanaël Beau, and Roman Plaud. 2025. Ignore the KL penalty! Boosting exploration on critical tokens to enhance RL fine-tuning. In *Findings of the Association for Computational Linguistics: NAACL 2025*.

Chen Wang, Zhaochun Li, Jionghao Bai, Yuzhi Zhang, Shisheng Cui, Zhou Zhao, and Yue Wang. 2025a. SCOPE-RL: Second-order curvature-aware policy optimization for llm reasoning. *arXiv preprint arXiv:2510.08141*.

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. 2025b. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for LLM reasoning. *arXiv preprint arXiv:2506.01939*.

Zhenting Wang, Guofeng Cui, Yu-Jhe Li, Kun Wan, and Wentian Zhao. 2025c. DUMP: Automated distribution-level curriculum learning for RL-based LLM post-training. *arXiv preprint arXiv:2504.09710*.

Ronald J. Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8(3–4):229–256.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. 2024. Can LLMs express their uncertainty? An empirical evaluation of confidence elicitation in LLMs. In *International Conference on Learning Representations*.

An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, Keming Yang, Le Lu, Mingfeng Liu, Runji Cheng, Yang Yang, Yichang Xiao, Zeyu Tu, Wenru Zhang, Shanghaoran Jiang, and 83 others. 2024. Qwen2.5-Math technical report: Toward mathematical expert model via self-improvement. *arXiv preprint arXiv:2409.12122*.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng,

Yuxuan Tong, Zhang Chi, Mofan Zhang, Wang Wang, Hang Zhu, and 15 others. 2025. DAPO: An open-source LLM reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*.

A Entropy Analyses

Entropy–correctness correlation and difficulty scaling. Pooled rollout entropy is measured on base-model samples and compared against binary correctness and benchmark difficulty. These diagnostics support the use of entropy as an online confidence proxy and complement the quadrant distribution in Table 5.

Training dynamics. Figure 5 tracks validation metrics across checkpoints. Internal metrics in Figure 6 show that MGE-PO preserves higher late-stage policy entropy, produces shorter responses, and keeps KL bounded throughout training.

B Extended Implementation Details of Entropy-Aware Advantage Injection

This appendix provides the full mathematical descriptions, tracking mechanics, and optimization safeguards for the Entropy-Aware Advantage Injection outlined in Section 4.2.

B.1 Confidence Score from Pooled Entropy

The confidence score $u_i \in [0, 1]$ used in Eq. (5) maps the pooled rollout entropy $\bar{H}_i$ to a calibrated certainty by first standardizing against running statistics and then applying a logistic squashing:

$$u_i = 1 - \sigma(z_i), \qquad z_i = \frac{\bar{H}_i - \mu_H}{\tau_H + \varepsilon_u}, \tag{11}$$

where $\sigma(\cdot)$ is the standard logistic function, $\varepsilon_u = 10^{-6}$ prevents division by zero, and (μ_H, τ_H) are running estimates of the batch-level mean and standard deviation of pooled entropy (Section B.2). This construction guarantees that u_i remains well-scaled across the dynamic shifts of policy uncertainty during training: rollouts that are unusually confident relative to the current batch ($z_i \ll 0$) approach $u_i \to 1$, while rollouts that are unusually uncertain ($z_i \gg 0$) approach $u_i \to 0$.

B.2 Global Exponential Moving Average (EMA) Tracking

To map the raw pooled token-entropy $\bar{H}_i$ to a stable, standardized score z_i across varying training phases, we dynamically maintain the global running mean μ_H and running standard deviation τ_H. Let $\bar{h}_b$ and s_b denote the sample mean and sample standard deviation of $\{\bar{H}_i\}_{i=1}^{B}$ computed within the current training batch b, respectively:

$$\bar{h}_b = \frac{1}{B}\sum_{i=1}^{B} \bar{H}_i, \tag{12}$$

$$s_b = \sqrt{\frac{1}{B-1}\sum_{i=1}^{B}(\bar{H}_i - \bar{h}_b)^2}. \tag{13}$$

At each optimizer step, μ_H and τ_H are updated via a synchronized exponential moving average:

$$\mu_H \leftarrow \lambda_{\text{ema}}\mu_H + (1 - \lambda_{\text{ema}})\bar{h}_b, \tag{14}$$

$$\tau_H \leftarrow \lambda_{\text{ema}}\tau_H + (1 - \lambda_{\text{ema}})s_b, \tag{15}$$

where the momentum coefficient is fixed at $\lambda_{\text{ema}} = 0.99$. Both variables are initialized using the empirical statistics $(\bar{h}_1, s_1)$ derived from the very first training batch to prevent transient scaling biases during early training iterations.

B.3 Derivation of the Distribution Collapse Limits

We justify the two limits in Eq. (6). Let $p = \pi_\theta(a_{\text{wrong}})$ denote the probability mass on a confident-incorrect token, so that the remaining mass $(1 - p)$ is distributed over the other $|\mathcal{V}| - 1$ tokens. The Shannon entropy decomposes as

$$\begin{aligned}\mathcal{H}(\pi_\theta) = &-p\log p - (1-p)\log(1-p) \\ &+ (1-p)\,\mathcal{H}(\pi_\theta^{\setminus a_{\text{wrong}}}),\end{aligned} \tag{16}$$

where $\mathcal{H}(\pi_\theta^{\setminus a_{\text{wrong}}})$ is the entropy of the renormalized distribution over the remaining vocabulary.

Entropy reservoir limit. As $p \to 1$, each of the three terms vanishes: $-p\log p \to 0$ (since $\log 1 = 0$), $(1-p)\log(1-p) \to 0$ (the standard $x \log x \to 0$ limit as $x \to 0$), and $(1-p)\,\mathcal{H}(\pi_\theta^{\setminus a_{\text{wrong}}}) \to 0$ because the prefactor $(1-p)$ vanishes and $\mathcal{H}(\pi_\theta^{\setminus a_{\text{wrong}}}) \le \log(|\mathcal{V}| - 1)$ is bounded. Hence $\mathcal{H}(\pi_\theta) \to 0^+$: the entropy reservoir empties as the policy concentrates on the incorrect token, leaving no budget for any further compressive update.

Marginal deviation cost limit. We deliberately track the *marginal* (not total) KL cost: when π_θ already sits near the corner of the simplex, the total $\mathrm{KL}(\pi_\theta \,\|\, \pi_{\text{ref}})$ is bounded above by

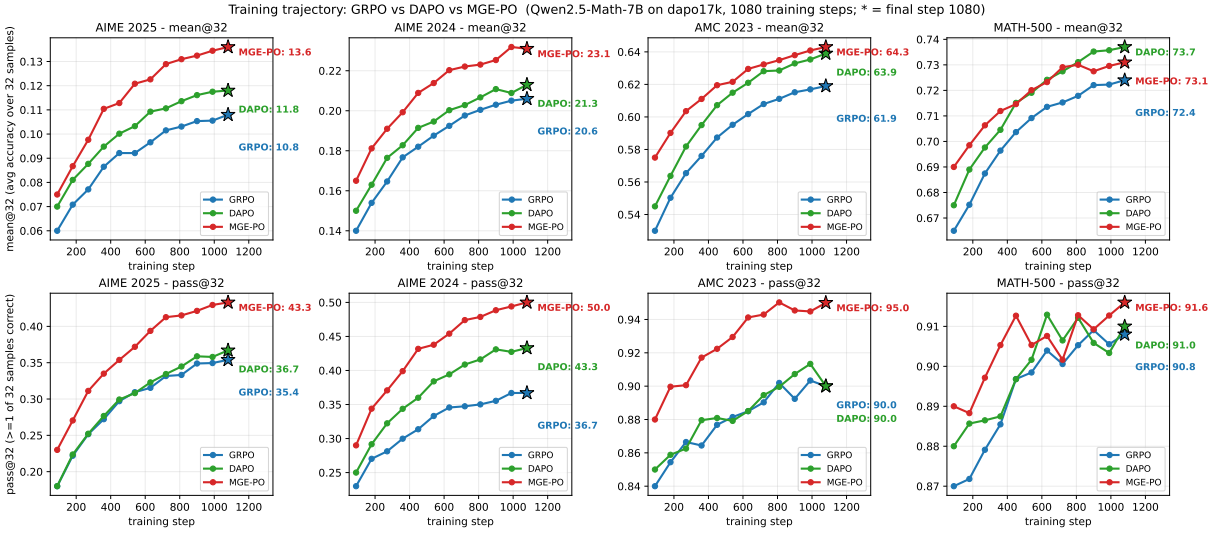

Figure 5: **Validation dynamics.** Training trajectory on Qwen2.5-Math-7B (DAPO-Math, $\sim$ 1080 steps); 32 samples/problem at $T = 1.0$. Final-step values match Table 2. Blue = GRPO, green = DAPO, red = MGE-PO.

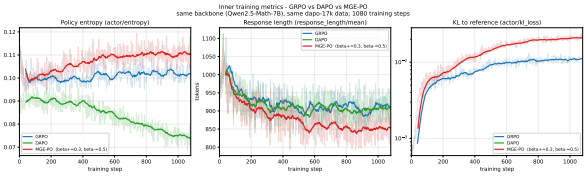

Figure 6: **Internal metrics.** Policy entropy, response length, and KL to reference. MGE-PO retains the highest entropy, the shortest responses, and bounded KL drift.

$-\log \pi_{\text{ref}}(a_{\text{wrong}})$, but its derivative with respect to $p = \pi_\theta(a_{\text{wrong}})$ diverges, making any further compressive update arbitrarily expensive per unit of probability mass moved.

Concretely, restrict attention to the two-outcome projection $\{a_{\text{wrong}}, \text{rest}\}$ (which lower-bounds the full KL by the data-processing inequality). Writing $q = \pi_{\text{ref}}(a_{\text{wrong}}) \in (0, 1)$ and using the binary KL

$$\mathrm{KL}_{\text{bin}}(p \,\|\, q) = p \log \frac{p}{q} + (1-p) \log \frac{1-p}{1-q}, \quad (17)$$

the derivative admits the closed form

$$\frac{\partial\, \mathrm{KL}_{\text{bin}}(p \,\|\, q)}{\partial p} = \log \frac{p\,(1-q)}{q\,(1-p)} \xrightarrow{p \to 1} +\infty. \quad (18)$$

Equivalently, the Fisher information $I_F(p) = 1/[p(1-p)] \to \infty$ as $p \to 1$: a fixed KL trust-region budget allows only vanishingly small further increments in p, so each additional NLR penalty step purchases progressively less probability movement at exploding KL cost.

A vanishing entropy reservoir together with a diverging marginal KL cost destabilizes the local softmax manifold and formalizes the distribution collapse argument in Section 4.2: there is no entropy left to absorb compressive updates, and the trust-region geometry punishes further compression super-linearly. EA-KL (Section 4.3) is designed precisely to absorb this regime: by anchoring sharp tokens more strongly to π_{ref} via $w_{i,t} > 1$, it lets the reference policy gently restore exploration rather than letting an additional NLR penalty drive the trust-region collapse further.

B.4 Generalized Four-Quadrant Mathematical Formulation

In the comprehensive framework, advantage corrections can be symmetrically parameterized across all four combinations of rollout correctness and confidence. The full definition of the additive modifications Δ_i^+ and Δ_i^- is given by:

$$\Delta_i^+ = q_i \big[\gamma_{\text{PLR}}\, u_i + \gamma_{\text{PHR}}(1 - u_i)\big], \quad (19)$$

$$\Delta_i^- = (1 - q_i) \big[\gamma_{\text{NHR}}(1 - u_i) + \gamma_{\text{NLR}}\, u_i\big], \quad (20)$$

where γ_{PLR}, γ_{PHR}, γ_{NHR}, and γ_{NLR} correspond to Positive-Low-Reward (Confident-Correct), Positive-High-Reward (Lucky-Correct), Negative-High-Reward (Uncertain-Incorrect), and Negative-Low-Reward (Confident-Incorrect) quadrants, respectively. The default configuration deactivates the unstable quadrants by setting $(\gamma_{\text{PHR}}, \gamma_{\text{NLR}}) = (0, 0)$, reducing the general form to Eq. (5).

B.5 Degenerate-Group Safeguards and Cap-Aware Gating

When a GRPO group encounters a degenerate state, the sample variance of the group rewards collapses

to zero ($\sigma_{g(i)} = 0$), yielding a base GRPO advantage of $A_i^{\text{GRPO}} = 0$. To insulate optimization from these edge cases, we use the cap-aware gate

$$g_i = 1 - \mathbf{1}[o_i \text{ truncated}], \tag{21}$$

which forces entropy-aware modifications to zero for truncated responses. The dynamic curriculum sampler further downweights trivially easy or intractable questions by decaying their sampling weights toward the floor η.

B.6 Batch-Level Re-Centering Optimization

Because the asymmetric coefficients ($\gamma_{\text{PLR}} \neq \gamma_{\text{NHR}}$) can introduce a non-zero mean shift to the advantage distribution, we apply batch-level zero-mean re-centering:

$$\widetilde{A}_i = A_i^{\text{EA}} - \frac{1}{B}\sum_{j=1}^{B} A_j^{\text{EA}}. \tag{22}$$

The token-level advantage used in the surrogate objective is then

$$\widetilde{A}_{i,t} = \widetilde{A}_i \cdot \text{mask}_{i,t}, \tag{23}$$

which excludes padding tokens from credit assignment.

C Extended Implementation Details of KL Regularization and Curriculum Sampling

C.1 Mathematical Formulations of EA-KL

To guarantee numerical stability and accurate gradient tracking during the execution of the Entropy-Aware KL loss defined in Eq. (7), we deploy the following formal implementations:

1. **Low-Variance KL Estimator:** Following standard policy optimization paradigms, the per-token divergence is calculated using the unbiased, low-variance form:

$$\begin{aligned} \text{KL}_{i,t} &= e^{\Delta_{i,t}} - \Delta_{i,t} - 1, \\ \text{with} \quad \Delta_{i,t} &= \log \pi_{\text{ref}}(a_{i,t} \mid c_{i,t}) \\ &\quad - \log \pi_\theta(a_{i,t} \mid c_{i,t}). \end{aligned} \tag{24}$$

2. **Bounded Stabilization Floor:** To insulate the relative sharpness ratio from extreme numerical fluctuations in late-stage convergence, the batch mean token entropy $\bar{H}$ in Eq. (8) is bounded from below via a strict numerical floor $H_{\min} = 5 \times 10^{-3}$:

$$\bar{H} \leftarrow \max\left(\frac{\sum_{i,t} \text{mask}_{i,t} H_{i,t}}{\sum_{i,t} \text{mask}_{i,t}}, H_{\min}\right) + 10^{-8}. \tag{25}$$

The upper clipping boundary for the weight factor is constrained at $w_{\max} = 1.5$, preventing single highly convergent sequences from monopolizing the global regularization budget.

C.2 Tracking and Initialization of Question Difficulty Statistics

The question-level difficulty indicator $\xi_q^{(t)}$ utilized in the Gaussian sampling weight (Eq. (9)) is dynamically updated at each training iteration using the per-question mean token entropy. Let $\mathcal{B}_q$ denote the set of parallel rollouts generated for question q within the current batch. The instant batch-level mean entropy ξ_q^{current} is computed as:

$$\xi_q^{\text{current}} = \frac{1}{|\mathcal{B}_q|}\sum_{i \in \mathcal{B}_q} \bar{H}_i, \tag{26}$$

where $\bar{H}_i = \frac{1}{T_i}\sum_{t=1}^{T_i} H_{i,t}$ is the mean token entropy of rollout i. The historical tracking EMA is updated via:

$$\xi_q^{(t)} = \lambda_q \xi_q^{(t-1)} + (1 - \lambda_q)\xi_q^{\text{current}}, \tag{27}$$

with the tracking momentum set to $\lambda_q = 0.9$. For newly encountered questions or at the inception of the optimization process ($t = 0$), the historical values are initialized to the target center ($\xi_q^{(0)} \leftarrow \mu_{\text{ZPD}} = 0.30$), ensuring a smooth transition from uniform exploration to curriculum concentration as the moving averages accumulate entropy statistics.

D Hyper-parameters

Table 6 lists all hyper-parameters held fixed across ablations.

E Coefficient Sensitivity

Representative slices of the $(\gamma_{\text{PLR}}, \gamma_{\text{NHR}})$ sweep on AIME 2024 are summarized by Table 4: the default $(0.5, 0.3)$ attains the best pass@32 (50.0), the swapped ordering $(0.3, 0.5)$ drops to 40.0, and either single-active configuration $(0, 0.3)$ or $(0.5, 0)$ falls to 36.7. The asymmetric ordering $\gamma_{\text{PLR}} > \gamma_{\text{NHR}} > 0$ is therefore the load-bearing choice, robust within a ± 0.2 envelope around the default point but degrading sharply once any of the silent quadrants is activated (Section 5.4).

Table 6: **Principal MGE-PO hyper-parameters** (held fixed across all reported runs). The baseline GRPO uses $\beta_{\mathrm{KL}} = 0.04$ (Section 5.1); MGE-PO uses a smaller global $\beta_{\mathrm{KL}} = 10^{-3}$ because per-token re-weighting via EA-KL ($w_{i,t} \in [0, w_{\max}]$, Section 4.3) already provides the regularization budget locally where it is needed.

Component	Symbol	Value
Training (shared with GRPO)		
Global batch size	B	128
Rollouts per question	G	16
Training temperature	T_{train}	1.0
Max sequence length	$L_{\max}$	12,288
Training epochs	—	8
EA-Adv (Section 4.2)		
Quadrant gain (PLR)	γ_{PLR}	0.5
Quadrant gain (NHR)	γ_{NHR}	0.3
Quadrant gain (PHR, NLR)	$\gamma_{\mathrm{PHR}}, \gamma_{\mathrm{NLR}}$	0
Top-K ratio / cap	$r_K, K_{\max}$	0.1, 4096
EMA decay	λ_{ema}	0.99
EA-KL (Section 4.3)		
Reweighting slope	α	0.20
Entropy floor	$H_{\min}$	5×10^{-3}
Weight bounds	$w_{\mathrm{floor}}, w_{\max}$	0, 1.5
KL coefficient	β_{KL}	10^{-3}
DynZPD (Section 4.4)		
Centre / bandwidth	$\mu_{\mathrm{ZPD}}, \sigma_{\mathrm{ZPD}}$	0.30, 0.15
Sampling floor	η	0.10
EMA decay (question)	λ_q	0.9

F Over-optimisation Analysis

The joint action of the clipped surrogate and EA-KL (Section 4.3) imposes a stable fixed point on the logit growth of confident-correct tokens. The clip bounds prevent $|\rho_{i,t}|$ from exceeding $1+\varepsilon_h = 1.28$, while EA-KL anchors low-entropy token distributions toward π_{ref} with weight up to $w_{\max} = 1.5$. Removing KL entirely ($\beta_{\mathrm{KL}} = 0$) raises training-end KL from 0.5 to ~ 3 nats and drops all benchmarks by 4–7pp, confirming that the KL term is non-redundant with the clip alone.

G Compute Budget

All experiments are run on a single node equipped with 2×NVIDIA H200 (141 GB HBM3e) GPUs. Each Qwen2.5 (1.5B or 7B) training run trains for 8 epochs over the 17,398-question DAPO-Math set with $G = 16$ on-policy rollouts per question at training temperature $T_{\mathrm{train}} = 1.0$ and global batch size 128 (max sequence length 12,288); evaluation uses 32 samples per problem at $T = 1.0$, $\mathrm{top}_p = 0.7$ on each of AIME 2024/2025, AMC 2023, and MATH-500. The end-to-end wall-clock for a single MGE-PO run on Qwen2.5-7B-Base under this protocol is approximately $\sim 1.0\times$ that of the matched GRPO baseline, since the entropy signal $H_{i,t}$ is computed once during rollout and cached, so EA-Adv, EA-KL, and DynZPD add no rollout compute, only a small constant cost during the policy update (advantage reshaping, per-token weight, and per-question sampling-weight EMA). Exact GPU-hour accounting per run will be released alongside the code.

H Paper Checklist

1. For all authors:
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] Section 6.
   (c) Have you read and adhered to the ACL submission guidelines? [Yes]
   (d) Have you cited relevant prior work? [Yes]
2. If you ran experiments:
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results? [Yes, upon acceptance] Implementation builds on the open-source DAPO codebase (Yu et al., 2025) with our entropy-aware extensions; the full training and evaluation pipeline will be released alongside the camera-ready version.
   (b) Did you specify all the training details? [Yes] Section 5.1 and Appendix D.
   (c) Did you report error bars or confidence intervals? [No] Compute constraints limited each (model, method) cell to a single training run; we instead report avg@32 and pass@32 over 32 inference samples per problem.
   (d) Did you include the total amount of compute and the type of resources used? [Yes] See Appendix G.