LNPO-ATC: LENGTH-NORMALIZED POLICY OPTIMIZATION WITH ADAPTIVE TOKEN CREDIT FOR LONG-CONTEXT REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement learning has improved reasoning in large language models, but long reasoning trajectories remain challenging under finite generation budgets. Uniformly applying a terminal length penalty can turn positive content advantages into negative update coefficients throughout a response, coupling budget control with penalties on earlier reasoning tokens. We propose Length-Normalized Policy Optimization with Adaptive Token Credit (LNPO-ATC) to address this conflict. The method combines length-normalized sequence weighting, a smooth boundary penalty, and token-level allocation based on position relative to the available budget, while retaining the content advantage at every token. Our analysis separates penalty-strength and positional contributions and derives sufficient conditions for local improvement through alignment with the true content gradient. The evaluation covers six reasoning benchmarks, long-input generalization, and allocation controls. LNPO-ATC improves the six-task macro-average score on Qwen3.6-35B-A3B by an absolute 5.53% over GSPO; ATC adds an absolute 2.27% over uniform LNPO. The ordered-versus-shuffled comparison, which also varies EOS weighting, yields an absolute 1.79% gain. At equal counts of 2,000 updates, training wall time is 1.4–1.5% higher than GSPO on the two Qwen3.6 backbones.

1 INTRODUCTION

Reinforcement learning (RL) with verifiable rewards has improved reasoning in large language models (LLMs), eliciting behaviors such as self-reflection and answer verification (Guo et al., 2025). These reasoning processes operate within finite generation budgets: a hard limit can truncate an unfinished response, and inappropriate penalties can introduce reward noise even when intermediate reasoning is sound (Yu et al., 2025). Thinking-budget mechanisms and RL-based length control offer ways to balance reasoning cost and task performance (Yang et al., 2025; Aggarwal & Welleck, 2025; Wu et al., 2025). During training, this balance also depends on how budget-related penalties interact with the token updates that learn from task outcomes.

We focus on two coupled problems in this interaction. First, uniformly broadcasting a terminal length penalty can reverse positive content coefficients throughout a response when the penalty exceeds the content advantage, including at earlier reasoning positions (Eq. (4); Appendix B.7.5). Second, positional allocation can change both where the penalty acts and its average strength. An improvement over uniform allocation may therefore reflect weaker regularization, more favorable placement, or both. This motivates a budget-relative allocation rule and an analysis that distinguishes the contributions of penalty strength and position.

We propose **Length-Normalized Policy Optimization with Adaptive Token Credit (LNPO-ATC)** to address these problems. A cubic soft barrier activates the length penalty within a warning region near the generation limit. ATC allocates this penalty according to token position relative to the available budget, retaining the outcome-based content advantage at every token. Tokens before the warning region receive no explicit length penalty, while coefficients increase within the warning region. The method uses the length-normalized sequence weighting and token-gradient construction of GSPO and GSPO-token (Zheng et al., 2025a). Uniform LNPO broadcasts the terminal coefficient within the same objective family and serves as the corresponding control.

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.

Our analysis separates the local gradient difference into mean-strength and positional covariance terms (Proposition 1). This decomposition motivates response-wise mass matching and coefficient permutations to distinguish strength, concentration, and temporal placement. Under explicit structural, local smoothness, and step-size conditions, we derive sufficient conditions for a positive content-direction margin and a finite-step improvement bound relative to uniform allocation (Theorem 1). A conditional extension accounts for a specified optimizer displacement under the same curvature assumptions. These results characterize when allocation can benefit the local content objective under the stated assumptions.

The evaluation covers six reasoning tasks, long-input generalization, allocation controls, penalty sensitivity and training resources. The complete method gains 5.53 macro-average points over GSPO on Qwen3.6-35B-A3B; ATC adds 2.27 points over uniform LNPO. Ordered allocation exceeds shuffling by 1.79 points in a comparison that also varies EOS weighting. Exact enumeration provides a separate mathematical check of the local identities in a two-logit policy.

2 Related work

PPO uses clipped surrogates (Schulman et al., 2017), GRPO uses group-relative rewards and token ratios (Shao et al., 2024), and GSPO provides the sequence and token constructions used here (Zheng et al., 2025a). DAPO includes soft overlong punishment (Yu et al., 2025). Recent RLVR work studies outlier-sensitive gradient weighting in GMPO (Zhao et al., 2026) and mixtures of token and sequence importance ratios in DHPO (Min et al., 2026). Our comparison holds the GSPO weighting mechanism fixed to identify the effects of boundary coefficients.

Length control methods include LCPO's user-specified constraints (Aggarwal & Welleck, 2025) and LAPO's learned reasoning-length adaptation (Wu et al., 2025). CurES analyzes prompt selection and rollout allocation through gradient optimization (Zeng et al., 2026). These are complementary allocation levels: our coefficients act within a response. Length bias and long-context RL motivate reporting accuracy jointly with generated tokens (Singhal et al., 2023; Wan et al., 2025).

Classical policy-gradient analysis defines our expected verifier objective (Sutton et al., 1999); TRPO provides a reference for relating surrogate updates to policy improvement (Schulman et al., 2015). DiCE establishes correct higher-order differentiation of stochastic objectives (Foerster et al., 2018), relevant to our curvature derivation. We prove the boundary-allocation identities and sufficient conditions for the present surrogate rather than transferring a guarantee from these algorithms.

3 LNPO-ATC: objective and allocation

3.1 A shared objective and its uniform control

For prompt x, sample G responses from the frozen policy π_{θ_0}. A response has L_i generated tokens, including a sampled EOS when present, and available generation budget $H_x \le C - |x|$ for hardware context cap C. Forced termination has no policy score. A fixed verifier returns $R_i \in [0,1]$; incomplete responses receive the same declared treatment across methods. Define

$$A_i^C = \frac{R_i - \bar{R}_x}{\sigma_x + \delta}, \quad \bar{R}_x = G^{-1}\sum_j R_j, \quad \sigma_x^2 = G^{-1}\sum_j (R_j - \bar{R}_x)^2, \tag{1}$$

$$s_i(\theta) = \exp\left[\frac{1}{L_i}\sum_t \left(\log \pi_\theta(y_{i,t} \mid x, y_{i,<t}) - \log \pi_{\theta_0}(y_{i,t} \mid x, y_{i,<t})\right)\right]. \tag{2}$$

Rewards, lengths, group statistics, and old log-probabilities are detached; $\delta > 0$ is fixed. Equation (2) is exactly the GSPO weight, whereas the exact trajectory change-of-measure ratio is $s_i^{L_i}$ (Zheng et al., 2025a). Let $\ell(s,a) = \min\{sa, \mathrm{clip}(s, 1-\epsilon_-, 1+\epsilon_+)a\}$ with $\epsilon_\pm > 0$.

For $0 \le \alpha < 1$, put $w_x = (1-\alpha)H_x$ and define the cubic boundary cost

$$b_\alpha(u; H_x) = 3z^2 - 2z^3, \qquad z = \mathrm{clip}((u - \alpha H_x)/w_x, 0, 1). \tag{3}$$

The cost is zero before the warning region and reaches one at the cap. The system enforces the hard cap H_x, while the bounded soft barrier supplies a smooth regularization coefficient. The separate

hard-cut control uses $b_{\mathrm{hard}}(u; H_x) = \mathbf{1}\{u \geq H_x\}$ directly, avoiding division by a zero warning width at $\alpha = 1$. The uniform LNPO control adds the raw coefficient $A_i^L = -\beta c_i$, with $c_i = b_\alpha(L_i; H_x)$ and $\beta \geq 0$:

$$J_{\mathrm{LNPO}} = \mathbb{E}\left[G^{-1} \sum_i \ell(s_i, A_i^C - \beta c_i)\right]. \tag{4}$$

Within each prompt group, correctness rewards are normalized to obtain A_i^C. The length coefficient $-\beta c_i$ is then added at the fixed scale β, and the resulting coefficient $A_i^C - \beta c_i$ enters the clipped surrogate directly.

3.2 The ATC allocation rule

Use the existing GSPO-token construction, with sg denoting stop-gradient:

$$s_{i,t} = \mathrm{sg}[s_i] \exp\{\log \pi_\theta(y_{i,t} \mid x, y_{i,<t}) - \mathrm{sg}[\log \pi_\theta(y_{i,t} \mid x, y_{i,<t})]\}, \tag{5}$$

$$J(q) = \mathbb{E}\left[\frac{1}{G} \sum_i \frac{1}{L_i} \sum_t \ell(s_{i,t}, A_i^C - \beta q_{i,t})\right]. \tag{6}$$

Numerically $s_{i,t} = s_i$, but $\nabla s_{i,t} = s_i \nabla \log \pi_\theta(y_{i,t} \mid x, y_{i,<t})$. A constant $q_{i,t} = c_i$ recovers both the value and gradient of Eq. (4). For the same policy, response batch, detached statistics and clipping settings, $q = 0$ or $\beta = 0$ recovers the GSPO objective and gradient. Identical training trajectories additionally require matched initialization, optimizer state, data order and random draws. LNPO-ATC uses $q_{i,t} = b_\alpha(t; H_x)$ and retains the outcome-based content coefficient A_i^C at every token. Its allocation adapts to position relative to the available budget. Token coefficient signs determine the active clipping branches.

ATC changes two things: the penalty's position and its mean strength, since $\bar{q}_i = L_i^{-1} \sum_t b_\alpha(t; H_x) \leq c_i$. To hold the original LNPO strength fixed, use the diagnostic control

$$q_{i,t}^{\mathrm{match}} = \frac{L_i c_i b_\alpha(t; H_x)}{\sum_u b_\alpha(u; H_x)}, \qquad q_{i,t}^{\mathrm{match}} = 0 \text{ if the denominator is zero.} \tag{7}$$

Its response mean is c_i; mass conservation can give suffix coefficients above one, which are used as defined. A uniform-at-ATC-mean control uses $q_{i,t}^{\mathrm{mean}} = \bar{q}_i$. Comparing ATC with this control isolates nonuniform weighting at the same mean strength, including both concentration and temporal arrangement.

For a direct placement comparison, uniformly permute *original* ATC's valid positions: $q_{i,t}^{\mathrm{shuffle}} = q_{i,\rho_i(t)}$. This preserves its complete coefficient multiset, mean and squared-weight sum. Independently shuffle q^{match} in the same way. Table 4's six controls thus compare strength, nonuniform weighting and temporal placement within $J(q)$. Each permutation includes sampled EOS, excludes padding, and stays fixed across update epochs of that rollout.

A sampled EOS receives the positional length coefficient and participates in the permutation. Consequently, the shuffled comparison includes changes to EOS weighting as well as to other token positions; it does not isolate placement with the EOS coefficient held fixed. Appendix B.7.5 discusses coefficient-level protection. Algorithm 1 specifies the masks and update procedure. The reported default is $(\alpha, \beta) = (0.8, 0.9)$, with parameter comparisons in Section 5.

4 Boundary allocation and local policy improvement

Fix the prompt distribution, verifier, and generation budgets. Our reference is the true content objective

$$F(\theta) = \mathbb{E}_{x \sim \mathcal{D},\, y \sim \pi_\theta}[R(x, y)], \qquad v = \nabla F(\theta_0) = \mathbb{E}[R \textstyle\sum_t S_t], \quad S_t = \nabla \log \pi_{\theta_0}(y_t \mid x, y_{<t}). \tag{8}$$

The score expression follows policy-gradient differentiation (Sutton et al., 1999). We assume positive support for stochastic actions and interchange of gradient and expectation. Every comparison below starts at θ_0, where surrogate ratio clipping is inactive, with shared detached content advantages and raw gradient ascent. Let G_q be the gradient of one entire prompt group, $g_q = \mathbb{E}G_q$,

$D = G_q - G_c$, and $d = \mathbb{E}D$. Group normalization couples responses; independent sampling units are *prompt groups*. No identity between v and the normalized content update is assumed. All gradient scales are fixed in advance and shared. The theorem compares single raw-gradient steps at this common policy. AdamW preconditioning, gradient-norm clipping and repeated updates are outside that comparison; the displacement corollary applies only when its own directional and curvature conditions hold.

4.1 An exact separation of total strength and location

For a response, let $\bar{S} = L^{-1}\sum_t S_t$, $\bar{q} = L^{-1}\sum_t q_t$, and $\mathrm{Cov}_t(q, S) = L^{-1}\sum_t (q_t - \bar{q})(S_t - \bar{S})$. Here $\mathbb{E}$ over a response also averages its uniformly selected group index.

Proposition 1 (Mass–location decomposition)**.** *Under the local conditions above,*

$$d = \beta\mathbb{E}\left[(c - \bar{q})\bar{S} - \mathrm{Cov}_t(q, S)\right]. \tag{9}$$

For response-wise mass matching, $\bar{q} = c$ and only the covariance term remains. For any fixed weight vector, uniform permutation has conditional expected local gradient equal to broadcasting its response mean: $\mathbb{E}_\rho[G_{\rho q} \mid \text{group}] = G_{\bar{q}}$. Matched weights recover G_c.

The first term measures the change in mean regularization and the second its positional coupling to policy scores. Original ATC includes both terms, and their sum determines the projection on v. Original ATC versus its shuffled weights holds both strength and concentration fixed. The permutation identity concerns conditional expected raw gradients at the shared old policy; finite-step and optimizer effects require additional assumptions.

4.2 A structural condition and a stochastic one-step guarantee

Put $a_t = \langle v, S_t\rangle$ and $\mu_t = \mathbb{E}[a_t \mid x, L]$, the directional score projections. Suppose q is deterministic given (x, L, H_x) and mass matched. Its content-direction margin is

$$M = \langle v, d\rangle = -\beta\mathbb{E}[\mathrm{Cov}_t(q, a)]. \tag{10}$$

Theorem 1 (Sufficient allocation condition and local content benefit)**.** *On each penalized response stratum, split positions into a zero-weight prefix P and a suffix Q of size m, with $0 < m < L$. Suppose the conditional prefix projection exceeds the suffix projection by at least $\delta_{x,L} > 0$ and $\mathrm{Cov}_Q(q, \mu) \le 0$. Then*

$$M \ge \beta\mathbb{E}\left[c(1 - m/L)\delta_{x,L}\right], \tag{11}$$

with zero contribution on unpenalized strata. For $\beta > 0$ and positive probability of such penalized strata, the margin is strictly positive.

Further suppose ∇F is K-Lipschitz on a region containing all segments traversed by the two candidate updates. For N iid prompt groups, let $\bar{G}_j$ denote their mean, $V_j = \mathrm{tr}\,\mathrm{Var}(G_j)$, and $Q_N = \|g_q\|^2 + \|g_c\|^2 + (V_q + V_c)/N$. For $\eta > 0$,

$$\mathbb{E}F(\theta_0 + \eta\bar{G}_q) - \mathbb{E}F(\theta_0 + \eta\bar{G}_c) \ge \eta M - \tfrac{1}{2}K\eta^2 Q_N. \tag{12}$$

Thus $M > 0$ and $0 < \eta < 2M/(KQ_N)$ suffice for a positive expected difference when $KQ_N > 0$.

For any zero-prefix positional rule, including original ATC, put $\Delta = \mu_P - \mu_Q$, $C_Q = \mathrm{Cov}_Q(q, \mu)$, and $\bar{\mu} = L^{-1}\sum_t \mu_t$. The same calculation gives

$$M_q = \beta\mathbb{E}Z_q, \qquad Z_q = (c - \bar{q})\bar{\mu} + \bar{q}(1 - m/L)\Delta - \frac{m}{L}C_Q.$$

The three terms identify strength, prefix–suffix separation, and within-suffix coupling. Matched allocation sets $\bar{q} = c$ and recovers Eq. (11). With $Z_q^{\pm} = \max(\pm Z_q, 0)$, the exact margin is $\beta(\mathbb{E}Z_q^+ - \mathbb{E}Z_q^-)$. Thus mixed-sign strata can still yield positive total benefit; their magnitudes matter as well as their frequency. Corollary 1 formalizes a tolerance bound and the link from matched allocation to original ATC.

The finite-step argument retains its curvature assumption. Lemma 1 derives a sufficient local bound from token log-probability derivatives: for horizon $H_{\max}$, reward bound $R_{\max}$, score bound B_1, and log-probability Hessian bound B_2 on the update region,

$$K_{\mathcal{U}} \le R_{\max}(H_{\max}^2 B_1^2 + H_{\max}B_2).$$

This concerns F, separately from the smoothness of b_α. For actual displacements $U_j = \eta\bar{G}_j + e_j$, Corollary 2 adds the signed optimizer correction $\langle v, \mathbb{E}(e_q - e_c)\rangle$ and uses the actual displacement second moments. The correction is a term in a conditional bound, rather than an estimated optimizer effect. Appendix B.8 checks the raw-gradient identities and finite-step bounds within an exactly enumerable finite policy.

4.3 Variance and bias decomposition

Proposition 2 (Paired variance and common-target MSE)**.** *For square-integrable paired group gradients and a fixed reference v,*

$$V_q - V_c = \operatorname{tr}\operatorname{Var}(D) + 2\operatorname{tr}\operatorname{Cov}(G_c, D), \tag{13}$$

$$\Delta\mathrm{MSE}_N = 2\langle g_c - v, d\rangle + \|d\|^2 + (V_q - V_c)/N. \tag{14}$$

Here $\mathrm{MSE}_{j,N} = \mathbb{E}\|\bar{G}_j - v\|^2$*. Writing $G_j = C + \Lambda_j$, with the negative penalty included in Λ_j, gives $V_j = \operatorname{tr}\operatorname{Var} C + \operatorname{tr}\operatorname{Var}\Lambda_j + 2\operatorname{tr}\operatorname{Cov}(C, \Lambda_j)$.*

Concentrating matched mass on m positions gives $\sum_t q_t^2 \geq L^2c^2/m$. Its effect on total variance depends on the cross term in Eq. (13); MSE also includes bias. Appendix B gives the conditional concentration-cost example and proofs.

Prefix reversal and boundary sensitivity. Define coefficient reversal by $A^C > 0$ but $A^C - \beta c < 0$, with rate conditional on $A^C > 0$. ATC and the matched rule have zero explicit length coefficient for $t \leq \alpha H_x$, preserving the positive content coefficient there before clipping. This property concerns the coefficient assigned to each score term; whole-batch probability changes depend on the combined parameter update (Appendix B.7). The cubic extension is C^1, with Lipschitz constant $3/(2w_x)$, derivative Lipschitz constant $6/w_x^2$, and quadratic onset. These bounds describe length perturbations with fixed score factors. Appendix B.7 analyzes the GSPO backbone's normalization bias and conditional $1/L$ variance under explicit dependence assumptions.

5 Experiments

5.1 Experimental setup

The evaluation uses Qwen3.6-35B-A3B and dense Qwen3.6-27B on HMMT February 2026, IMO-AnswerBench, LiveCodeBench (LCB) v6, LCB Pro, ZebraLogic, and GPQA Diamond (HMMT, 2026; IMO-Bench Team, 2025; Zheng et al., 2025b; Lin et al., 2025; Rein et al., 2024). GSPO, LCPO, LAPO, AEPO, LNPO and LNPO-ATC form the common comparison set; Table 1 also includes PPO, GRPO and DAPO. Base denotes the starting checkpoint under the shared evaluation setting. Published Base (ref.) scores retain their provider protocols and are excluded from shared-protocol contrasts (Qwen Team, 2026b;a).

The LNPO-ATC defaults are AdamW, learning rate 10^{-6}, group size 16, batch size 128, 2,000 updates, $(\alpha, \beta) = (0.8, 0.9)$ and a generated-token budget $H_x = 32{,}768$. Prompts and outputs satisfy $|x| + H_x \leq C$; the boundary penalty becomes positive at generated-token position 26,215.

Data and comparison settings. The stated training allocation contains 80,000 prompts: 32,000 mathematics, 24,000 coding, 12,000 logic and 12,000 science questions. Appendix C.1 lists source partitions and the allocation. The comparison uses a common prompt-sampling rule, templates and correctness metrics, with method-specific responses. Appendix C.2 summarizes the optimization settings and baseline scope.

Baseline tuning. All trained baselines—PPO, GRPO, DAPO, GSPO, LCPO, LAPO and AEPO—underwent method-specific hyperparameter tuning. The reported comparisons use each baseline's tuned configuration. Table 6 separately characterizes the sensitivity of LNPO-ATC to penalty strength and onset. Appendix C.2 states the tuning scope.

Evaluation and aggregation. Scores are percentages: exact-answer verification for mathematics, test-based pass@1 for coding, whole-puzzle accuracy for ZebraLogic, and answer-choice accuracy for GPQA. Non-reference entries in Table 1 report means and sample SDs over five runs. Six-task macros use equal task weights. Macro SDs use within-run aggregation to retain cross-task covariance; pointwise 95% paired intervals use unrounded run-level differences within predefined randomization blocks (Appendix C). LongBench v2 is a custom 330-question subset with input-bin counts 60/90/100/80 and a separate 32,768-token output cap. Its results describe comparisons within that subset. Wall time and output volume are reported under their respective units in Table 8.

5.2 Core reasoning results

Table 1: **Six-task reasoning comparison on Qwen3.6-35B-A3B.** Base (ref.) quotes published scores (Qwen Team, 2026b); missing reference cells are dashed. Other rows show mean ± SD over five runs; trained baselines use their respective tuned configurations. Reference scores are excluded from the shared-protocol macro and contrasts.

Method	HMMT Feb26	IMO Answer	LCB v6	LCB Pro	Zebra Logic	GPQA Diamond
Base (ref.)	83.6	78.9	80.4	–	–	86.0
Base	82.4±0.87	77.6±1.03	79.8±0.68	41.2±1.28	62.4±1.12	85.1±0.52
PPO	83.8±0.92	79.4±0.96	80.7±0.71	42.9±1.21	64.5±1.06	85.9±0.48
GRPO	84.7±0.83	80.6±0.91	81.9±0.64	44.4±1.16	66.7±0.98	86.5±0.45
DAPO	85.6±0.79	81.5±0.87	82.8±0.62	45.7±1.09	68.4±1.03	87.0±0.47
GSPO	85.4±0.76	82.3±0.83	83.4±0.59	46.5±1.03	68.9±0.94	86.8±0.43
LCPO	86.3±0.72	82.0±0.85	84.1±0.56	47.3±0.98	68.7±0.91	87.4±0.38
LAPO	86.9±0.68	83.2±0.78	83.9±0.53	47.1±1.02	70.6±0.86	87.5±0.41
AEPO	87.4±0.71	83.9±0.74	84.7±0.51	48.6±0.94	71.5±0.83	87.8±0.36
LNPO	88.5±0.63	85.0±0.69	86.3±0.48	50.6±0.89	73.8±0.79	88.7±0.35
LNPO-ATC	90.3±0.58	87.1±0.65	88.4±0.42	53.5±0.86	77.2±0.74	90.0±0.31

Averaging the task means displayed in Table 1 gives macro scores of 75.55 for GSPO, 78.82 for LNPO and 81.08 for LNPO-ATC: a 5.53-point difference from GSPO and a 2.27-point difference from LNPO, computed before rounding the macros. These are descriptive summaries of the displayed task means. LCPO, LAPO and AEPO provide additional length-control comparisons (Aggarwal & Welleck, 2025; Wu et al., 2025; Shen et al., 2025). Appendix C defines the run-level summaries and the pointwise paired intervals. The score comparisons are descriptive; no multiplicity-adjusted significance claim is made. Official reference cells do not enter those comparisons or the macro.

Per-task gains are unevenly distributed. The two tasks with the lowest GSPO baselines receive the largest absolute improvements: LCB Pro rises from 46.5 to 53.5 (+7.0, a 15.1% relative gain) and Zebra Logic from 68.9 to 77.2 (+8.3, 12.0%), whereas GPQA Diamond, already at 86.8, gains 3.2 points (3.7%). This ordering is consistent with available headroom and remains a descriptive summary of the displayed means. Run-to-run dispersion narrows on every task: the mean per-task SD falls from 0.76 under GSPO to 0.59 under LNPO-ATC, a 22% reduction, and Table 4 reports the smallest macro SD for original ATC (0.19) among the six allocation variants. The length reduction compounds the score gain. Macro score per thousand generated tokens rises from 7.88 to 10.85 (+37.8%), and the truncation rate falls from 7.4% to 3.0%, a 4.4-point absolute reduction (Table 2). Under the evaluation protocol, answers left incomplete at the generation cap receive zero.

5.3 Default output length and truncation

Table 2 reports final evaluation results. During training, boundary penalties from rollouts with $L_i > \alpha H_x$ enter shared parameter updates and can affect later responses below the threshold. GSPO averages 9.59k tokens with 7.4% truncation; LNPO-ATC averages 7.47k with 3.0%. Incomplete or unparseable answers score zero and remain in all denominators. Appendix D.1 defines token counting and aggregation.

Table 2: **Default accuracy, length and truncation.** Qwen3.6-35B-A3B with a 32,768-token output cap. Macro is the arithmetic mean of the displayed task means in Table 1. Tokens (k) and truncation use equal task weights (Appendix D.1).

Method	Macro	Tokens (k)	Truncation (%)
GSPO	75.55	9.59	7.4
LCPO	75.97	8.86	6.0
LAPO	76.53	8.65	5.6
AEPO	77.32	8.51	5.3
LNPO	78.82	8.08	4.6
LNPO-ATC	81.08	7.47	3.0

5.4 LONG-INPUT GENERALIZATION

Table 3: **LongBench v2 custom-subset comparison.** The 32k output cap is separate from input length. Bin counts are 60/90/100/80. Weighted scores use these counts within each run, followed by mean $\pm$ sample SD across five runs (Appendix D.2). The scores describe this subset.

Method	$\leq$ 16k	16–32k	32–64k	64–128k	Weighted
Base	62.0 ± 1.4	57.1 ± 1.5	49.0 ± 1.9	41.0 ± 2.2	51.64 ± 0.92
GSPO	63.3 ± 1.2	58.7 ± 1.4	51.2 ± 1.8	43.5 ± 2.1	53.58 ± 0.87
LCPO	64.0 ± 0.9	59.6 ± 1.5	52.8 ± 1.9	45.8 ± 2.3	54.97 ± 0.85
LAPO	64.7 ± 0.7	60.9 ± 1.4	54.0 ± 1.7	46.5 ± 2.1	56.00 ± 0.81
AEPO	65.0 ± 1.2	61.1 ± 1.4	54.8 ± 1.6	47.5 ± 2.0	56.61 ± 0.78
LNPO	67.3 ± 0.9	63.6 ± 1.2	57.8 ± 1.6	51.5 ± 1.9	59.58 ± 0.70
LNPO-ATC	69.0 ± 0.9	66.4 ± 1.2	61.8 ± 1.5	55.3 ± 1.6	62.79 ± 0.66

LongBench v2 evaluates long-input understanding and reasoning (LongBench Team, 2024). Inputs are grouped by tokenized length before candidate evaluation, retained in full, and paired across methods. A fixed 32k output allowance is independent of the input bin. Decoding uses temperature 0.6 and top-p 0.95, with one response per question in each run. The 330-question evaluation uses inputs up to 128k tokens and is separate from a full-release benchmark score. Within this custom subset, the weighted ATC score is 62.79, a descriptive 9.21-point difference from GSPO. Appendix D.2 defines the input bins and weighted aggregation.

5.5 ALLOCATION AND PARAMETER SENSITIVITY

Table 4: **Allocation comparisons.** Five paired runs; C–D compares ordered and shuffled ATC coefficients, including sampled EOS. Each mass/multiset constraint applies per shared response. Reported contrasts and pointwise 95% intervals use paired run-level scores (Appendix C).

ID	Allocation	Mean	Macro	Paired contrast (95% interval)
A	Uniform LNPO	c	78.82 ± 0.34	Reference
B	Uniform at ATC mean	$\bar{q}$	79.24 ± 0.22	B–A: +0.42 [0.04, 0.80]
C	Original ATC	$\bar{q}$	81.08 ± 0.19	C–B: +1.85 [1.60, 2.10]
D	Shuffled original ATC	$\bar{q}$	79.30 ± 0.27	C–D: +1.79 [1.41, 2.16]
E	Terminal-matched ATC	c	80.86 ± 0.25	E–A: +2.05 [1.69, 2.40]
F	Shuffled terminal-matched	c	78.87 ± 0.28	E–F: +2.00 [1.40, 2.59]

C–D compares ordered and shuffled weights while preserving each shared response's coefficient multiset. The comparison includes the EOS coefficient and therefore combines placement and stopping-related weighting effects. B–A tests uniform weakening; C–B tests nonuniform allocation at ATC's mean strength. E–A and E–F repeat these checks at terminal mass. The reported paired C–D estimate is 1.79 macro points, versus 0.42 for B–A; these estimates follow the unrounded-score convention in Appendix C.

Table 5: **Boundary-profile comparison.** Reported scores and the gradient-norm variance diagnostic. The norm-variance column is not the gradient covariance trace in Proposition 2; it is not used to estimate that proposition's variance or MSE decomposition.

Shape	HMMT Feb26	LCB v6	Macro	Norm variance
Hard cut	87.3 ± 0.4	84.6 ± 0.4	77.48 ± 0.10	2.06
Linear	88.7 ± 0.4	86.4 ± 0.4	79.17 ± 0.08	1.19
Cosine	90.2 ± 0.4	88.2 ± 0.4	80.92 ± 0.15	0.69
Cubic	90.3 ± 0.6	88.4 ± 0.4	81.08 ± 0.19	0.65

Table 6: **Strength and onset sensitivity.** Scores and generated tokens in thousands (Tok.). Default and zero-penalty accuracies reuse Table 1; their lengths reuse the 32k endpoints of Table 2; $\alpha = 1$ denotes the separate hard-cut rule.

Strength sweep: fixed $\alpha = 0.8$				Onset sweep: fixed $\beta = 0.9$			
β	LCB v6	Macro	Tok.	α	LCB v6	Macro	Tok.
0	83.4	75.55	9.59	0.5	86.1	79.68	8.3
0.5	87.5	80.46	12.1	0.8	88.4	81.08	7.47
0.8	88.2	80.97	11.5	0.9	87.7	80.76	12.4
0.9	88.4	81.08	7.47	1 (hard)	84.6	77.48	13.0
0.95	88.2	80.96	11.0				
0.99	87.9	80.80	10.8				
2	85.6	79.53	8.6				

The cosine and cubic profiles have similar displayed scores in Table 5. With z as in Eq. (3), the linear profile is z, the cosine profile is $(1 - \cos(\pi z))/2$, and the cubic profile is $3z^2 - 2z^3$; hard cut uses $\mathbf{1}\{u \geq H_x\}$. The cubic profile is 1.91 macro points above linear and 3.60 above hard cut, while its difference from cosine is 0.16 points. Its norm-variance value is also close to cosine's (0.65 versus 0.69). The reported norm-variance column is a descriptive diagnostic, not an estimate of the covariance trace or the MSE terms in Proposition 2.

Table 6 retains all strength and onset settings. At fixed $\alpha = 0.8$, the displayed macro stays within 0.28 points of the default for $\beta = 0.8$–0.99. At $\beta = 2$, the macro is 79.53 (1.55 points below the default) and the reported length is 8.6k tokens. The length summaries are not monotone in β. In the onset sweep, $\alpha = 0.8$ gives both the highest displayed macro and the shortest mean output among the tested settings. Compared with $\alpha = 0.9$, it raises the macro by 0.32 points and reduces mean length from 12.4k to 7.47k tokens. Earlier onset at $\alpha = 0.5$ lowers the macro to 79.68. Both sweeps reuse the primary-table default accuracy and the 7.47k length from Table 2.

Table 7: **MoE and dense backbone comparison.** MoE macros average the displayed task means in Table 1. Dense entries report macro mean $\pm$ SD over five runs. Macro SD uses within-run aggregation (Appendix C); contrasts precede final rounding.

Method	Qwen3.6-35B-A3B	Qwen3.6-27B
GSPO	75.55	76.91 ± 0.30
LCPO	75.97	77.51 ± 0.29
LAPO	76.53	77.82 ± 0.33
AEPO	77.32	78.02 ± 0.31
LNPO	78.82	79.29 ± 0.33
LNPO-ATC	81.08	81.06 ± 0.20
Δ(ATC–GSPO)	+5.53	+4.15

Table 8: **Reported wall time and output-volume equivalents.** Wall-hour summaries correspond to 2,000-update workloads. The MoE output-volume equivalent scales the default evaluation length by 4,096,000 response slots per run ($\mathrm{B} = 10^9$ tokens). Changes use displayed values and are descriptive; scope and units are defined in Appendix D.3.

Metric	Backbone	GSPO	ATC	Change (%)
Wall time (h)	35B-A3B	80.00	81.10	+1.4
Wall time (h)	27B dense	110.00	111.65	+1.5
Output equivalent (B/run)	35B-A3B	39.3	30.6	−22.1

5.6 Training-reward dynamics

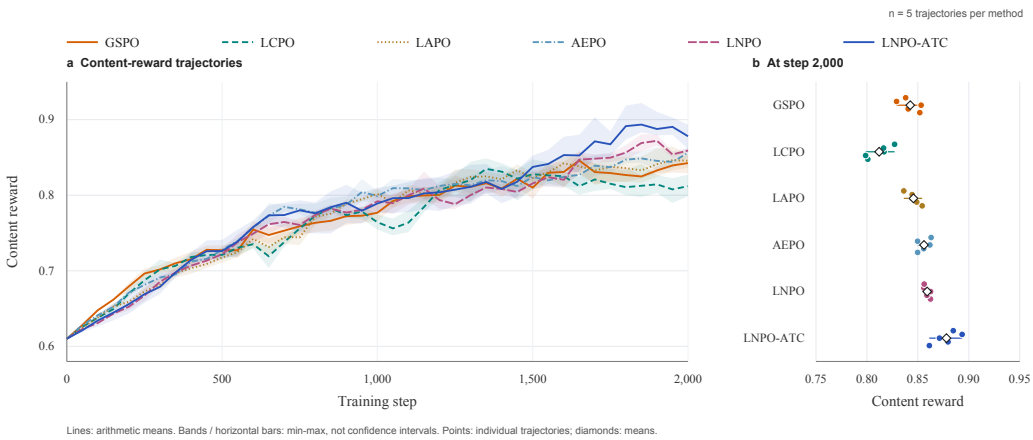

Figure 1: Training reward over five runs per method. Left: arithmetic mean reward versus optimizer updates, with pointwise min–max bands. Right: rewards at update 2,000; points show individual runs, diamonds show means, and horizontal bars show min–max ranges.

Figure 1 reports content reward during training on a $[0, 1]$ scale. Each method is summarized over five runs, and the endpoint panel uses their rewards at update 2,000. LNPO-ATC attains the highest final mean reward among the compared methods in these runs. The min–max ranges describe between-run variation. Training reward is reported separately from benchmark accuracy.

5.7 Backbone transfer and training resources

Table 7 gives descriptive ATC–GSPO differences of 5.53 and 4.15 macro points on the MoE and dense backbones, respectively. The LNPO row supplies the corresponding uniform-allocation comparison.

The displayed wall-time differences are 1.10 h for MoE and 1.65 h for dense. The MoE output-volume equivalents differ by 22.1%; they scale the default evaluation lengths by the scheduled response count. These quantities describe different comparisons. Equal update counts do not enforce equal token workloads, so the wall-time percentages are descriptive ratios, not estimates of overhead at a matched token budget (Appendix D.3).

6 Conclusion

LNPO-ATC combines smooth boundary penalties with position-dependent token credit. The analysis separates penalty strength from positional coupling and gives sufficient conditions for a positive local content-objective difference. The reported comparisons summarize task scores, allocation variants and length behavior under the default generation cap. Their interpretation distinguishes descriptive score differences from the conditional one-step result and treats EOS allocation as part of the shuffled comparison.

AI USE STATEMENT

This research and the accompanying manuscript were developed without assistance from generative AI tools. The authors conducted the research, prepared and reviewed the manuscript, and assume full responsibility for the work presented.

ETHICS STATEMENT

This work studies reinforcement learning methods for long-context reasoning using publicly available datasets. It does not involve human-subject experiments or the recruitment of human participants. Dataset and model sources are identified in the paper. As improvements in general-purpose reasoning may support both beneficial and potentially harmful applications, we encourage responsible use of the resulting methods and models.

REPRODUCIBILITY STATEMENT

We provide the `lnpo-atc-release` code package as supplementary material submitted with this paper. The core implementation is located in `lnpo_atc/`, including boundary penalties (`boundary.py`), token-level penalty allocation (`allocation.py`), policy objectives (`objective.py`), and parameter updates (`trainer.py`). Default training settings are provided in `configs/default.yaml`, with the training entry point in `scripts/train.py`. Evaluation, answer verification, and statistical aggregation are implemented in `eval/`. The script `scripts/run_local_theory.py` provides numerical checks of the local theoretical identities. The package includes usage instructions in `README.md` and software dependencies in `requirements.txt`. Algorithmic details, theoretical assumptions and proofs, and evaluation protocols are documented in the corresponding sections and appendices of the paper.

REFERENCES

Pranjal Aggarwal and Sean Welleck. L1: Controlling how long a reasoning model thinks with reinforcement learning. *arXiv preprint arXiv:2503.04697*, 2025. URL https://arxiv.org/abs/2503.04697.

Jakob Foerster, Gregory Farquhar, Maruan Al-Shedivat, Tim Rocktäschel, Eric Xing, and Shimon Whiteson. DiCE: The infinitely differentiable Monte Carlo estimator. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pp. 1529–1538, 2018. URL https://proceedings.mlr.press/v80/foerster18a.html.

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

HMMT. February 2026 tournament archive. https://www.hmmt.co/www/archive/292, 2026.

IMO-Bench Team. IMO-Bench: Towards robust mathematical reasoning. https://imobench.github.io/, 2025.

Bill Yuchen Lin, Ronan Le Bras, Kyle Richardson, Ashish Sabharwal, Radha Poovendran, Peter Clark, and Yejin Choi. Zebralogic: On the scaling limits of llms for logical reasoning. *arXiv preprint arXiv:2502.01100*, 2025.

LongBench Team. LongBench v2: Towards deeper understanding and reasoning on realistic long-context multitasks. https://longbench2.github.io/, 2024.

Zijun Min, Bingshuai Liu, Ante Wang, Long Zhang, Anxiang Zeng, Haibo Zhang, and Jinsong Su. Orchestrating tokens and sequences: Dynamic hybrid policy optimization for RLVR. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 37782–37795, July 2026. URL https://aclanthology.org/2026.findings-acl.1883/.

Qwen Team. Qwen3.6-27B model card. https://huggingface.co/Qwen/Qwen3.6-27B, 2026a. Accessed 22 September 2026.

Qwen Team. Qwen3.6-35B-A3B model card. https://huggingface.co/Qwen/Qwen3.6-35B-A3B, 2026b. Accessed 22 September 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 *First Conference on Language Modeling*, 2024.

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37, pp. 1889–1897, 2015. URL https://proceedings.mlr.press/v37/schulman15.html.

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

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Weizhou Shen, Ziyi Yang, Chenliang Li, Zhiyuan Lu, Miao Peng, Huashan Sun, Yingcheng Shi, Shengyi Liao, Shaopeng Lai, Bo Zhang, Dayiheng Liu, Fei Huang, Jingren Zhou, and Ming Yan. QwenLong-L1.5: Post-training recipe for long-context reasoning and memory management, 2025.

Prasann Singhal, Tanya Goyal, Jiacheng Xu, and Greg Durrett. A long way to go: Investigating length bias in rlhf. *arXiv preprint arXiv:2310.03716*, 2023.

Richard S. Sutton, David A. McAllester, Satinder P. Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Advances in Neural Information Processing Systems*, volume 12, 1999. URL https://proceedings.neurips.cc/paper/1999/hash/464d828b85b0bed98e80ade0a5c43b0f-Abstract.html.

Fanqi Wan, Weizhou Shen, Shengyi Liao, Yingcheng Shi, Chenliang Li, Ziyi Yang, Ji Zhang, Fei Huang, Jingren Zhou, and Ming Yan. Qwenlong-l1: Towards long-context large reasoning models with reinforcement learning. *arXiv preprint arXiv:2505.17667*, 2025.

Xingyu Wu, Yuchen Yan, Shangke Lyu, Linjuan Wu, Yiwen Qiu, Yongliang Shen, Weiming Lu, Jian Shao, Jun Xiao, and Yueting Zhuang. LAPO: Internalizing reasoning efficiency via length-adaptive policy optimization. *arXiv preprint arXiv:2507.15758*, 2025. URL https://arxiv.org/abs/2507.15758.

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

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025.

Yongcheng Zeng, Zexu Sun, Bokai Ji, Erxue Min, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, Haifeng Zhang, Xu Chen, and Jun Wang. CurES: From gradient analysis to efficient curriculum learning for reasoning LLMs. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/035ff7b446719c6d8c54e1967b784f89-Abstract-Conference.html.

Yuzhong Zhao, Yue Liu, Junpeng Liu, Jingye Chen, Xun Wu, Yaru Hao, Tengchao Lv, Shaohan Huang, Lei Cui, Qixiang Ye, Fang Wan, and Furu Wei. Geometric-mean policy optimization. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/44a1f18afd6d5cc34d7e5c3d8a80f63b-Abstract-Conference.html.

Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, et al. Group sequence policy optimization. *arXiv preprint arXiv:2507.18071*, 2025a.

Zihan Zheng, Zerui Cheng, Zeyu Shen, Shang Zhou, Kaiyuan Liu, Hansen He, Dongruixuan Li, Stanley Wei, Hangyi Hao, Jianzhu Yao, Peiyao Sheng, Zixuan Wang, Wenhao Chai, Aleksandra Korolova, Peter Henderson, Sanjeev Arora, Pramod Viswanath, Jingbo Shang, and Saining Xie. LiveCodeBench Pro: How do olympiad medalists judge LLMs in competitive programming? https://arxiv.org/abs/2506.11928, 2025b.

A Implementation and update procedure

Algorithm 1 LNPO / LNPO-ATC: update procedure

```
Sample G responses per prompt from a frozen π_{θ_0}; retain masks and EOS/truncation flags.
Compute detached A_i^C using Eq. (1) and available budgets H_x.
for each permitted optimizer pass over this rollout batch do
    Recompute current token log-probabilities and s_i using valid response tokens only.
    if sequence LNPO then
        Maximize Eq. (4) with final-length penalty −β b_α(L_i; H_x).
    else
        Build s_{i,t} with Eq. (5); maximize Eq. (6).
    end if
    Apply the configured optimizer, gradient clipping, and any shared KL term.
end for
Refresh θ_0 only before the next rollout batch.
```

The reference implementation uses contiguous valid-response masks and includes a sampled EOS. Forced termination carries zero policy score; padding never enters length or ratio reductions. Matched and shuffle rules are explicit diagnostic alternatives within $J(q)$, with the same content normalization and clipping settings.

B Core proofs and a finite-policy example

B.1 Sampling, scales, and the target

A prompt group consists of $x \sim \mathcal{D}$ and G iid responses conditional on x. Within-group gradients share $\bar{R}_x$ and σ_x and may be dependent. Expectations of single-response expressions include a uniform choice of group member. A batch contains N independent groups, with the task mixture included in $\mathcal{D}$. Any fixed reference scaling applies the same predeclared constant to both candidate gradients and their moments.

The target F is the expected unnormalized verifier reward under fixed generation-cap and forced-transition rules. On finite, parameter-independent support with differentiable positive stochastic probabilities, differentiating the finite sum yields $v = \mathbb{E}[R \sum_t S_t]$. More general models require an integrable envelope for differentiation under the expectation. Deterministic transitions have zero score. The one-step comparison additionally assumes smoothness of F.

B.2 Proof of Proposition 1

At θ_0, the numerical ratio is one and lies strictly inside the clipping interval. With detached coefficients, the local gradient of a response is

$$G_q^{(i)} = \frac{1}{L_i} \sum_t (A_i^C - \beta q_{i,t}) S_{i,t}. \tag{15}$$

Subtracting its paired uniform version cancels the content term even when the content coefficient depends on other group members. For one response,

$$G_q^{(i)} - G_c^{(i)} = \beta \left[c_i \bar{S}_i - \frac{1}{L_i} \sum_t q_{i,t} S_{i,t} \right] \tag{16}$$

$$= \beta[(c_i - \bar{q}_i)\bar{S}_i - \mathrm{Cov}_t(q_i, S_i)]. \tag{17}$$

Average over the group and its sampling law to obtain Eq. (9). If $\beta = 0$ or $q = c$ at every position, the difference is zero pathwise. Outside the warning region, monotonicity gives $c = 0$ and zero matched weights.

Fix the entire rollout group and any allocation weights. For a uniform random permutation ρ_i within response i, each coefficient visits each position with probability $1/L_i$, so $\mathbb{E}_\rho q_{i,\rho_i(t)} = \bar{q}_i$. Linearity

gives $\mathbb{E}_\rho[G_{\rho q} \mid \text{group}] = G_{\bar{q}}$, where $G_{\bar{q}}$ broadcasts the original weights' response mean. The law of total covariance gives

$$\operatorname{tr}\operatorname{Var}(G_{\rho q}) = V_{\bar{q}} + \mathbb{E}\operatorname{tr}\operatorname{Var}(G_{\rho q} \mid \text{group}) \geq V_{\bar{q}}. \tag{18}$$

Consequently, $\langle v, g_{\mathrm{ATC}} - g_{\mathrm{shuffle}} \rangle = -\beta \mathbb{E}\operatorname{Cov}_t(q, a)$, equal to the ATC-minus-uniform-mean raw margin. Original ATC recovers the uniform-at-ATC-mean gradient in permutation expectation; terminal-matched ATC recovers G_c and V_c in the corresponding terms. Permutation preserves $\sum_t q_t^2$, holding the coefficient multiset and concentration fixed. Equation (18) accounts for its additional variance in stochastic finite-step comparisons. This identity applies to old-policy raw gradients with independent permutations and a common group law.

B.3 Proof of the structural part of Theorem 1

For response-wise matched mass, project Proposition 1 onto v. Since q is deterministic given x, L, H_x, taking conditional expectation replaces a_t by μ_t, giving $M = -\beta \mathbb{E}\operatorname{Cov}_t(q, \mu)$. Let μ_P and μ_Q be the respective unweighted means on prefix and suffix. The suffix mean weight is Lc/m. Directly,

$$-\operatorname{Cov}_t(q, \mu) = c\bar{\mu} - \frac{m}{L}\left(\frac{Lc}{m}\mu_Q + \operatorname{Cov}_Q(q, \mu)\right) \tag{19}$$

$$= c(1 - m/L)(\mu_P - \mu_Q) - \frac{m}{L}\operatorname{Cov}_Q(q, \mu) \tag{20}$$

$$\geq c(1 - m/L)\delta_{x,L}. \tag{21}$$

For unpenalized strata the contribution is zero. Integrating and multiplying by β proves Eq. (11). Strict positivity requires a penalized stratum with positive probability, $\beta > 0$, and a nonempty prefix.

An alternative ordered sufficient condition follows from

$$\operatorname{Cov}_t(q, \mu) = \frac{1}{2L^2}\sum_{t,u}(q_t - q_u)(\mu_t - \mu_u). \tag{22}$$

If q is nondecreasing and μ nonincreasing, every summand is nonpositive. A pair with strict opposite differences on a positive-probability stratum gives strict positivity of M for $\beta > 0$. The prefix–suffix condition allows the weaker requirement of nonpositive within-suffix covariance. Appendix B.8 evaluates aligned and opposed examples.

Why random length matters. If length is fixed independently of policy actions and q_t depends only on the prompt and position, the conditional score identity gives $\mathbb{E}[S_t \mid x] = 0$ at each position, so the expected allocation change is zero. Policy-dependent realized length can break this cancellation, as illustrated by the length-selecting action in Appendix B.8. Original ATC additionally contributes the signed mass term $\beta \mathbb{E}[(c - \bar{q})\langle v, \bar{S}\rangle]$.

B.4 Proof of the stochastic one-step comparison

Assume F is continuously differentiable and ∇F is K-Lipschitz along every candidate update segment and the segments joining points used in the bound. The fundamental theorem of calculus gives, for every admissible displacement u,

$$|F(\theta_0 + u) - F(\theta_0) - \langle v, u\rangle| \leq \tfrac{1}{2}K\|u\|^2. \tag{23}$$

Apply the lower bound with $u = \eta\bar{G}_q$ and the upper bound with $u = \eta\bar{G}_c$, subtract, and take expectation. Finite second moments and independence of whole groups give

$$\mathbb{E}\|\bar{G}_j\|^2 = \|g_j\|^2 + V_j/N. \tag{24}$$

The expected linear difference is $\eta\langle v, d\rangle$, proving Eq. (12) for paired as well as independently sampled candidates. If $KQ_N > 0$, the stated threshold makes the lower bound positive. If $KQ_N = 0$, the quadratic remainder vanishes and $\eta M > 0$ suffices within the assumed region. When $Q_N = 0$, both updates and M are zero almost surely.

The sufficient step-size condition uses a uniform upper bound K over all possible update segments; a statistical estimate requires a valid upper-confidence bound.

Actual optimizers. Corollary 2 conditionally extends the comparison to a specified optimizer displacement through its directional correction and second moment. This mathematical extension does not assert empirical validation of optimizer dynamics.

B.5 Proof of Proposition 2 and concentration cost

From $G_q = G_c + D$, expand the covariance matrix and take its trace. The two cross-covariance traces are equal, yielding Eq. (13). Similarly, the expansion $G_j = C + \Lambda_j$ preserves the full content–penalty cross term, regardless of dependence. For N independent groups,

$$\mathrm{MSE}_{j,N} = \|g_j - v\|^2 + V_j/N. \tag{25}$$

Subtract the two expressions and expand $\|g_c + d - v\|^2$ to obtain Eq. (14). Here variance denotes the trace of the gradient covariance.

For nonnegative weights supported on m positions with $\sum_t q_t = Lc$, Cauchy–Schwarz gives

$$(Lc)^2 = \left(\sum_{t:q_t>0} q_t \right)^2 \le m \sum_t q_t^2. \tag{26}$$

Uniform allocation has squared weight sum Lc^2. For the conditional noise corollary, fix a stratum where L, c, q are deterministic and token score noises are pairwise uncorrelated with common covariance trace σ^2. The length-only response gradient then has variance trace

$$\operatorname{tr}\operatorname{Var}\left(-\frac{\beta}{L} \sum_t q_t S_t \,\middle|\, \text{stratum} \right) = \frac{\beta^2\sigma^2}{L^2} \sum_t q_t^2 \ge \frac{\beta^2\sigma^2 c^2}{m}. \tag{27}$$

For $c > 0$ and nonzero noise, this conditional variance is at least L/m times the uniform value $\beta^2\sigma^2c^2/L$. Correlated rollout gradients are evaluated with the complete covariance identity.

B.6 Supporting conditions for the allocation-to-benefit argument

The following results support Theorem 1 and specialize the same objective family. Throughout, comparisons use the original prompt law, a common reference gradient v, and detached coefficients.

Corollary 1 (Mixed-sign contributions and original ATC)**.** *Let q be deterministic on each (x, L, H_x) stratum and nonnegative, with $q = 0$ whenever $c = 0$. On exposed strata suppose it is zero on a prefix P, with a suffix Q of size m, $0 < m < L$. On unexposed strata set $Z_q = 0$. For exposed strata define*

$$Z_q = (c - \bar{q})\bar{\mu} + \bar{q}(1 - m/L)(\mu_P - \mu_Q) - \frac{m}{L}\operatorname{Cov}_Q(q, \mu). \tag{28}$$

Then $M_q = \beta\mathbb{E}Z_q = \beta(\mathbb{E}Z_q^+ - \mathbb{E}Z_q^-)$, whenever these expectations are finite. More quantitatively, let ρ be exposure probability. If, conditional on exposure, a set of probability $1 - \varepsilon$ has $Z_q \ge a > 0$ and the remainder has $Z_q \ge -b$ for $b \ge 0$, then

$$M_q \ge \beta\rho[(1 - \varepsilon)a - \varepsilon b]. \tag{29}$$

For original ATC on exposed strata with $c > 0$ and $\bar{q} > 0$, define $r = \bar{q}/c$. Its matched version satisfies

$$Z_{\mathrm{ATC}} = (c - \bar{q})\bar{\mu} + rZ_{\mathrm{match}}. \tag{30}$$

Thus matched positional benefits and the signed mass term jointly determine the margin of the implemented ATC rule.

If the prefix is empty ($m = L$), use $Z_q = (c - \bar{q})\bar{\mu} - \operatorname{Cov}_t(q, \mu)$ directly, without defining a mean on the empty prefix. The mixed-sign and ATC-transfer identities still hold; the positive-gap sufficient condition requires a nonempty prefix.

Proof. The unweighted response mean is $(1 - m/L)\mu_P + (m/L)\mu_Q$, and the suffix mean weight is $L\bar{q}/m$. Hence $-\operatorname{Cov}_t(q, \mu) = \bar{q}(1 - m/L)(\mu_P - \mu_Q) - (m/L)\operatorname{Cov}_Q(q, \mu)$. Project Proposition 1 on v and condition on the stratum to obtain Eq. (28). Splitting a real number into positive and negative parts gives the expectation identity. Integrating the two stratum lower bounds gives Eq. (29). Finally $q_{\mathrm{match}} = q/r$, so its covariance term and mean weight are both $1/r$ times the original values. This proves Eq. (30). Unexposed responses contribute zero without division. □

The tolerance bound controls adverse strata through both their frequency ε and magnitude b. Matched mass, a positive prefix gap and nonpositive suffix covariance recover the original sufficient condition. For original ATC, Eq. (30) combines this positional margin with the signed mass contribution.

Lemma 1 (A sufficient local curvature bound for the verifier objective)**.** *Consider a finite prompt pool, finite vocabulary and fixed horizon $H_{\max}$, with parameter-independent trajectory support and fixed rewards $|R| \le R_{\max}$. Let stochastic token probabilities be positive and twice continuously differentiable on a neighborhood of a compact convex region $\mathcal{U}$. On all feasible histories and $\theta \in \mathcal{U}$, suppose*

$$\|\nabla \log \pi_\theta(y_t \mid h_t)\| \le B_1, \qquad \|\nabla^2 \log \pi_\theta(y_t \mid h_t)\|_{\mathrm{op}} \le B_2.$$

With $T_\theta = \sum_t \nabla \log \pi_\theta(y_t \mid h_t)$ and $J_\theta = \sum_t \nabla^2 \log \pi_\theta(y_t \mid h_t)$, the true content objective satisfies

$$\nabla^2 F(\theta) = \mathbb{E}_{\pi_\theta}\left[R(T_\theta T_\theta^\top + J_\theta)\right], \tag{31}$$

$$\sup_{\theta \in \mathcal{U}} \|\nabla^2 F(\theta)\|_{\mathrm{op}} \le R_{\max}(H_{\max}^2 B_1^2 + H_{\max} B_2). \tag{32}$$

Consequently its gradient is Lipschitz on $\mathcal{U}$ with this sufficient constant.

Proof. Differentiate the finite sum $F = \sum_x \mathcal{D}(x) \sum_y P_\theta(y \mid x) R(x, y)$ twice. Since $\nabla P_\theta = P_\theta T_\theta$, differentiating again gives $\nabla^2 P_\theta = P_\theta(T_\theta T_\theta^\top + J_\theta)$. Forced transitions have zero score and Hessian. Summing yields Eq. (31). Triangle inequality gives $\|T_\theta\| \le H_{\max} B_1$ and $\|J_\theta\|_{\mathrm{op}} \le H_{\max} B_2$; use $\|TT^\top\|_{\mathrm{op}} = \|T\|^2$ and the reward bound. Integrating the Hessian along a segment in convex $\mathcal{U}$ gives the Lipschitz bound. □

Compactness and smooth positive probabilities give finite B_1, B_2. The lemma permits discrete verifier rewards fixed on sampled outcomes; general prompt laws require an integrable envelope. The gradient identity follows the score-function argument (Sutton et al., 1999), with the score outer product contributing to the second derivative (Foerster et al., 2018).

This local setting requires smooth computation and fixed trajectory support throughout $\mathcal{U}$, including stable expert routing and decoding support where applicable. Curvature concerns F; the cubic allocation coefficient is detached during updates. Appendix B.8 derives uniform curvature bounds for the finite policy; these conditions remain assumptions for language-model checkpoints.

Corollary 2 (Transfer to a specified optimizer displacement)**.** *At a shared checkpoint and fixed optimizer state, let the actual random displacement be $U_j = \eta \bar{G}_j + e_j$, $j \in \{c, q\}$, with finite second moments. Assume the curvature condition of Theorem 1 holds on all traversed segments. Put $E_{\mathrm{opt}} = \langle v, \mathbb{E}(e_q - e_c)\rangle$ and $Q_{\mathrm{opt}} = \mathbb{E}(\|U_q\|^2 + \|U_c\|^2)$. Then*

$$\mathbb{E}F(\theta_0 + U_q) - \mathbb{E}F(\theta_0 + U_c) \ge \eta M_q + E_{\mathrm{opt}} - \tfrac{1}{2} K Q_{\mathrm{opt}} \tag{33}$$

$$\ge \eta M_q - \|v\| \|\mathbb{E}(e_q - e_c)\| - \tfrac{1}{2} K Q_{\mathrm{opt}}. \tag{34}$$

The first lower bound is positive whenever its signed direction contribution exceeds its quadratic remainder.

Proof. Apply the lower Taylor bound to U_q and the upper bound to U_c. The expected linear difference is $\langle v, \mathbb{E}(U_q - U_c)\rangle = \eta M_q + E_{\mathrm{opt}}$. Taking expectation gives the first inequality; Cauchy–Schwarz gives the second. With $e_j = 0$, $Q_{\mathrm{opt}} = \eta^2 Q_N$ and Eq. (12) is recovered. □

This conditional bound specializes the local improvement approach in policy optimization (Schulman et al., 2015) to a specified displacement. It makes no empirical claim about a particular optimizer trajectory.

B.7 Normalization and coefficient properties

B.7.1 Normalization properties

Proposition 3 (Normalization changes the policy-gradient target)**.** *Let $S_\theta(y) = \sum_t \nabla \log \pi_\theta(y_t \mid x, y_{<t})$ and let $r(x, y)$ be a fixed, parameter-independent scalar reward. At $\theta = \theta_0$, the unclipped*

normalized surrogate with coefficient r has gradient $\mathbb{E}[rS/L]$, equal to $\nabla\mathbb{E}_{\pi_\theta}[r/L]$. The original expected-reward gradient is $\mathbb{E}[rS]$. For fixed reference length $h > 0$, their rescaled difference is

$$B_h = \mathbb{E}\left[\left(\frac{h}{L} - 1\right) rS\right], \qquad \|B_h\| \le \sqrt{\mathbb{E}[(h/L-1)^2]\mathbb{E}[r^2\|S\|^2]}. \tag{35}$$

The following derivation treats the additional coupling from group-normalized coefficients, random-length baselines and off-policy weighting, and evaluates bias and variance against a common target.

Proposition 4 (Conditional variance under weak token dependence)**.** *Fix a response length L and put $Z_t = AS_t$, where $S_t = \nabla \log \pi_{\theta_0}(y_t \mid x, y_{<t})$ and A is the common, possibly trajectory-dependent coefficient. Suppose $|\operatorname{tr}\operatorname{Cov}(Z_t, Z_u \mid L)| \le \sigma^2\kappa(|t-u|)$, where $\kappa(0) = 1$, $\kappa \ge 0$, and $C_\kappa = 1 + 2\sum_{k\ge1}\kappa(k) < \infty$. Then*

$$\operatorname{tr}\operatorname{Var}\left(\frac{1}{L}\sum_t Z_t \,\middle|\, L\right) \le \frac{\sigma^2 C_\kappa}{L}. \tag{36}$$

The bound applies at the old policy, where $s = 1$ and clipping is inactive, under the stated covariance summability condition. For random length, total variance additionally contains $\operatorname{tr}\operatorname{Var}(\mathbb{E}[L^{-1}\sum_t Z_t \mid L])$, which accounts for between-length variation. This local normalization property is shared with GSPO.

B.7.2 Relation to the original policy gradient

For a finite trajectory with positive probability, write $P_\theta(y \mid x)$ for its autoregressive probability, including any sampled EOS and excluding forced non-actions. Under an integrable score,

$$\nabla_\theta \sum_y P_\theta(y \mid x) r(x,y) = \sum_y P_\theta(y \mid x) r(x,y) \nabla_\theta \log P_\theta(y \mid x) = \mathbb{E}[rS_\theta]. \tag{37}$$

Because sampled $L(y)$ is part of the outcome, the same argument with reward r/L yields $\nabla\mathbb{E}[r/L] = \mathbb{E}[rS_\theta/L]$. At θ_0, $s = 1$ and $\nabla s = S_{\theta_0}/L$; for a clipping interval with 1 strictly inside, clipping is inactive locally. Subtracting the original gradient from h times this normalized gradient gives Eq. (35). Triangle inequality followed by Cauchy–Schwarz gives its norm bound.

Baselines under random length. For a prompt-only baseline $b(x)$, $\mathbb{E}[S_\theta \mid x] = 0$, so subtracting $b(x)$ in ordinary REINFORCE does not change the target. But

$$\mathbb{E}[b(x)S_\theta/L \mid x] = b(x)\nabla_\theta \mathbb{E}[1/L \mid x], \tag{38}$$

When the policy affects termination, this baseline contribution is generally nonzero. A sample also participates in its own group mean and standard deviation in Eq. (1). Detachment specifies the resulting estimator, including those shared statistics. A leave-one-out baseline removes self-inclusion in the mean; random-length division and reward scaling remain part of the target analysis.

Away from the old policy. Exact off-policy correction uses $W = P_\theta/P_{\theta_0} = s^L$. The derivative of the unclipped GSPO-style surrogate is $\mathbb{E}_{\theta_0}[s\, rS_\theta/L]$, whereas the exact gradient of $\mathbb{E}_{\pi_\theta}[r]$ is $\mathbb{E}_{\theta_0}[W\, rS_\theta]$. Consequently even a constant scaling does not generally recover the exact gradient. For an estimator $\hat{g}$ and deterministic target $g_\star$,

$$\mathbb{E}\|\hat{g} - g_\star\|^2 = \|\mathbb{E}\hat{g} - g_\star\|^2 + \operatorname{tr}\operatorname{Var}(\hat{g}). \tag{39}$$

B.7.3 Conditional covariance proof

Conditional on length L, covariance is bilinear:

$$\operatorname{tr}\operatorname{Var}\left(L^{-1}\sum_t Z_t \mid L\right) = L^{-2}\sum_{t,u}\operatorname{tr}\operatorname{Cov}(Z_t, Z_u \mid L) \tag{40}$$

$$\le \frac{\sigma^2}{L^2}\left[L + 2\sum_{k=1}^{L-1}(L-k)\kappa(k)\right] \le \frac{\sigma^2 C_\kappa}{L}. \tag{41}$$

The covariance condition applies to the weighted scores $Z_t = AS_t$. With perfect common noise $Z_t = Z$, the average equals Z and has constant variance, illustrating the role of summable dependence.

For random L, total covariance adds a between-length term. A weighted token sum with summable covariance depends on $\sum_t a_t^2$ and the correlations. Concentrating conserved penalty mass on a shorter suffix increases the squared-weight lower bound; the covariance identity determines the resulting total variance.

B.7.4 Boundary sensitivity proof

For fixed $H > 0$ and $0 \le \alpha < 1$, let $w = (1-\alpha)H$ and $f(z) = 3z^2 - 2z^3$ on $[0,1]$, extended by zero and one to the left and right. On $0 < z < 1$, $f'(z) = 6z(1-z)$ and $f''(z) = 6 - 12z$, so $0 \le f'(z) \le 3/2$ and $|f''(z)| \le 6$. The outer pieces match $f'(0) = f'(1) = 0$. The chain rule and mean-value inequality therefore give, for any real u, u',

$$|b_\alpha(u;H) - b_\alpha(u';H)| \le \frac{3}{2w}|u-u'|, \qquad |b'_\alpha(u;H) - b'_\alpha(u';H)| \le \frac{6}{w^2}|u-u'|.$$

Thus the barrier is globally C^1, with the stated function and derivative Lipschitz constants; the bounds hold across the joins and at integer lengths.

For $0 \le d \le w$ at onset, $b_\alpha(\alpha H + d; H) = 3(d/w)^2 - 2(d/w)^3 \le 3(d/w)^2$. The cubic therefore has quadratic onset, compared with the linear ramp's d/w, while its slope is larger in the middle of the warning zone.

These inequalities describe sensitivity to the length argument. The one-step policy theorem separately assumes Lipschitz continuity of ∇F along its update segments. Clipping and adaptive optimizer state enter through the actual update direction.

B.7.5 Coefficient reversal and prefix protection

For a uniform sequence coefficient, $A_i^C > 0$ and $\beta b(L_i) > A_i^C$ give $A_i^C - \beta b(L_i) < 0$ directly. In the active branch of the surrogate, the sign of the scalar coefficient attached to every token score is reversed. For $t \le \alpha H$, $b(t) = 0$ and ATC's coefficient remains A_i^C. The mass-matched rule has the same zero prefix weights. If $\sum_t b(t) = 0$, monotonicity and inclusion of the final position imply $b(L) = 0$, so its zero case is consistent. Otherwise Eq. (7) sums to $Lb(L)$, as required.

Coefficient-level interpretation. The zero-prefix property concerns the explicit length coefficient. Token probabilities after an update depend on the combined batch gradient through shared parameters, while content coefficients use sequence-level verifier outcomes. Prefix-level behavior can be evaluated through verified continuations. For an action-independent positional cost c_t, an ordinary unnormalized on-policy update satisfies $\mathbb{E}[c_t \nabla \log \pi(a_t \mid s_t) \mid s_t] = 0$. The stopping effect of the full surrogate additionally depends on length normalization, future-dependent coefficients, clipping, and terminal censoring.

B.8 A finite-policy example

Let $p = \sigma(\theta_1)$ and $r = \sigma(\theta_2)$. The policy samples $z \sim \mathrm{Bernoulli}(p)$ at position 1, sets $L = 4 + 4z$, and samples $u \sim \mathrm{Bernoulli}(r)$ at position L, independently of z. Intermediate tokens are deterministic; termination is forced. Thus there are four trajectories with scores $S_1 = (z-p, 0)$ and $S_L = (0, u-r)$; filler tokens and forced termination have zero score. Filler tokens count toward length.

Table 9 compares mass-matched and uniform allocation with $H = 8$, $\alpha = 0.5$, $\beta = 0.9$, $\delta = 10^{-6}$ and group size $G = 2$. Enumerating all 16 ordered iid response groups includes their exact shared content normalization. One group is used per update ($N = 1$), with $\eta = 0.1$. The first three reward cases use $(p, r) = (0.5, 0.7)$; the mixed case uses $(0.5, 0.8)$.

For $L = 8$, the original suffix weights are $(5/32, 1/2, 27/32, 1)$; scaling by $16/5$ matches the uniform response mean of one. All weights are zero for $L = 4$. For an additive reward $R = az + bu + d_0$, $F = ap + br + d_0$ and $v = (ap(1-p), br(1-r))$. Define $v_1 = \partial F/\partial\theta_1 = ap(1-p)$, the first coordinate of the content gradient; the first-token projection is $\langle v, S_1\rangle = v_1(z-p)$.

Conditional on $L = 8$, the four-position prefix has mean projection $\mu_P = v_1(1-p)/4$. Independence of u and z gives $\mu_Q = 0$ and $\mathrm{Cov}_Q(q, \mu) = 0$ on the four-position suffix. Since $P(L = 8) = p$, the exact margin is

$$M = \beta p\Big(1 - \frac{4}{8}\Big)(\mu_P - \mu_Q) = \frac{\beta p(1-p)v_1}{8}.$$

For $R = z$, this becomes $M = \beta[p(1-p)]^2/8$. At $p = 0.5$, the prefix–suffix gap is $1/32$ and $M = 0.00703125$; at $p = 0.2$ or 0.8, $M = 0.00288$. Reversing the reward reverses the margin; the flat case shows that zero margin need not give identical finite-step outcomes.

Table 9: **Executed exact finite-policy expectations.** Matched versus uniform allocation, one independent two-response group, $\eta = 0.1$. M uses the true content gradient; the final column is the proved lower bound.

Reward	M	ΔF	ΔV	Lower bound
Prefix aligned ($R = z$)	0.007031	0.000703	0.004120	0.000683
Flat ($R = u$)	0.000000	-0.000003	0.007043	-0.000028
Opposed ($R = 1 - z$)	-0.007031	-0.000703	0.007368	-0.000737
Mixed ($R = .25z + .75u$)	0.001758	0.000173	0.006163	0.000154

For these additive rewards, the diagonal Hessian and $|\sigma''| \le 1/4$ give the global bound $K = \max(|a|, |b|)/4$: $1/4$ for the first three cases and $3/16$ for the mixed case. This bound applies along every update segment in the reported one-step comparison. Positive margins alone do not guarantee improvement at arbitrary step sizes. Also, for $R = z$, both content reward and expected boundary cost equal p, so improving content need not reduce boundary cost.

The scope of these calculations is the specified finite policy, whose expectations and curvature bounds are available exactly. Applicability to another policy is governed by the structural and smoothness assumptions in the theorem.

In the submitted `lnpo-atc-release` package, `scripts/run_local_theory.py` checks the mass–location and permutation identities and enumerates the finite-policy mean gradients and margins. Its `--out` option selects the output directory for numerical results and a configuration/source-hash manifest.

C Evaluation definitions, aggregation and protocol

Task metrics and reference scores. For the core suite in Section 5.1, scores are percentages: exact-answer verification for mathematics, test-based pass@1 for code, whole-puzzle accuracy for ZebraLogic, and answer-choice accuracy for GPQA. IMO-AnswerBench uses verifiable short answers; LCB v6 and LCB Pro are evaluated separately. Base denotes the released post-trained checkpoint before additional RL. Published Base (ref.) cells retain model-card protocols (Qwen Team, 2026b;a) and are excluded from shared-protocol contrasts.

Run-level aggregation. For method a, independent training run s and core task j, define

$$M_{as} = \frac{1}{6}\sum_{j=1}^{6} Y_{asj}.$$

For $S = 5$ independent training runs, the macro mean and sample SD are

$$\bar{M}_a = \frac{1}{S}\sum_{s=1}^{S} M_{as}, \qquad s_{M,a} = \sqrt{\frac{1}{S-1}\sum_{s=1}^{S}(M_{as} - \bar{M}_a)^2}.$$

For the descriptive macro summaries of Table 1, average its displayed task means without intermediate rounding, compute contrasts, and round the final quantities to two decimal places. Task-level SDs use the sample-SD formula before rounding. Computing the macro within each run retains cross-task covariance; its SD is not the average of task-level SDs. Decoding repeats and checkpoints are nested within each run. LongBench uses the within-run question-count weights in Appendix D.2.

Paired contrasts. Paired analysis uses randomization blocks fixed before training. For n paired runs with differences d_s, the two-sided 95% interval is $\bar{d} \pm t_{n-1,0.975} s_d/\sqrt{n}$. Independently sampled runs use independent-sample analysis. Compute each contrast and its interval from unrounded paired run-level scores, then round each displayed estimate once to two decimal places. The intervals are pointwise. The within-run macro and paired-difference calculations preserve cross-task and cross-method dependence; separate task SDs are not used to reconstruct these intervals.

C.1 Data allocation and source scope

Table 10 states the domain allocation and source partitions for the comparison.

Table 10: **Training-pool allocation.** Allocation under the protocol in Section 5.1; validation and reward-monitoring partitions are additional and disjoint.

Domain	Source and allowed partition	Prompts	Share
Math	NuminaMath 1.5, `train`; verifiable final answers only	32,000	40%
Coding	TACO, `ALL/train`, and CodeContests, `train`; deduplicated across sources	24,000	30%
Logic	SynLogic, `hard` training partition; exclude Zebra/Einstein task families	12,000	15%
Science	OpenScienceReasoning-2: 8,000 natural-science questions; ARC: 2,000 from Easy/Challenge training splits; OpenBookQA: 2,000 from `main/train`	12,000	15%
Total	Unique training prompts; reference solutions and repeated rollouts do not add unique questions	80,000	100%

Input and partition definitions. Inputs consist of the question, template, answer options or starter code; reference answers and hidden tests belong to the verifier. The input-length limit is 4,096 tokens. Uniform sampling with replacement from the training allocation gives domain probabilities of 40/30/15/15. The separate validation allocation is 4,000 questions (1,600/1,200/600/600), and the reward-monitoring allocation is 2,000 (800/600/300/300); together with training, these allocations total 86,000 questions.

C.2 Optimization settings and baseline tuning

The LNPO-ATC defaults are AdamW, learning rate 10^{-6}, 128 prompts per update, 16 responses per prompt, 2,000 updates and a 32,768-token output cap. ATC uses $(\alpha, \beta) = (0.8, 0.9)$. The LNPO and ATC objectives, content normalization and policy-ratio clipping are defined in Section 3.

PPO, GRPO, DAPO, GSPO, LCPO, LAPO and AEPO were each subjected to method-specific hyperparameter tuning. Each baseline is evaluated using its own tuned configuration, including the method-specific settings relevant to its objective. Table 1 reports the resulting comparisons. LCPO and LAPO provide length-control comparisons, and AEPO denotes its optimization component. Table 6 is a separate sensitivity analysis of LNPO-ATC's boundary parameters; the baseline comparisons use the tuned configurations described above.

C.3 Evaluation definitions

Core metrics use exact-answer verification for mathematics, test-based pass@1 for coding, whole-puzzle accuracy for ZebraLogic and answer-choice accuracy for GPQA. Incomplete or unparseable responses receive zero and remain in the denominator. Published Base (ref.) scores retain their source protocols and remain outside shared-setting contrasts.

The LongBench table is an internal comparison on a custom 330-question subset. The input-bin counts are fixed at 60, 90, 100 and 80 for the displayed aggregation; Appendix D.2 defines the bins and weighting. The output allowance is 32,768 tokens and is separate from input length. The reported decoding settings are temperature 0.6 and top-p 0.95, with one response per question for each run. These subset scores are not full-release LongBench v2 scores.

D EVALUATION METRICS AND RESOURCE ACCOUNTING

D.1 GENERATED LENGTH, TRUNCATION AND AGGREGATION

The evaluation uses a fixed 32,768-token output cap, with decoding and verification specified in Appendix C.3. Generated length counts tokens sampled from the model up to completion or the enforced cap, including a sampled EOS. It excludes prompt tokens, padding and externally inserted control text; model-generated control tokens count. Forced stopping adds no token. Truncated, failed and unparseable attempts remain in all metric denominators; a cap-stopped response contributes its actual emitted length. Truncation denotes an output-cap stop before completion. Incomplete or unparseable answers receive zero accuracy.

Within each run, accuracy, mean generated length and truncation are computed per task and averaged with equal task weights. Length and truncation summaries average these run-level metrics, using decimal thousands of tokens and percentages. The default Macro column reuses the task means in Table 1. The same token-counting rules apply to Tables 2 and 6.

D.2 LONGBENCH INPUT BINS AND AGGREGATION

Input length counts the complete rendered prompt, including instructions and answer choices. The disjoint bins are $(0, 16384]$, $(16384, 32768]$, $(32768, 65536]$, and $(65536, 131072]$. Complete inputs and the output allowance must fit the native context window. Bin membership is fixed before model evaluation.

For run s, the weighted score is

$$Y_s^{\mathrm{LB}} = \frac{\sum_b n_b Y_{sb}}{\sum_b n_b},$$

where n_b is the number of questions in bin b and Y_{sb} is its score in that run. The overall mean and sample SD are computed after forming Y_s^{LB} for each run. The counts and the fixed output allowance are given in Table 3; its subset score is separate from a full-release LongBench v2 score.

D.3 RESOURCE QUANTITIES AND COMPARISON SCOPE

Table 8 lists the reported wall-hour summaries for four 2,000-update workloads. The change column is $100(T_{\mathrm{ATC}}/T_{\mathrm{GSPO}} - 1)$ within each backbone. It describes the displayed durations at equal update counts; it is not a comparison at equal token consumption or a study-wide cost estimate. The wall-time entries carry no run-level uncertainty estimate.

Output-volume equivalent. The schedule contains $2{,}000 \times 128 \times 16 = 4{,}096{,}000$ response slots per run. Multiplying this count by the default evaluation lengths, 9.59k and 7.47k tokens, gives 39.3B and 30.6B tokens after rounding. The table calls these values output-volume equivalents: they are length-scaled workload quantities, not independent training-token counter measurements. Their 22.1% difference is computed from the displayed output-volume values. Evaluation length, generated training tokens and tokens processed by the optimizer represent different quantities.

The same schedule contains 256,000 prompt draws, or 3.2 draws per item of the 80,000-question allocation on average. At the 32,768-token cap, its response slots have a maximum output capacity of 134,217,728,000 tokens. This capacity is determined by the configuration and is not observed token consumption.

The resource comparison is restricted to the quantities in Table 8; it does not report peak memory, allocation-derived GPU-hours or total study cost.