LEARNING WHERE TO CONTINUE: BUDGETED CORRECTION OF CENSORED POLICY GRADIENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

In language-model training, reinforcement learning with verifiable rewards improves reasoning through outcome-based feedback. Generation limits leave some trajectories unfinished, while completing every rollout is costly. Imputation provides low-cost updates but leaves open where further generation most reduces gradient estimation error. We propose a budgeted continuation framework that selects rollout groups using joint prefix–tail residual risk and corrects their imputed updates using selected completions. Inverse-probability weighting over all arriving groups accounts for selective feedback when learning this risk. We establish unbiasedness for the unclipped, finite-horizon gradient of a frozen policy, derive the optimal allocation under an expected budget constraint, and bound the excess risk of its learned counterpart. An exact autoregressive construction shows that prefix–tail coupling can make allocation based on segment-wise second moments strictly suboptimal, even with a perfectly calibrated imputation head. Across six benchmarks in mathematics, coding, knowledge and long-context reasoning, joint-residual allocation increases mean accuracy by 6.27 and 6.18 percentage points on Qwen3.6-27B and Qwen3.6-35B-A3B, respectively, and reduces gradient correction error by 24.67% and 28.02% relative to fixed continuation at the same training budget. In ablations, accounting for prefix–tail coupling and correcting selective feedback each yields lower correction error and higher mean accuracy on both dense and mixture-of-experts models.

1 INTRODUCTION

Training reasoning language models with verifiable rewards relies on the outcomes of generated solutions (Guo et al., 2025; Jaech et al., 2024; Team, 2025). A generation cap C controls training cost but leaves some trajectories unfinished. Their prefixes reveal neither a terminal answer nor its reward, although the desired objective concerns success by a longer horizon H. Completing every trajectory provides this feedback at additional cost; imputing missing outcomes gives an inexpensive update whose error depends on the unobserved continuation.

Existing approaches improve incomplete-rollout updates or adapt generation budgets (§2). We ask which unfinished groups to continue when the value of that decision must itself be learned from selectively revealed outcomes.

Figure 1 summarises the framework. Every group first contributes an imputed update. A learned allocator selects some groups for continuation to H, and an inverse-probability-weighted residual corrects their updates. The newly revealed residuals update risk predictions for later batches. This control-variate construction (Horvitz & Thompson, 1952; Robins et al., 1994; Owen, 2013) preserves the finite-horizon gradient in expectation; the learning problem is to spend the continuation budget where the correction is most valuable.

Two structures determine that value. First, the residual contains interacting prefix and tail components. Their separate second moments can be identical while the joint residual moments differ, even with a calibrated imputation head (Theorem 2). Second, only selected groups reveal a residual. Learning from those groups alone can reverse the risk ranking (Proposition 2); averaging inverse-probability-weighted contributions over all arrivals corrects this feedback.

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.

Contributions. Our contributions are threefold:

1. **A budgeted framework for correcting censored policy gradients.** We combine imputed updates with selectively acquired completions to obtain an unbiased estimator of the frozen-policy, unclipped finite-horizon gradient, and derive its exact correction risk under a continuation budget (§4.1).
2. **Joint-residual allocation that captures prefix–tail coupling.** We identify the joint residual second moment as the quantity governing optimal continuation and establish that segment-wise second moments alone are insufficient. An exact autoregressive construction establishes a strict risk gap between segment-moment and joint-residual allocation at the same budget (§4.2–4.3).
3. **Risk learning from selective continuation feedback.** We construct an all-arrival, inverse-probability-weighted risk estimator that corrects the bias of learning only from continued groups, and bound the excess correction risk of the resulting learned allocation (§4.4).

Experimental validation. Across six mathematics, coding, knowledge and long-context reasoning benchmarks, we evaluate Qwen3.6-27B and Qwen3.6-35B-A3B with 64 million training tokens and five seeds. Relative to fixed continuation, mean accuracy improves by 6.27 and 6.18 percentage points, shared-checkpoint correction error falls by 24.67% and 28.02%, and all-in GPU time is $1.065\times$ and $1.075\times$, respectively. Ablations isolate coupling and feedback correction through risk-prediction error, correction error and task performance (§5); costs include residual computation, caching and risk learning.

2 Related work

2.1 Censored feedback in language-model reinforcement learning

GRPO (Shao et al., 2024) centres rewards within a group instead of learning a critic (Schulman et al., 2017); GSPO (Zheng et al., 2025) normalises sequence ratios; DAPO (Yu et al., 2025) addresses zero-advantage groups and noisy truncated rewards through dynamic sampling and overlong masking. Generation caps nevertheless leave outcomes unobserved for trajectories that could finish within a longer horizon: a Type-I censoring problem (Kalbfleisch & Prentice, 2002). Boundary-value bootstrapping (Pardo et al., 2018) and advantage estimation from unfinished responses (Fan et al., 2025) improve the feedback available without completing every rollout. We study which missing outcomes to acquire under a budget, valuing completions by their correction of the imputed update and preserving the frozen-policy, unclipped gradient (Theorem 1).

2.2 Budget allocation for gradient correction

Variance-guided methods such as VIP (Nguyen et al., 2026), DynaMO (Fang et al., 2026) and VIGOR (Jiang et al., 2026) adapt rollout counts to prompt-level learning signals, while TRACE (Zou et al., 2026) allocates budget across prompts and intermediate prefixes. PAIR (Nomand et al., 2026) corrects pair gradients through joint inclusion probabilities and optimises a graph-coupled pair-energy/covariance surrogate. Our whole-group continuation formulation yields an exact correction-risk objective governed by the conditional second moment of the joint prefix–tail update residual. We show that separate segment moments can conceal different joint risks even with a calibrated prediction head (Theorem 2), derive the budget-optimal allocation, and quantify the cost of ignoring this coupling (§4.2–4.3).

2.3 Learning continuation risk from selective feedback

Inverse-probability weighting and regression imputation provide classical tools for estimation from incomplete data (Horvitz & Thompson, 1952; Robins et al., 1994; Little & Rubin, 2019). PAIR (Nomand et al., 2026) also uses IPW for selectively observed predictor labels. Here, continuation reveals the squared joint update residual, whose prediction governs later selection. Learning from continued groups alone can reverse risk rankings (Proposition 2). We analyse all-arrival IPW learning with predictions fixed before selection and bound the induced excess correction risk (Theorem 3). Matched-

log and online ablations (§5.4) isolate feedback learning from the choice of joint or segment-wise targets.

3 BUDGETED CONTINUATION OF CENSORED ROLLOUTS

Objective and observation model. For a prompt x, an autoregressive policy π_θ emits tokens until a terminal symbol at the natural stopping time T. Fix a finite target horizon H and let $Y = \mathbf{1}[T \le H,\ \text{answer correct}]$, so $J^H(\theta) = \mathbb{E}[Y]$. A rollout that overruns H has $Y = 0$ by definition. Training runs at a smaller budget $C < H$, giving $\delta = \mathbf{1}[T \le C]$. Write the trajectory score as $s = s^{\mathrm{P}} + s^{\mathrm{T}}$, the prefix part observed at C and the tail part beyond it; for a finished rollout $s^{\mathrm{T}} = 0$. Throughout, rewards are outcome-only: if a process reward model (Lightman et al., 2023) scores prefixes, Y is partially observed on censored rollouts and this section does not apply.

The estimand. A group is $G \ge 2$ rollouts drawn independently from the frozen π_θ. With leave-one-out baselines $A_i = Y_i - \frac{1}{G-1}\sum_{j \ne i} Y_j$, the complete-group update is

$$g^\star = \frac{1}{G}\sum_{i=1}^{G} A_i\, s_i, \qquad \text{and} \qquad \mathbb{E}[g^\star] = \nabla_\theta J^H(\theta), \tag{1}$$

because $\mathbb{E}[s_i] = 0$ and Y_j is independent of s_i for $j \ne i$. This is the estimand: what an unlimited budget would have produced.

The cheap surrogate. At C only prefixes are available. A frozen head supplies $\hat{v}_i$ for censored members; let $\tilde{Y}_i = \delta_i Y_i + (1-\delta_i)\hat{v}_i$ and $s_i^{\mathrm{obs}} = \delta_i s_i + (1-\delta_i)s_i^{\mathrm{P}}$. Substituting into equation 1 gives h, computable with no extra generation. It is not unbiased, and its error is the object of this paper.

Budgeted correction. Let $\mathcal{F}$ collect all batch prefixes and the pre-batch frozen state. Choose an $\mathcal{F}$-measurable $\nu \in [\alpha, 1]$, draw one coin $Z \sim \mathrm{Bern}(\nu)$ independently of the continuations given $\mathcal{F}$, and if $Z = 1$ resume *every* censored member of that group to H, which reveals $g^\star$. The update is

$$\boxed{\hat{g} = h + \frac{Z}{\nu}\big(g^\star - h\big).} \tag{2}$$

A whole-group coin reveals the complete residual required by equation 2. Member-level designs can also be unbiased with appropriate joint inclusion weights (Nomand et al., 2026); they require a different estimator and allocation analysis. The capacity charged is

$$c = \#\{i : \delta_i = 0\} \cdot (H - C), \tag{3}$$

which is known before the coin is drawn. A group with nothing censored has $c = 0$ and $h = g^\star$ already.

3.1 LEARNING AND APPLYING CONTINUATION PRIORITIES

The allocator estimates residual risk from earlier arrivals and uses its ratio to cost to prioritise correction (Figure 1). §4 derives the rule; Algorithm 1 in Appendix G specifies the update order.

Budget and statistical memory. The budget is an external resource constraint. The sole independently varied allocation setting is the per-bin arrival capacity N, with shared default $N = 256$. The fixed rule $\alpha = \min\{\beta, N^{-1/2}\}$ aims at $\sqrt{N}$ expected continued arrivals per full bin window when the budget permits; actual counts $N_k \le N$ determine the available history. The condition $\alpha \le \beta$ ensures feasibility. Very small budgets can leave little room for adaptation. Bins are fixed by task type and censoring count. An empty bin borrows the pooled mean of retained historical IPW records; until positive estimates exist, allocation is uniform at the same budget. Head, optimiser and decoding settings are shared across arms and documented in the evaluation protocol.

Snapshots and delayed commits prevent a coin from depending on its own outcome. Feasibility is checked before selection, and every selected continuation completes. Appendix G gives the complete update procedure.

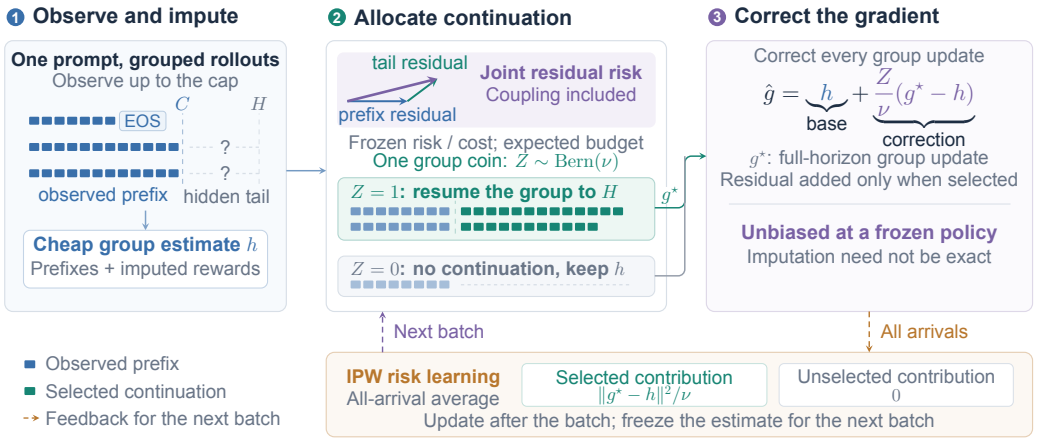


Figure 1: **Budgeted correction with delayed risk learning.** Prefixes up to C yield h. Frozen joint-residual estimates and costs determine ν under an expected budget. One coin per censored group selects all its unfinished members for continuation to H, revealing $g^\star$. The correction is unbiased for the frozen-policy, unclipped target (Theorem 1). The IPW module averages selected contributions $\|g^\star - h\|^2/\nu$ and unselected zeros over all arrivals in each retained bin window, as in equation 9. Updates enter subsequent batches only. Group subscripts are omitted.

4 Why joint residuals determine continuation

4.1 An exact risk identity

Definition 1 (Residual moment). $m = \mathbb{E}\big[\|g^\star - h\|^2 \,\big|\, \mathcal{F}\big]$.

Theorem 1 (Correctness and exact selection cost). *Let $\nu > 0$ be $\mathcal{F}$-measurable and $Z \perp g^\star \mid \mathcal{F}$. Then, for any h that is $\mathcal{F}$-measurable:* (i) $\mathbb{E}[\hat{g} \mid \mathcal{F}] = \mathbb{E}[g^\star \mid \mathcal{F}]$, *hence* $\mathbb{E}[\hat{g}] = \nabla_\theta J^H(\theta)$; (ii) $\mathbb{E}\big[\|\hat{g} - g^\star\|^2 \mid \mathcal{F}\big] = (\nu^{-1} - 1)\, m$; (iii) *for B independent identically distributed complete groups, draw selection coins independently of all potential tails and of each other given $\mathcal{F}$. With $\hat{g} = B^{-1}\sum_b \hat{g}_b$,*

$$\mathbb{E}\big\|\hat{\bar{g}} - \nabla_\theta J^H\big\|^2 = \underbrace{\frac{1}{B}\,\mathrm{tr}\,\mathrm{Cov}(g^\star)}_{\textit{irreducible}} + \underbrace{\frac{1}{B^2}\sum_{b=1}^{B}\mathbb{E}\big[(\nu_b^{-1} - 1)\, m_b\big]}_{\textit{selection term}}. \tag{4}$$

Three consequences shape the rest of the paper. Unbiasedness is unconditional in the head, so a badly calibrated head costs capacity, not correctness. The head enters only through m, so improving imputation and allocating capacity are complementary. And the *only* allocation-dependent quantity is $\sum_b(\nu_b^{-1} - 1)m_b$, so the design problem is fully specified by m and c.

Remark 1 (Scope). *Theorem 1 characterises the conditional mean and risk of a single unclipped score-function update at a frozen policy, with the target defined in equation 1. Other update objectives are treated separately in Appendix F.5.*

4.2 The capacity-constrained oracle allocation

Dropping the constant $\sum_b m_b$ from equation 4, and writing $b_{\max}$ for the extra capacity a batch may spend, the design problem is

$$\min_{\nu} \sum_b \frac{m_b}{\nu_b} \qquad \text{s.t.} \qquad \sum_b \nu_b c_b \le b_{\max}, \quad \alpha \le \nu_b \le 1. \tag{5}$$

Proposition 1 (Clipped square-root allocation)**.** *equation 5 is convex and feasible exactly when* $\alpha \sum_b c_b \le b_{\max}$*. Set* $\nu_b = 1$ *if* $c_b = 0$*, and choose* $\nu_b = \alpha$ *if* $c_b > 0, m_b = 0$*. For the remaining coordinates and* $\lambda > 0$*,*

$$\nu_b(\lambda) = \min\Big\{1, \max\big\{\alpha, \sqrt{m_b/(\lambda c_b)}\big\}\Big\}. \tag{6}$$

Define $\nu_b(0) = 1$ *on positive-cost, positive-moment groups and* $S(\lambda) = \sum_b c_b \nu_b(\lambda)$*. If* $b_{\max} \ge S(0)$*, take* $\lambda = 0$*; unused capacity is permitted. Otherwise choose any multiplier satisfying* $S(\lambda) = b_{\max}$*. Such a multiplier exists and bisection applies. Positive-moment coordinates have a unique optimum, but flat parts of* S *can make the multiplier non-unique; they yield the same allocation.*

The exploration floor $\alpha > 0$ gives every eligible group a positive selection probability and bounds the inverse weights used in residual estimation and the excess-risk analysis (§4.4). Feasibility is checked before selection, and every selected continuation completes, preserving the Bernoulli sampling rule in Theorem 1.

4.3 Segment marginals do not determine the allocation

Everything now rests on m. The natural attempt is to predict it from separate statistics of the two segments. Write $\epsilon_j = Y_j - \tilde{Y}_j$, which is zero on finished members, and $\beta_j = G^{-1}\big(s_j^{\mathrm{obs}} - \frac{1}{G-1}\sum_{i\neq j} s_i^{\mathrm{obs}}\big)$.

Lemma 1 (Residual decomposition)**.** $g^\star - h = R^{\mathrm{P}} + R^{\mathrm{T}}$ *with* $R^{\mathrm{P}} = \sum_j \epsilon_j \beta_j$ *and* $R^{\mathrm{T}} = G^{-1}\sum_{i:\delta_i=0} A_i\, s_i^{\mathrm{T}}$*, hence*

$$m = \underbrace{\mathbb{E}[\|R^{\mathrm{P}}\|^2 \mid \mathcal{F}]}_{\textit{prefix marginal}} + \underbrace{\mathbb{E}[\|R^{\mathrm{T}}\|^2 \mid \mathcal{F}]}_{\textit{tail marginal}} + \underbrace{2\,\mathbb{E}[\langle R^{\mathrm{P}}, R^{\mathrm{T}}\rangle \mid \mathcal{F}]}_{\textit{cross term}}. \tag{7}$$

Shared score coordinates let the cross term change the optimal allocation.

Theorem 2 (Insufficiency of segment-wise second moments)**.** *There is a two-parameter autoregressive instance with* $G = 2$ *and a perfectly calibrated head whose two prefix realisations have identical segment second moments* $(65/192, 1)$*, cross terms* $-5/8$ *and* $+5/8$*, and joint moments* $m_1 = 137/192$ *and* $m_2 = 377/192$*. For two equal-cost groups with an expected budget of one continuation and* $\alpha \le 0.376$*, a common rule based on segment moments and cost has best allocation* $(1/2, 1/2)$*, which is strictly suboptimal.*

For probabilities $(\nu, 1-\nu)$, strict convexity of the summed selection risk gives

$$\begin{aligned} \mathcal{R}(\nu) &= (\nu^{-1} - 1)m_1 + ((1-\nu)^{-1} - 1)m_2, \\ \mathcal{R}'(\nu) &= -\frac{m_1}{\nu^2} + \frac{m_2}{(1-\nu)^2} = 0 \quad \Longrightarrow \quad \nu^\star = \frac{\sqrt{m_1}}{\sqrt{m_1} + \sqrt{m_2}}, \\ \mathcal{R}(1/2) - \mathcal{R}(\nu^\star) &= m_1 + m_2 - 2\sqrt{m_1 m_2} = \big(\sqrt{m_1} - \sqrt{m_2}\big)^2 > 0. \end{aligned} \tag{8}$$

For this instance, joint allocation shifts the probabilities from $(0.5, 0.5)$ to approximately $(0.376, 0.624)$, assigning more capacity to the group with larger joint residual risk. This motivates allocation by the joint residual second moment. Appendix A.4 gives the full construction.

Remark 2 (When prefix–tail coupling matters)**.** *The theorem establishes a strict separation in a shared-parameter instance. The cross term vanishes if the vectors* R_{P} *and* R_{T} *occupy disjoint coordinates. The language-model probe in §5.4 measures coupling and its effect on allocation.*

4.4 Learning m under selective feedback

$\|g^\star - h\|^2$ is observed only when $Z = 1$. Partition groups into finitely many bins by a statistic that is predictable before the coin is drawn — we use (prompt stratum, number of censored members) — and estimate each bin by

$$\hat{m}_k = \frac{1}{|W_k|}\sum_{a\in W_k} X_a, \qquad X_a = \frac{Z_a}{\nu_a}\big\|g_a^\star - h_a\big\|^2, \tag{9}$$

over a window W_k of the most recent arrivals in bin k. Unselected arrivals contribute an explicit zero; they are not pseudo-labelled.

Proposition 2 (The selected-only average targets the wrong quantity)**.** *For a fixed arrival-index window, $\mathbb{E}[X_a \mid \mathcal{F}_a] = m_a$, so equation 9 is unbiased for the window mean $|W_k|^{-1}\sum_a m_a$. By contrast $\big(\sum_a Z_a X_a \nu_a\big)/\sum_a Z_a$ concentrates on the ν-weighted mean $\sum_a \nu_a m_a/\sum_a \nu_a$. If ν is constant within a bin the two agree; otherwise they can differ enough to invert a ranking. In the two-bin instance of Appendix A.5, the true means are $3/5$ and $4/5$ while the selected-only estimands are $41/45$ and $1/3$, so the ranking reverses.*

The multiplier λ in equation 6 depends on the whole batch, so within-bin continuation probabilities can vary across batches and covary with residual moments under drift. The IPW estimator in equation 9 accounts for this selective observation.

Theorem 3 (Excess risk of the learned allocation)**.** *Let $\mathcal{R}(\nu) = \sum_b (\nu_b^{-1} - 1) m_b$, let $\nu^\star$ minimise equation 5 with the true m, and let $\hat{\nu}$ be the output of equation 6 on predictions $\hat{m}$ with a feasible solution whose objective gap is at most $\varepsilon_{\mathrm{sol}}$. Then*

$$\mathcal{R}(\hat{\nu}) - \mathcal{R}(\nu^\star) \;\le\; \Big(\frac{1}{\alpha} - 1\Big) \sum_b \big|m_b - \hat{m}_b\big| + \varepsilon_{\mathrm{sol}}, \tag{10}$$

and the contribution to equation 4 is this divided by B^2.

Appendix A.6 gives a finite-window bound separating sampling error from bin mismatch and drift.

5 Experiments and analysis

We evaluate task accuracy, gradient correction, mechanism ablations and cost.

5.1 Shared evaluation setup

The two checkpoints are `Qwen3.6-27B` (dense) and `Qwen3.6-35B-A3B` (mixture of experts)(Qwen Team, 2026a;b). The six benchmarks are IMO-AnswerBench, HMMT Feb 26, LiveCodeBench v6, SuperGPQA, HLE and LongBench v2. HLE adds difficult cross-disciplinary reasoning; LongBench v2 tests transfer to long-context tasks. The six measured continuation arms use the same single-step, fixed-normalisation GRPO target (Appendix F.5), a 64M generated-token budget, $C = 16384$, $H = 32768$, $B = 16$ prompts per update, $G = 16$, seeds 42–46, $\beta = 0.10$, and per-bin capacity $N = 256$, giving $\alpha = 0.0625$. The budget includes prefixes, continuations and calibration; 600 updates is a safeguard. Training-source resource checks assess update count and effective history before comparative runs (Appendix F.3). All checkpoint evaluations and independent audits enter the all-in cost.

Baselines are uniform Bernoulli continuation (Fixed), Native C, value bootstrapping, length/cost and uncertainty allocation. Our group-level allocation controls share the corrected update, head architecture and fitting rule, bins, floor and normalisation; settings are shared across models and tasks. For each model, measured arms pair seeds 42–46, policy initialisation, training splits, candidate prompt order, optimiser settings and token accounting. Evaluation pairs the same questions, decoding settings and per-question decoding seeds. Baseline quotes published checkpoint scores under their source protocols. Appendix E specifies scoring and paired aggregation. Appendix B examines sensitivity to continuation budget, prefix cap and memory size.

5.2 Task performance across both model configurations

Table 1 uses mean $\pm$ SD for per-benchmark scores and 95% intervals for paired contrasts, weighting the six tasks equally. Joint allocation improves mean accuracy over Fixed by 6.27 and 6.18 percentage points on the dense and mixture-of-experts models, respectively. The corresponding gains over Uncertainty are 5.35 and 4.95 points. Table 7 reports per-task correction errors and censoring. Joint attains higher accuracy than every trained comparator on all six benchmarks.

Comparisons with established allocators. We select PAIR (Nomand et al., 2026), VIP (Nguyen et al., 2026) and VIGOR (Jiang et al., 2026) for direct comparison: they allocate member continuations, predict prompt-level rollout counts, and refine counts from observed rewards, respectively.

Table 1: **Qwen3.6 baselines and continuation comparison.** Published Baseline scores use their source protocols (Appendix F.1). Trained arms share 64M-token budgets and evaluation at H. Scores are mean $\pm$ sample SD over seeds 42–46; paired six-task contrasts give Δ [95% CI] in percentage points (Appendix E.1). PAIR, VIP and VIGOR are specified in Appendix E.3. LCB denotes LiveCodeBench.

Arm	IMO-AnswerBench	HMMT Feb 26	LCB v6	Super GPQA	HLE	LongBench v2
Qwen/Qwen3.6-27B						
Baseline	80.8	84.3	83.9	66.0	24.0	62.6
Native C	80.0 ± 0.65	82.4 ± 0.85	82.1 ± 0.69	65.7 ± 0.53	23.0 ± 1.05	61.4 ± 0.55
Value bootstrap	81.1 ± 0.80	83.6 ± 0.89	83.4 ± 0.75	66.4 ± 0.51	24.6 ± 0.85	62.5 ± 0.53
Fixed probability	82.0 ± 0.69	84.8 ± 0.97	84.5 ± 0.68	67.0 ± 0.49	25.3 ± 0.86	63.3 ± 0.45
Length/cost	82.4 ± 0.68	85.5 ± 0.80	84.8 ± 0.85	67.1 ± 0.53	25.8 ± 0.94	63.8 ± 0.52
Uncertainty	83.0 ± 0.81	86.1 ± 0.82	85.3 ± 0.74	67.6 ± 0.58	26.4 ± 1.03	64.0 ± 0.46
PAIR (aligned)	82.6 ± 0.57	84.3 ± 1.04	85.7 ± 0.67	67.3 ± 0.44	27.4 ± 0.81	63.9 ± 0.51
VIP (RLOO)	81.9 ± 0.78	85.6 ± 0.93	87.6 ± 0.87	67.9 ± 0.59	28.2 ± 0.92	65.7 ± 0.48
VIGOR (GRPO)	84.2 ± 0.93	86.4 ± 0.83	87.8 ± 0.72	68.2 ± 0.46	31.2 ± 0.83	64.8 ± 0.54
Learned joint	86.9 ± 0.69	87.8 ± 1.00	89.6 ± 0.72	72.9 ± 0.52	37.5 ± 0.85	69.8 ± 0.72
Qwen/Qwen3.6-35B-A3B						
Baseline	78.9	83.6	80.4	64.7	21.4	57.7
Native C	78.0 ± 0.82	81.8 ± 0.83	79.2 ± 0.62	64.1 ± 0.46	20.5 ± 0.94	56.7 ± 0.49
Value bootstrap	79.7 ± 0.65	83.7 ± 1.12	80.1 ± 0.71	65.8 ± 0.57	21.9 ± 0.96	57.8 ± 0.47
Fixed probability	80.0 ± 0.70	83.6 ± 1.18	81.2 ± 0.82	65.5 ± 0.51	22.7 ± 0.98	58.6 ± 0.54
Length/cost	81.6 ± 0.74	84.2 ± 0.89	80.7 ± 0.72	63.8 ± 0.42	21.2 ± 0.75	59.0 ± 0.51
Uncertainty	82.2 ± 0.72	84.8 ± 0.93	82.3 ± 0.77	66.2 ± 0.50	23.9 ± 0.80	59.6 ± 0.58
PAIR (aligned)	81.0 ± 0.60	84.4 ± 1.12	81.9 ± 0.63	68.1 ± 0.46	25.1 ± 0.85	60.1 ± 0.49
VIP (RLOO)	80.7 ± 0.65	85.4 ± 0.79	82.4 ± 0.73	67.2 ± 0.44	26.3 ± 0.97	62.4 ± 0.52
VIGOR (GRPO)	82.3 ± 0.72	86.0 ± 0.92	83.4 ± 0.73	67.9 ± 0.46	28.0 ± 0.76	61.2 ± 0.66
Learned joint	85.1 ± 0.82	87.0 ± 0.87	88.5 ± 0.98	69.3 ± 0.53	32.5 ± 0.77	66.3 ± 0.71

Paired Δ [95% CI]	27B	35B-A3B
Joint − Fixed	+6.27 [+5.48, +7.05]	+6.18 [+5.46, +6.90]
Joint − Uncertainty	+5.35 [+4.93, +5.77]	+4.95 [+4.45, +5.45]
Joint − PAIR	+5.55 [+4.98, +6.12]	+4.68 [+4.38, +4.98]
Joint − VIP	+4.60 [+4.13, +5.07]	+4.05 [+3.52, +4.57]
Joint − VIGOR	+3.65 [+2.91, +4.39]	+3.31 [+3.07, +3.56]

All runs use seeds 42–46, a 64M-token budget, and matched initial checkpoints, data and evaluation settings (Appendix E.3). VIGOR has the highest six-task mean among these three comparators; Joint exceeds it by 3.65 and 3.31 points on 27B and 35B-A3B, with paired 95% CIs of [2.91, 4.39] and [3.07, 3.56], respectively. Native allocation and update rules are retained.

5.3 Gradient correction at shared checkpoints

For each model and seed, the Fixed policy and head remain frozen at the common final token checkpoint. Allocators see identical prefixes and the same complete GRPO target U_H, with calibration, reference and test prompts disjoint. Define

$$\begin{aligned} \mathrm{CE}_U(\nu) &= \mathbb{E}_Z \left\| B^{-1} \sum_b (\hat{U}_b - U_{H,b}) \right\|^2 \\ &= B^{-2} \sum_b (\nu_b^{-1} - 1) \|U_{H,b} - h_{U,b}\|^2. \end{aligned} \tag{11}$$

Conditioning on complete test trajectories and the frozen pre-batch state gives this exact selection-risk audit; allocation sees no tails. Coin replay checks the identity on measured gradients.

Table 2 compares Fixed, Uncertainty, PAIR and Joint at equal expected capacity. After equal-task averaging, the reduction $100(1 - \overline{\mathrm{CE}}_{\mathrm{Joint}}/\overline{\mathrm{CE}}_{\mathrm{Fixed}})$ is 24.67% for 27B and 28.02% for 35B-A3B. The paired Joint-minus-Fixed CE intervals lie below zero in both models. Appendix C details aggregation, per-task errors and gradient coordinates. CE measures error against the complete surrogate. PAIR uses the same target via pair-inclusion correction (Appendix E.3).

Table 2: **Gradient correction at a shared GRPO checkpoint.** All allocators use the same Fixed-policy checkpoint within each model and seed, with predictors frozen. CE measures error against U_H, aggregated across six tasks before normalisation. The group-level identity is equation 11; the aligned PAIR audit is defined in Appendix E.3. Reduction is relative to Fixed. Absolute CE gives mean $\pm$ sample SD across seeds 42–46, in units of 10^{-6}. Paired intervals use the same units.

Model	Allocation	Relative CE	Reduction (%)	Absolute CE
27B	Fixed	1.000	0.00	12.067 ± 0.140
	Uncertainty	0.876	12.40	10.570 ± 0.289
	PAIR (aligned)	0.812	18.80	9.798 ± 0.506
	Joint	0.753	24.67	9.090 ± 0.482
35B-A3B	Fixed	1.000	0.00	12.450 ± 0.091
	Uncertainty	0.855	14.50	10.645 ± 0.152
	PAIR (aligned)	0.788	21.20	9.811 ± 0.337
	Joint	0.720	28.02	8.961 ± 0.480
27B: Joint − Fixed		-2.977 $[-3.666, -2.287]$ (95% CI)		
35B-A3B: Joint − Fixed		-3.489 $[-4.037, -2.941]$ (95% CI)		

5.4 Contributions of coupling and feedback correction

Online ablation under the main GRPO target. Table 3 crosses joint (J) versus marginal-sum (M) risk targets with IPW versus selected-only (SO) risk learning. For GRPO, fix the full-group statistics and write $U_H = U_H^{\mathrm{P}} + U_H^{\mathrm{T}}$; the residual components are $U_H^{\mathrm{P}} - h_U$ and U_H^{T} (Appendix F.5). J uses their squared sum, while M sums their squared norms. Every arm retains the same inverse-probability update correction. J-IPW reuses the main Joint run; J-IPW minus M-IPW isolates coupling and J-IPW minus J-SO isolates feedback learning. Initialisation, prompt order, optimiser and budget remain paired. J-IPW improves macro accuracy over M-IPW by 3.35/3.21 points and over J-SO by 1.63/1.40 points (27B/35B-A3B). The paired 95% intervals in Table 3 exclude zero for both contrasts in each model. Appendix D defines the interaction contrast.

Table 3: **Residual coupling and selective-feedback correction.** All arms share the GRPO target and update correction. M/J denote marginal-sum/joint residual risk; SO/IPW denote selected-only/weighted risk learning. Score is six-task macro accuracy (mean $\pm$ sample SD across seeds); CE uses the shared Fixed reference of Table 2. J-IPW reuses the main Joint run. Paired contrasts give Δ [95% CI] in percentage points.

		27B		35B-A3B	
Target	Risk learning	Score	CE	Score	CE
M	SO	70.38 ± 0.42	0.969	67.87 ± 0.29	0.957
M	IPW	70.73 ± 0.30	0.872	68.23 ± 0.39	0.848
J	SO	72.45 ± 0.14	0.843	70.05 ± 0.43	0.818
J	IPW	74.08 ± 0.45	0.753	71.45 ± 0.20	0.720
Paired Δ [95% CI]		27B		35B-A3B	
J-IPW − M-IPW		+3.35 [+2.55, +4.15]		+3.21 [+2.72, +3.71]	
J-IPW − J-SO		+1.63 [+1.16, +2.10]		+1.40 [+0.89, +1.91]	

Frozen probes that explain the online contrasts. The unclipped probe compares joint, marginal-sum, prefix-only and tail-only targets in Lemma 1, stratifying test comparisons by independently estimated coupling. Figure 2b measures allocation changes and CE on these same trajectories. At

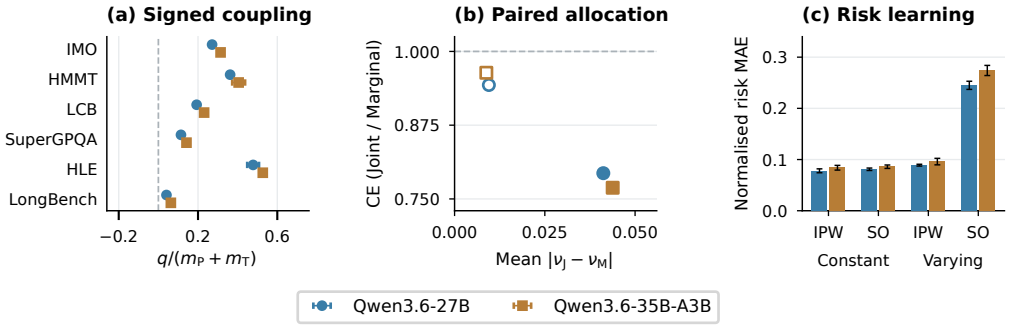


Figure 2: **Language-model mechanism diagnostics.** (a) Signed coupling by task. (b) Allocation change and Joint/Marginal CE; open/filled markers denote low/high coupling. (c) IPW versus selected-only (SO) risk MAE under constant/varying propensities (GRPO). Means use seeds 42–46; error bars in (a,c) show sample SD.

matched expected capacity, Joint-minus-Marginal CE is -0.211 for 27B and -0.268 for 35B-A3B, with paired 95% CIs $[-0.329, -0.093]$ and $[-0.374, -0.162]$, respectively (Table 8). Matched-log replay compares IPW and SO on identical histories. Under varying propensities, IPW lowers normalised risk MAE from 0.245 to 0.089 for 27B and from 0.274 to 0.096 for 35B-A3B; the constant-propensity control gives similar errors for the two learners (Figure 2c; Table 9). These probes connect residual coupling to allocation quality and propensity weighting to risk-estimation accuracy. Appendix D specifies the matched-log and online factorial protocols.

5.5 Performance and Computational Cost

Table 4 pairs final task accuracy with actual resource use for Fixed and Joint. At the shared 64M-token budget, Joint improves six-task mean accuracy by 6.27/6.18 points while using $1.065\times$/$1.075\times$ Fixed's all-in GPU time (27B/35B-A3B), including evaluation and audits. Tables 11 and 12 provide the per-arm resource counts and cost breakdown on matched hardware.

Table 4: **Performance and computational cost at matched token budgets.** Score is six-task mean accuracy (%), reported as mean ± sample SD over seeds 42–46. Accuracy gains are computed from unrounded seed means. Tokens include calibration. Relative GPU time includes training, evaluation and audits, normalised to Fixed within each model.

Model	Allocation	Score	Tokens (M)	Updates	Relative all-in GPU time
27B	Fixed	67.82 ± 0.34	64.13	48	$1.000\times$
	Joint	74.08 ± 0.45	64.23	45	$1.065\times$
35B-A3B	Fixed	65.27 ± 0.49	64.29	47	$1.000\times$
	Joint	71.45 ± 0.20	64.38	44	$1.075\times$

6 Conclusion

Budgeted continuation learns where additional generation most reduces correction error. Joint prefix–tail residuals determine its priorities, and all-arrival weighting addresses selective feedback. The resulting update is unbiased for the frozen-policy, finite-horizon gradient. Shared-checkpoint error measurements, target-matched ablations and cost accounting connect this local guarantee to the task-performance evaluation.

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

Our benchmark evaluation uses publicly available datasets. The study consists of computational experiments and does not involve human participants.

REPRODUCIBILITY STATEMENT

Code for the core algorithm and its verification is included with this submission in `atc_supplementary`. Within this directory, the batch procedure is implemented in `src/atc/trainer.py`, with residual correction in `src/atc/update.py` and risk estimation and budget allocation in `src/atc/risk.py`. Configuration definitions are provided in `src/atc/config.py`, and runnable presets are in `configs/`. Verification and correction-error audit scripts are in `scripts/`, including `scripts/audit_ceu.py`; dependencies and execution instructions are provided in `requirements.txt` and `README.md`. Appendix A provides the theoretical proofs, while Appendices E and F specify the benchmark evaluation, paired-seed statistics, training settings, and resource accounting.

REFERENCES

Yushi Bai, Shangqing Tu, Jiajie Zhang, Hao Peng, Xiaozhi Wang, Xin Lv, Shulin Cao, Jiazheng Xu, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. LongBench v2: Towards deeper understanding and reasoning on realistic long-context multitasks, 2024. URL https://huggingface.co/datasets/zai-org/LongBench-v2.

BottleCapAI. ThinkingCap: Qwen 3.6 27B, 2026. URL https://huggingface.co/bottlecapai/ThinkingCap-Qwen3.6-27B. Original Qwen3.6-27B baseline column; revision 0c5557fdf61f; accessed September 22, 2026.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

BytedTsinghua-SIA. DAPO-Math-17k training dataset. https://huggingface.co/datasets/BytedTsinghua-SIA/DAPO-Math-17k, 2025. Revision 65877096c24ffa7abc4e4fa5edb95cf3413a5674.

Center for AI Safety. MMLU auxiliary training split. https://huggingface.co/datasets/cais/mmlu, 2021. Revision c30699e8356da336a370243923dbaf21066bb9fe.

DeepMind. CodeContests dataset. https://huggingface.co/datasets/deepmind/code_contests, 2022. Revision 802411c3010cb00d1b05bad57ca77365a3c699d6.

Tiantian Fan, Lingjun Liu, Yu Yue, Jiaze Chen, Chengyi Wang, Qiying Yu, Chi Zhang, Zhiqi Lin, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Bole Ma, Mofan Zhang, Gaohong Liu, Ru Zhang, Haotian Zhou, Cong Xie, Ruidong Zhu, Zhi Zhang, Xin Liu, Mingxuan Wang, Lin Yan, and Yonghui Wu. Truncated proximal policy optimization. *arXiv preprint arXiv:2506.15050*, 2025. URL https://arxiv.org/abs/2506.15050v1.

Yangyi Fang, Jiaye Lin, Xiaoliang Fu, Cong Qin, Haolin Shi, Chaowen Hu, Lu Pan, Ke Zeng, and Xunliang Cai. How to allocate, how to learn? dynamic rollout allocation and advantage modulation for policy optimization. *arXiv preprint arXiv:2602.19208*, 2026. URL https://arxiv.org/abs/2602.19208.

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.

Daniel G. Horvitz and Donovan J. Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685, 1952.

InternScience. Agents-A1: Model card and evaluation results, 2026. URL https://huggingface.co/InternScience/Agents-A1. Qwen3.6-35B-A3B comparison column; revision e4a30b76e439; accessed September 22, 2026.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Baker-Whitcomb, Alex Beutel, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.

Heyang Jiang, Henry Liu, and Baharan Mirzasoleiman. Learning as reasoning unfolds: Progressive rollout allocation for efficient reinforcement learning. *arXiv preprint arXiv:2607.22002*, 2026. URL https://arxiv.org/abs/2607.22002.

John D. Kalbfleisch and Ross L. Prentice. *The Statistical Analysis of Failure Time Data*. Wiley, 2nd edition, 2002.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. *arXiv preprint arXiv:2305.20050*, 2023.

Roderick J. A. Little and Donald B. Rubin. *Statistical Analysis with Missing Data*. Wiley, 3rd edition, 2019.

LiveCodeBench Team. LiveCodeBench: Dataset versions and evaluation, 2025. URL https://github.com/LiveCodeBench/LiveCodeBench/blob/28fef95ea8c9f7a547c8329f2cd3d32b92c1fa24/README.md. Full release_v6: 1055 problems, May 2023–April 2025. Release specification at repository revision 28fef95; accessed September 26, 2026.

Thang Luong, Dawsen Hwang, Hoang H Nguyen, Golnaz Ghiasi, Yuri Chervonyi, Insuk Seo, Junsu Kim, Garrett Bingham, Jonathan Lee, Swaroop Mishra, Alex Zhai, Huiyi Hu, Henryk Michalewski, Jimin Kim, Jeonghyun Ahn, Junhwi Bae, Xingyou Song, Trieu Hoang Trinh, Quoc V Le, and Junehyuk Jung. Towards robust mathematical reasoning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 35418–35442. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.1794. URL https://aclanthology.org/2025.emnlp-main.1794/.

M-A-P. SuperGPQA: Graduate-level knowledge evaluation dataset, 2025. URL https://huggingface.co/datasets/m-a-p/SuperGPQA. Official dataset; accessed September 21, 2026.

MathArena. HMMT February 2026: Evaluation dataset, 2026. URL https://huggingface.co/datasets/MathArena/hmmt_feb_2026. 33 released problems; accessed September 21, 2026.

Hieu Trung Nguyen, Bao Nguyen, Wenao Ma, Yuzhi Zhao, Ruifeng She, and Viet Anh Nguyen. Adaptive rollout allocation for online reinforcement learning with verifiable rewards. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2602.01601.

Pixel Nomand, Elena Voss, Marcus Hale, and Sofia Reyes. PAIR: Pairwise-aware inclusion reweighting for adaptive rollout allocation in RLVR. *arXiv preprint arXiv:2608.11368*, 2026. URL https://arxiv.org/abs/2608.11368v2.

Art B. Owen. *Monte Carlo Theory, Methods and Examples*. 2013. available from the author.

Fabio Pardo, Arash Tavakoli, Vitaly Levdik, and Petar Kormushev. Time limits in reinforcement learning. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 4045–4054, 2018.

Long Phan, Alice Gatti, Ziwen Han, et al. Humanity's last exam, 2025. URL https://huggingface.co/datasets/cais/hle. Public test split; accessed September 22, 2026.

Qwen Team. Qwen3.6-27B: Official model card, 2026a. URL https://huggingface.co/Qwen/Qwen3.6-27B. Accessed September 17, 2026; revision 6a9e13bd6fc8.

Qwen Team. Qwen3.6-35B-A3B: Official model card, 2026b. URL https://huggingface.co/Qwen/Qwen3.6-35B-A3B. Accessed September 17, 2026; revision 995ad96eacd9.

James M. Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427):846–866, 1994.

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

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

Kimi Team. Kimi k1.5: Scaling reinforcement learning with llms. *arXiv preprint arXiv:2501.12599*, 2025.

Joel A. Tropp. Freedman's inequality for matrix martingales. *Electronic Communications in Probability*, 16:262–270, 2011. doi: 10.1214/ECP.v16-1624.

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

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*, 2025.

Heming Zou, Qi Wang, Yun Qu, Yuhang Jiang, Lizhou Cai, Yixiu Mao, Ru Peng, Xin Xu, Weijie Liu, Kai Yang, Saiyong Yang, and Xiangyang Ji. TRACE: A unified rollout budget allocation framework for efficient agentic reinforcement learning. *arXiv preprint arXiv:2606.11119*, 2026. URL https://arxiv.org/abs/2606.11119.

A Proofs

A.1 Theorem 1

Proof. Work conditionally on $\mathcal{F}$, which fixes h, ν and the costs. Since $Z \perp g^\star \mid \mathcal{F}$ and $\mathbb{E}[Z \mid \mathcal{F}] = \nu$,

$$\mathbb{E}[\hat{g} \mid \mathcal{F}] = h + \frac{1}{\nu}\,\mathbb{E}[Z \mid \mathcal{F}]\,\mathbb{E}[g^\star - h \mid \mathcal{F}] = \mathbb{E}[g^\star \mid \mathcal{F}],$$

which is (i); taking expectations and using $\mathbb{E}[g^\star] = \nabla J^H$ from equation 1 gives the unconditional statement. Note that h cancels identically, so no property of the head is used.

For (ii), $\hat{g} - g^\star = (Z/\nu - 1)(g^\star - h)$, and $Z/\nu - 1$ is independent of $g^\star - h$ given $\mathcal{F}$ with mean zero and

$$\mathbb{E}\Big[\big(\tfrac{Z}{\nu} - 1\big)^2\Big] = \nu\,\frac{(1-\nu)^2}{\nu^2} + (1-\nu) = (1-\nu)\Big(\frac{1-\nu}{\nu} + 1\Big) = \frac{1}{\nu} - 1.$$

Hence $\mathbb{E}[\|\hat{g} - g^\star\|^2 \mid \mathcal{F}] = (\nu^{-1} - 1)m$.

For (iii), condition on all complete potential group trajectories and on the pre-batch history. The batch allocation may share a multiplier, so the corrected updates need not be independent. Nevertheless the errors $e_b = (Z_b/\nu_b - 1)(g_b^\star - h_b)$ have conditional mean zero and are conditionally independent under independent selection coins. Hence cross-group error products and products with any complete-group update vanish. The complete groups are independent at the frozen policy, so their average contributes $B^{-1}\,\mathrm{tr}\,\mathrm{Cov}(g^\star)$; summing the conditional error variances from (ii) gives equation 4. □

A.2 PROPOSITION 1

Proof. Each $\nu \mapsto m_b/\nu$ is convex on $\nu > 0$ and the constraints are linear, so the problem is convex (Boyd & Vandenberghe, 2004); the feasible set is non-empty exactly when the floor is affordable, i.e. $\alpha \sum_b c_b \le b_{\max}$, and it is compact, so a minimiser exists. For $\lambda > 0$, minimising each scalar Lagrangian term $m_b/\nu_b + \lambda c_b \nu_b$ on $[\alpha, 1]$ gives equation 6 for $c_b m_b > 0$. Zero-cost coordinates take 1; zero-moment, positive-cost coordinates can be fixed at α. At $\lambda = 0$, all remaining coordinates take 1. Thus

$$S(0) = \sum_{c_b>0, m_b>0} c_b + \alpha \sum_{c_b>0, m_b=0} c_b.$$

If the budget exceeds this value the stated choice is optimal with slack. Otherwise continuity and monotonicity of S yield a root, including the floor-only boundary, which is attained at a sufficiently large finite multiplier. The scalar Lagrangian minimisers and complementary slackness certify optimality without needing strict feasibility. Strict convexity of m_b/ν_b when $m_b > 0$ gives uniqueness of those coordinates. The multiplier can be non-unique, and excess budget can leave the constraint slack. □

A.3 LEMMA 1

Proof. Write $\tilde{A}_i$ for the leave-one-out advantage built from $\tilde{Y}$. Then $A_i - \tilde{A}_i = \epsilon_i - \frac{1}{G-1}\sum_{j\neq i}\epsilon_j$ with $\epsilon_j = Y_j - \tilde{Y}_j$, and $s_i - s_i^{\mathrm{obs}} = (1-\delta_i)s_i^{\mathrm{T}}$. Therefore

$$g^\star - h = \frac{1}{G}\sum_i (A_i - \tilde{A}_i)s_i^{\mathrm{obs}} + \frac{1}{G}\sum_i A_i(1-\delta_i)s_i^{\mathrm{T}}.$$

Exchanging the order of summation in the first term,

$$\frac{1}{G}\sum_i \Big(\epsilon_i - \frac{1}{G-1}\sum_{j\neq i}\epsilon_j\Big)s_i^{\mathrm{obs}} = \sum_j \epsilon_j \cdot \frac{1}{G}\Big(s_j^{\mathrm{obs}} - \frac{1}{G-1}\sum_{i\neq j} s_i^{\mathrm{obs}}\Big) = \sum_j \epsilon_j \beta_j = R^{\mathrm{P}},$$

and the second term is R^{T}. Expanding $\|R^{\mathrm{P}} + R^{\mathrm{T}}\|^2$ and taking conditional expectations gives equation 7. Note $\epsilon_j = 0$ whenever $\delta_j = 1$, so R^{P} is supported on the censored members' label errors, while its directions β_j involve every member's observed score. □

A.4 THEOREM 2

The instance. There are three stochastic binary decisions (a, x_1, x_2) and eight leaves. Let $\Pr(a = 1) = \theta_1$ and let the independent outcome tokens have $\Pr(x_1 = 1) = \Pr(x_2 = 1) = \theta_2$. To realise a single fixed token cap, $a = 1$ emits $(a, x_1, x_2, \mathrm{EOS})$ and $a = 0$ emits $(a, \mathrm{pad}, \mathrm{pad}, \mathrm{pad}, x_1, x_2, \mathrm{EOS})$; set $C = 4, H = 7$. Padding and EOS are deterministic and have zero score. Thus the short branch finishes inside C, and only the route of the long branch has a nonzero observed score. Set $\theta_1 = 1/2$, $\theta_2 = 1/4$, $G = 2$ and $Y = \mathbf{1}[x_1 = x_2 = 1]$. Direct probability parametrisation keeps all moments rational. For the censored member the perfectly calibrated head predicts $p = \hat{v} = \Pr(Y_2 = 1 \mid \mathcal{F}) = 1/16$.

Member 2 is censored, with $s_2^{\mathrm{P}} = (-2, 0)$. Its four tails have probabilities $(9, 3, 3, 1)/16$, rewards $(0, 0, 0, 1)$, and second score coordinates $(-8/3, 8/3, 8/3, 8)$. The finished member has reward zero and either $(a, x_1, x_2) = (1, 1, 0)$ (class L) or $(1, 0, 0)$ (class H), with full scores $(2, 8/3)$ and $(2, -8/3)$. Let $w = s_1 - s_2^{\mathrm{P}} = (4, \sigma)$, $\sigma = \pm 8/3$. Lemma 1 gives

$$R^{\mathrm{P}} = -\tfrac{1}{2}(Y_2 - p)w, \qquad R^{\mathrm{T}} = \tfrac{1}{2}Y_2 s_2^{\mathrm{T}}.$$

Equal marginals and different joint moments. Since $\|w\|^2 = 208/9$ and $\mathrm{Var}(Y_2) = 15/256$,

$$\mathbb{E}\|R^{\mathrm{P}}\|^2 = \tfrac{1}{4}\tfrac{15}{256}\tfrac{208}{9} = \tfrac{65}{192}, \qquad \mathbb{E}\|R^{\mathrm{T}}\|^2 = \tfrac{1}{4}\tfrac{1}{16}64 = 1.$$

Writing σ_2 for the tail score coordinate,

$$2\mathbb{E}\langle R^{\mathrm{P}}, R^{\mathrm{T}}\rangle = -\tfrac{1}{2}\sigma\mathbb{E}[(Y_2 - p)Y_2\sigma_2] = -\tfrac{1}{2}\sigma\tfrac{15}{16}\tfrac{1}{16}8 = -\tfrac{15\sigma}{64}.$$

The two cross terms are $-5/8$ and $+5/8$, giving $m_{\mathrm{L}} = 137/192$ and $m_{\mathrm{H}} = 377/192$. Both marginals and the calibrated prediction coincide across classes.

The price. For two equal-cost groups and a budget of one continuation, the best marginal-measurable allocation is $(1/2, 1/2)$, since increasing a common probability reduces risk until the budget binds. For $\alpha \le 0.376$ the joint optimum is proportional to $(\sqrt{137}, \sqrt{377})$, approximately $(0.376101, 0.623899)$. Its selection risk is $\sqrt{51649}/96$, whereas the marginal rule pays $257/96$. The exact excess is

$$(\sqrt{m_{\mathrm{L}}} - \sqrt{m_{\mathrm{H}}})^2 = \frac{257 - \sqrt{51649}}{96} \approx 0.309748,$$

which is 13.0843% of the optimal selection risk. The mechanism persists under perfect calibration, with the head fixed at the exact conditional mean. It requires a mixed group in this instance: when both members are censored the observed and tail scores have disjoint support. □

A.5 Proposition 2

Proof. $\mathbb{E}[X_a \mid \mathcal{F}_a] = \nu_a^{-1}\mathbb{E}[Z_a \mid \mathcal{F}_a]\, m_a = m_a$, so $\hat{m}_k - |W_k|^{-1}\sum_a m_a$ is an average of martingale differences and equation 9 is unbiased for the window mean. The selected-only average is a ratio estimator whose numerator and denominator have conditional means $\sum_a \nu_a m_a$ and $\sum_a \nu_a$, so it concentrates on their ratio. The two coincide iff ν_a is uncorrelated with m_a across the window, in particular if ν is constant within the bin. □

The ranking-reversal instance. Two bins, each alternating between two regimes with equal frequency:

	regime 1		regime 2	
bin	ν	m	ν	m
A	$4/5$	1	$1/10$	$1/5$
B	$1/10$	$7/5$	$4/5$	$1/5$

The unweighted means are $m_{\mathrm{A}} = 3/5$ and $m_{\mathrm{B}} = 4/5$, so the truth ranks B above A. The ν-weighted means are $\frac{4/5\cdot 1 + 1/10\cdot 1/5}{9/10} = \frac{41}{45}$ and $\frac{1/10\cdot 7/5 + 4/5\cdot 1/5}{9/10} = \frac{1}{3}$, so the selected-only average ranks A above B. Both estimands are exact; only the sampling noise around them depends on the window length.

A.6 Theorem 3

Proof. Write $\hat{\mathcal{R}}(\nu) = \sum_b (\nu_b^{-1} - 1)\hat{m}_b$. Feasibility and $\varepsilon_{\mathrm{sol}}$-optimality give

$$\begin{aligned}\mathcal{R}(\hat{\nu}) - \mathcal{R}(\nu^\star) &\le \sum_b (\hat{\nu}_b^{-1} - \nu_b^{\star -1})(m_b - \hat{m}_b) + \varepsilon_{\mathrm{sol}} \\ &\le (\alpha^{-1} - 1)\sum_b |m_b - \hat{m}_b| + \varepsilon_{\mathrm{sol}}.\end{aligned}$$

Combining the two error terms before bounding them avoids an unnecessary factor of two. Dividing by B^2 gives the contribution to equation 4. □

Finite-window estimation. Fix a bin and a window of N_k consecutive, pre-specified arrival indices. Use the filtration that reveals each prefix and logged propensity before its selection coin and tail. For $d_a = \|g_a^\star - h_a\|^2 \in [0, M]$, $X_a = Z_a d_a/\nu_a$ and $m_a = \mathbb{E}[d_a \mid \mathcal{F}_a]$,

$$\mathbb{E}[X_a - m_a \mid \mathcal{F}_a] = 0, \quad |X_a - m_a| \le M/\alpha, \quad \mathrm{Var}(X_a \mid \mathcal{F}_a) = \frac{\mathbb{E}[d_a^2 \mid \mathcal{F}_a]}{\nu_a} - m_a^2 \le M^2/\alpha.$$

Freedman's inequality (Tropp, 2011), applied to both signs with a deterministic variance bound $N_k M^2/\alpha$, gives with probability $1 - \eta$

$$|m_b - \hat{m}_{k(b)}| \le \varepsilon_{\mathrm{bin}} + \varepsilon_{\mathrm{drift}} + M\sqrt{\frac{2\log(2/\eta)}{\alpha N_k}} + \frac{2M\log(2/\eta)}{3\alpha N_k}. \tag{12}$$

Table 5: **Memory-rule robustness at frozen checkpoints.** Both models share the default $N = 256$; $\alpha = \min\{\beta, N^{-1/2}\}$ in every row. Replay uses the same held-out complete trajectories. Entries are averaged across seeds 42–46. CE is relative to Fixed; Rank is Spearman correlation of predicted and reference m/c on positive-cost groups (undefined for constant ranks). All tasks use the same three settings; the reported endpoints are frozen-checkpoint CE and ranking.

Setting	N	CE 27B	Rank 27B	CE 35B	Rank 35B
Short memory	128	0.774	0.82	0.739	0.80
Shared default	256	0.753	0.87	0.720	0.86
Long memory	512	0.752	0.86	0.713	0.85

Here bin mismatch and window drift bound $|m_b - N_k^{-1} \sum_a m_a|$, the difference between the current group moment and the historical window mean. Equation equation 12 applies to a fixed arrival window with a population residual bound M. For K bins and T prespecified windows, a union bound replaces η by $\eta/(KT)$. Adaptive endpoints require a uniform bound over the candidate windows. The bin-specific result applies to occupied bins; pooled fallback uses a different reference population.

B Memory and budget sensitivity

B.1 Memory sensitivity

Memory rule and robustness check. The group-continuation arms use $N = 256$ arrivals per bin and $\alpha = \min\{\beta, N^{-1/2}\}$. Since $c_b \le G(H - C)$ and $\alpha \le \beta$, the floor is feasible under $b_{\max} = \beta BG(H - C)$. The rule targets $\sqrt{N}$ expected selected arrivals per full window when affordable.

Table 5 compares $N = 128, 256, 512$ by replaying common complete-feedback trajectories at frozen checkpoints. The floor follows the same rule in each condition; the measured endpoints are correction error and allocation ranking.

B.2 Continuation budget and censoring sensitivity

Table 6 uses the final shared Fixed checkpoint from Appendix C. At $C = 16384$, vary the external capacity fraction $\beta \in \{0.05, 0.10, 0.20\}$. At $\beta = 0.10$, vary the prefix cap $C \in \{8192, 16384, 24576\}$, retaining $H = 32768$. The common default is reused, giving five distinct conditions per model. These are separate one-dimensional sweeps, with $N = 256$, the same bins and head architecture, and the deterministic rule $\alpha = \min\{\beta, N^{-1/2}\}$ throughout.

The sweeps reuse complete calibration and test trajectories with the prefix appropriate to each cap. Risk estimates are recomputed from selected calibration feedback; test tails are revealed after allocation probabilities are fixed. Table 6 reports censoring, the Joint versus Fixed CE reduction and realised capacity. These are frozen-checkpoint replay measurements; full-training accuracy uses the default configuration.

C Shared-checkpoint correction error

C.1 Shared reference and correction-error aggregation

For each model and seed, the Fixed run supplies the reference policy and head at the final 64M-token boundary and at the 16M and 32M checkpoints. Each boundary uses the nearest completed batch with its actual token count. The policy, head, target rule, parameter coordinates, bins and predictor snapshots are fixed during the audit.

Calibration, reference and test prompts are disjoint. Calibration initialises each risk learner under its observation rule; reference prompts define risk strata. Allocators receive identical test prefixes

Table 6: **One-axis continuation-budget and censoring sensitivity.** Frozen GRPO reference, $H = 32768$, $N = 256$. Five distinct settings per model reuse the same complete trajectories. The default is shared between the two sweeps. CE reduction is Joint versus Fixed; Entries average seeds 42–46; selected capacity reports its realised mean divided by the expected budget.

Model	Sweep	β	C	Censor (%)	CE reduction (%)	Selected capacity ratio
27B	Default	0.10	16384	21.1	24.67	1.006
	Budget	0.05	16384	21.1	19.60	1.014
	Budget	0.20	16384	21.1	28.40	0.997
	Prefix cap	0.10	8192	31.8	32.70	1.009
	Prefix cap	0.10	24576	9.2	16.41	0.998
35B-A3B	Default	0.10	16384	23.4	28.02	1.008
	Budget	0.05	16384	23.4	23.10	1.011
	Budget	0.20	16384	23.4	31.99	0.996
	Prefix cap	0.10	8192	35.2	35.60	1.012
	Prefix cap	0.10	24576	11.4	19.69	1.003

and complete trajectories at H, with tails masked until probabilities are committed. Test labels are restricted to evaluation.

The full target is the GRPO surrogate U_H, with prefix counterpart h_U from Appendix F.5. Equation equation 11 gives the conditional CE on each complete batch. Gradient norms use the same trainable coordinates for every allocator and are accumulated blockwise. Independent coin replay compares the mean corrected update with the complete update.

Let $\mathrm{CE}_{a,s,q}$ be mean batch CE for allocation a, seed s and benchmark q. Define

$$\overline{\mathrm{CE}}_a = \frac{1}{5 \cdot 6} \sum_{s=1}^{5} \sum_{q=1}^{6} \mathrm{CE}_{a,s,q}, \qquad D_a = 100 \left(1 - \frac{\overline{\mathrm{CE}}_a}{\overline{\mathrm{CE}}_{\mathrm{Fixed}}} \right). \tag{13}$$

The ratio is computed after averaging absolute errors across seeds and tasks. Relative reductions are defined for positive aggregate Fixed error. Table 2 reports this reduction and the corresponding normalised CE; Appendix E.1 describes paired aggregation.

C.2 Correction error across six tasks

Table 7 reports absolute shared-checkpoint CE and pre-selection censoring for all six tasks. Both models use the same allocator settings and equal task weights. CE reductions range from 9% to 33% for 27B and from 12% to 38% for 35B-A3B, with the largest reduction on HLE in both configurations.

D Language-model mechanism experiments

D.1 Frozen-checkpoint coupling probe

The frozen-checkpoint probe uses the unclipped update in equation 1. Its metric is

$$\mathrm{CE} = \mathbb{E}_Z \left\| B^{-1} \sum_b (\hat{g}_b - g_b^\star) \right\|^2, \tag{14}$$

conditioned on the same complete batch; the full update is its reference.

Calibration logs initialise the moment estimates, a disjoint reference set estimates bin-level moments and fixes coupling strata, and the test set measures allocation and correction error. Both models cover mathematics, code and question answering.

The targets are $d_{\mathrm{J}} = \|R^{\mathrm{P}} + R^{\mathrm{T}}\|^2$ and $d_{\mathrm{M}} = \|R^{\mathrm{P}}\|^2 + \|R^{\mathrm{T}}\|^2$. Both use the same windowed estimator, bins, selected examples and arrival-memory length. Prefix-only and tail-only predictors

Table 7: **Task-level correction error and censoring.** Absolute CE is displayed in units of 10^{-6}, uses the shared GRPO reference, and is averaged over audit batches and training seeds 42–46. CE entries give mean ± sample SD across seeds; the aggregate first averages six tasks within each seed. Censoring is measured before selection at that reference.

Task	Fixed CE	Joint CE	CE reduction (%)	Censor (%)
Qwen3.6-27B				
IMO-AnswerBench	12.800 ± 0.345	9.600 ± 1.152	25.00	26.4
HMMT Feb 26	14.600 ± 0.160	10.366 ± 1.636	29.00	29.1
LiveCodeBench v6	10.900 ± 0.293	8.502 ± 0.455	22.00	21.7
SuperGPQA	7.400 ± 0.190	6.216 ± 0.397	16.00	8.6
HLE	18.500 ± 0.148	12.395 ± 1.912	33.00	34.2
LongBench v2	8.200 ± 0.211	7.462 ± 0.288	9.00	6.8
Six-task aggregate	12.067 ± 0.140	9.090 ± 0.482	24.67	21.1
Qwen3.6-35B-A3B				
IMO-AnswerBench	11.400 ± 0.187	8.208 ± 0.888	28.00	28.3
HMMT Feb 26	13.800 ± 0.327	9.384 ± 0.947	32.00	31.7
LiveCodeBench v6	12.300 ± 0.155	9.102 ± 0.807	26.00	24.9
SuperGPQA	8.100 ± 0.167	6.561 ± 0.193	19.00	10.2
HLE	19.600 ± 0.552	12.152 ± 2.274	38.00	37.4
LongBench v2	9.500 ± 0.244	8.360 ± 0.285	12.00	8.1
Six-task aggregate	12.450 ± 0.091	8.961 ± 0.480	28.02	23.4

Table 8: **Joint versus marginal allocation at matched expected capacity.** Frozen unclipped checkpoints; CE is defined in equation 14. CE gives mean ± sample SD across seeds 42–46; the paired Joint minus Marginal-sum contrast includes its 95% interval. Reference rank compares predicted and independently estimated bin-level m/c; tokens include allocation-related calibration. Constant Fixed ranks are undefined (n/a).

Model	Allocation	CE	Reference rank	Tokens (M)	GPU-hours
27B	Fixed	1.420 ± 0.011	n/a	32.00	24.00
	Uncertainty	1.288 ± 0.050	0.58	32.00	24.00
	Prefix-only	1.264 ± 0.052	0.62	32.00	24.00
	Tail-only	1.337 ± 0.043	0.54	32.00	24.00
	Marginal-sum	1.252 ± 0.054	0.77	32.00	24.00
	Joint	1.041 ± 0.052	0.88	32.00	24.00
Joint − Marginal-sum		-0.211 $[-0.329, -0.093]$ (95% CI)			
35B-A3B	Fixed	1.580 ± 0.009	n/a	34.00	20.00
	Uncertainty	1.397 ± 0.060	0.55	34.00	20.00
	Prefix-only	1.405 ± 0.061	0.60	34.00	20.00
	Tail-only	1.442 ± 0.056	0.56	34.00	20.00
	Marginal-sum	1.364 ± 0.067	0.74	34.00	20.00
	Joint	1.096 ± 0.058	0.87	34.00	20.00
Joint − Marginal-sum		-0.268 $[-0.374, -0.162]$ (95% CI)			

use $\|R^{\mathrm{P}}\|^2$ and $\|R^{\mathrm{T}}\|^2$. All risk labels use completed rewards, and every arm retains the corrected update in equation 2.

Allocation probabilities are fixed from observed prefixes before selection. Repeated draws estimate equation 14 at matched expected capacity. Table 8 reports CE, the Joint minus Marginal-sum contrast, reference ranking and resource costs.

The coupling diagnostic records $q = 2\mathbb{E}[\langle R^{\mathrm{P}}, R^{\mathrm{T}} \rangle \mid \mathcal{F}]$, estimated over reference bins, and its sign. The normalised diagnostic is $q/(m_{\mathrm{P}} + m_{\mathrm{T}})$ when the denominator is positive; if both marginal moments are zero, the ratio is undefined. Continuation probabilities and CE are compared within strata fixed using the independent reference set. Cross-term size, m/c dispersion and predictor error complement the online comparisons in Table 3.

Table 9: **Risk learning on identical selective-feedback logs.** Within each model and propensity condition, IPW and SO use the same arrival window and revealed targets. Bias, MAE and current error are normalised by the positive reference window mean; current-risk error is reported separately to distinguish drift. MAE gives mean ± sample SD across seeds 42–46; other entries give seed means. Reversal counts use independent reference estimates.

Model	Propensity	Learner	Bias	MAE	Reversal (%)	Current error
27B	Constant	IPW	+0.003	0.078 ± 0.004	1.8	0.112
	Constant	SO	+0.005	0.081 ± 0.002	2.1	0.121
	Varying	IPW	+0.004	0.089 ± 0.002	3.4	0.127
	Varying	SO	+0.168	0.245 ± 0.008	21.7	0.287
35B-A3B	Constant	IPW	+0.004	0.084 ± 0.005	2.0	0.126
	Constant	SO	+0.006	0.086 ± 0.003	2.4	0.132
	Varying	IPW	+0.005	0.096 ± 0.006	3.8	0.139
	Varying	SO	+0.187	0.274 ± 0.010	24.6	0.312

D.2 Matched-log risk learning

The primary log target is the joint GRPO surrogate residual from Appendix F.5. A common predictable behaviour allocator generates an arrival log containing prefix features, bin, probability ν_a, selection indicator Z_a, and the revealed scalar target only when $Z_a = 1$. IPW and SO read this identical log. Both use the same window of the most recent N arrivals:

$$\hat{m}_k^{\mathrm{IPW}} = \frac{1}{N_k} \sum_{a \in W_k} \frac{Z_a d_a}{\nu_a}, \qquad \hat{m}_k^{\mathrm{SO}} = \frac{\sum_{a \in W_k} Z_a d_a}{\sum_{a \in W_k} Z_a}. \tag{15}$$

With no revealed label, SO uses the same committed pooled IPW-history fallback as an empty IPW bin; the all-zero fallback is uniform at the same budget. Every estimate is frozen before the next batch, and arrivals are committed afterwards.

The comparison includes constant within-bin probabilities and naturally varying probabilities in training logs. Under the stationary constant-probability control, the two estimators target the same mean, while their variances can differ. Table 9 distinguishes signed bias and absolute error against the window mean from error against the current risk. The latter also reflects drift. Ranking reversals use independently estimated reference moments.

D.3 Online factorial contrasts

The four arms in Table 3 use the same single-step, fixed-normalisation GRPO objective as the main performance experiment. Initialisation, prompt order, predictor specification, head-learning rule, binning, floor and capacity fraction are paired. J-IPW reuses the main Joint run; the other arms are M-IPW, J-SO and M-SO. Their scalar risk labels use the complete GRPO partition defined in Appendix F.5. Only the risk target and its learning estimator change. The matched-log experiment isolates estimator effects, while the online ablation also captures their influence on subsequent feedback.

Primary contrasts are J-IPW minus M-IPW and J-IPW minus J-SO. The interaction is

$$\Delta_{\mathrm{interaction}} = (S_{\mathrm{J,IPW}} - S_{\mathrm{M,IPW}}) - (S_{\mathrm{J,SO}} - S_{\mathrm{M,SO}}), \tag{16}$$

where S is six-task macro accuracy. Table 3 reports the primary contrasts. The interaction is +1.28 points for 27B and +1.03 points for 35B-A3B, measuring the difference between coupling gains under IPW and SO risk learning.

E Evaluation protocol

E.1 Scoring and paired aggregation

The primary score is single-generation accuracy/pass@1 at H. For each backbone, the six measured continuation arms use seeds $\{42, 43, 44, 45, 46\}$ with the same policy initialisation, retained training

split and candidate prompt order. They share the optimiser configuration, B, G, C, H and the 64M-token stopping rule. Prefix generation, selected continuations and calibration all count towards that budget. Comparisons pair evaluation prompts and training seeds, with identical decoding settings and a shared decoding seed per prompt. Scores are averaged within each task and then across the six tasks with equal weight. For method a, seed s and benchmark q, let $S_{a,s,q}$ be the benchmark accuracy in percent. The paired contrast between methods a and b is

$$d_s^{a,b} = \frac{1}{6}\sum_{q=1}^{6}(S_{a,s,q} - S_{b,s,q}), \qquad \Delta_{a,b} = \frac{1}{5}\sum_{s=42}^{46} d_s^{a,b}.$$

The contrasts in Tables 1 and 3 give equal weight to the six benchmarks and the five paired training seeds. PAIR, VIP and VIGOR use the same seed and evaluation pairing, with their method-specific allocation and update rules specified in Appendix E.3.

Across-seed variability and paired intervals. For the five training seeds, the task-level sample standard deviation is

$$s_{a,q} = \sqrt{\frac{1}{4}\sum_{s=42}^{46}(S_{a,s,q} - \bar{S}_{a,q})^2}, \qquad \bar{S}_{a,q} = \frac{1}{5}\sum_{s=42}^{46} S_{a,s,q}.$$

For Table 3, compute each seed's six-task mean before calculating its sample SD. For a paired comparison, let s_d be the sample SD of the five differences $d_s^{a,b}$ defined above. The two-sided interval is

$$\mathrm{CI}_{95\%}(\Delta_{a,b}) = \left[\Delta_{a,b} - t_{0.975,4}\frac{s_d}{\sqrt{5}},\ \Delta_{a,b} + t_{0.975,4}\frac{s_d}{\sqrt{5}}\right].$$

This Student-t interval treats training seeds as independent units and assumes approximately normal seed-level differences. It conditions on the six-task evaluation suite and uses matched seeds across methods. Intervals are pointwise; no simultaneous-coverage claim is made. Published Baseline rows retain their source values and are excluded from these paired intervals.

Correction-error statistics. For CE, we first average the six task errors within each seed. Tables 2 and 7 report the mean and sample SD of those seed means; task rows use their five task-specific values. Relative CE and percentage reductions use the ratio of pooled means in equation 13. Paired CE intervals apply the same Student-t formula to seed-wise differences in absolute error. Joint minus PAIR gives 27B: -0.708 $[-0.927, -0.489]$; 35B-A3B: -0.849 $[-1.072, -0.627]$, in units of 10^{-6} (95% CIs). The factorial CE contrasts J-IPW minus M-IPW and J-IPW minus J-SO are 27B: -1.432 $[-2.017, -0.847]$ and -1.082 $[-1.475, -0.689]$; 35B-A3B: -1.597 $[-2.386, -0.808]$ and -1.223 $[-2.137, -0.309]$, respectively, in the same units. The unclipped probe in Table 8 applies this paired calculation to its own CE target in equation 14. Figure 2 uses five-seed means; error bars show sample SD for coupling and risk MAE.

E.2 Benchmarks and scoring rules

The primary suite uses the complete IMO-AnswerBench short-answer set (400 problems; Luong et al., 2025), the 33 MathArena HMMT February 2026 problems (MathArena, 2026), full LiveCodeBench `release_v6` (1055 problems, May 2023–April 2025; release specification at commit `28fef95`, LiveCodeBench Team, 2025), SuperGPQA (26529 questions; M-A-P, 2025), HLE's public test set (2500 questions; Phan et al., 2025), and LongBench v2 (503 questions; Bai et al., 2024). Dataset and evaluator revisions and the item manifest are fixed before training. All six sets are held out, including HMMT's and LongBench v2's publisher-named `train` splits. Mathematics uses a fixed final-answer checker, code uses execution tests, and SuperGPQA uses fixed multiple-choice prompts and answer parsing. HLE retains question images where supplied, with no external tools; answers and rationales are withheld, and a frozen answer-equivalence judge scores final answers. LongBench v2 uses its multiple-choice scorer over all 503 items. HLE and LongBench v2 reuse the QA allocation bins and the text-only training mixture.

For LongBench v2, the total context window is fixed at 262144 tokens, reserving $H = 32768$ for generation. If an input exceeds the remaining budget, the document context is middle-truncated,

preserving equal head and tail portions; the question, choices and chat template are retained. The same formatted input is used across arms, and retained lengths and truncation flags are logged. Input processing and HLE image processing enter evaluation GPU time.

Table 7 reports their pre-selection censoring rates. The prefix-cap sweep in Appendix B.2 examines the effect of censoring under common trajectories. Training-data filtering and evaluation-set separation are specified in Appendix F.2.

E.3 Direct comparison protocol for established allocators

Table 1 reports PAIR, VIP and VIGOR task scores under matched training and evaluation settings. Tables 11 and 12 report actual resource use; Table 2 reports the aligned PAIR audit.

Method-specific allocation. PAIR retains the graph-based probability design, correctness/cost predictors and pair-inclusion correction of Nomand et al. (2026). It uses $G = 16$ and the same C, H as Joint; completed members have inclusion probability one, and unfinished members use floor $\alpha = 0.0625$. Expected remaining capacity matches Joint. "Aligned" denotes these shared caps and the update scaling below.

VIP retains its GP predictor, RLOO variance objective and integer rounding (Nguyen et al., 2026), with 256 rollouts over $B = 16$ prompts and per-prompt bounds 4–32 (`https://github.com/HieuNT91/VIP`). VIGOR follows Algorithm 1 of Jiang et al. (2026): four refinement rounds, initial count 4, retention fraction $1/2$ and expansion factor 2, giving $64 + 64 + 64 + 64 = 256$ rollouts. Both complete allocated rollouts to H; VIP retains RLOO and VIGOR retains GRPO.

Paired budget and evaluation. All three methods use seeds 42–46, common initial checkpoints, candidate prompt streams, data splits, verifiers, outer optimisers and trainable modules within each model. Validation splits and tuning budgets are shared, with choices fixed before test evaluation. The 64M-token stopping rule and 600-update safeguard count calibration, initial/refinement rollouts, prefixes and tails; selected batches complete. All runs reached the token target before the update safeguard; actual token totals determine the budget match. The six tasks use identical questions, decoding settings and per-question seeds. All-in time includes predictor fitting, allocation, verification, policy work, evaluation and audits on matched hardware.

PAIR target alignment and audit. Let s_i be the complete trajectory score, $P_G = \binom{G}{2}$, $K_{ij} = \frac{1}{2}(Y_i - Y_j)(s_i - s_j)$ and $\kappa = (G-1)/(GHs_0)$. Equation equation 19 yields

$$U_H = \frac{\kappa}{P_G} \sum_{i<j} K_{ij}, \qquad \widehat{U}_{\mathrm{PAIR}} = \frac{\kappa}{P_G} \sum_{i<j} \frac{J_i J_j}{\rho_{ij}} K_{ij}. \tag{17}$$

Here J_i indicates an observed complete member and ρ_{ij} is its logged joint inclusion probability with j. Independent selection gives $\rho_{ij} = \pi_i \pi_j$, including $\pi_i = 1$ for completed members, and $\mathbb{E}_J[\widehat{U}_{\mathrm{PAIR}}] = U_H$ conditionally on the candidate population. The audit freezes the same Fixed checkpoint, complete held-out trajectories and gradient coordinates as Table 2, matching $\sum_{b,i:\delta_{bi}=0} \pi_{bi}(H - C) = \sum_b \nu_b c_b$. Pair-inclusion covariances or repeated selection draws estimate CE; absolute errors are aggregated before Fixed normalisation. Allocation sees prefixes only, with calibration and test data disjoint.

F Training protocol and resource accounting

F.1 Published checkpoint baselines

The published Baseline rows in Table 1 use the Qwen model cards for IMO-AnswerBench, HMMT February 2026, LiveCodeBench v6, SuperGPQA and HLE (Qwen Team, 2026a;b). LongBench v2 uses the original-checkpoint results from ThinkingCap (27B: 62.6; BottleCapAI, 2026) and Agents-A1 (35B-A3B: 57.7; InternScience, 2026). These retain their source protocols; trained-arm comparisons use Appendix E. Source URLs, revisions and evaluation settings are recorded in `results/official/qwen36_baselines.json`.

Table 10: **Shared training configuration.** Settings for the continuation and factorial arms. Allocator-specific adaptations are in Appendix E.3. Dataset revisions, preprocessing manifests and software records are stored with the protocol.

Component	Setting
Sampling and update	$B = 16$, $G = 16$; one accumulated update per batch; $C = 16384$, $H = 32768$
Budget and memory	64M generated training tokens including calibration; $\beta = 0.10$, $N = 256$ per bin, $\alpha = 0.0625$
Optimiser	AdamW; constant learning rate 10^{-6}; betas $(0.9, 0.999)$, epsilon 10^{-8}, weight decay 0; aggregate gradient norm cap 1
Trainable parameters	LoRA rank 16, scale 32, dropout 0 on text-decoder attention and feed-forward projections, including expert projections; exclude routers, embeddings, output head and vision encoder
Update target	Group-centred reward with fixed calibration scale s_0; fixed $1/H$ length factor; one on-policy step; no KL term
Imputation head	Sigmoid linear head on detached last-token hidden state at C; zero initial weights/bias; one AdamW BCE step at 10^{-3} on current revealed censored labels, weight decay 0
Generation	Thinking enabled; training uses full-softmax sampling at temperature 1.0, dropout disabled, and at most 8192 input tokens; evaluation uses temperature 1.0, top-p 0.95, top-k 20
Calibration and checkpoints	4M initial calibration target within the training allocation; checkpoints near 0, 16, 32, 48, 64M actual generated tokens

F.2 Training sources and shared settings

The training pool combines DAPO-Math-17k's training split, CodeContests' training split, and MMLU's auxiliary-training split in a $1:1:1$ ratio (BytedTsinghua-SIA, 2025; DeepMind, 2022; Center for AI Safety, 2021). The protocol pins dataset revisions and licences; their metadata are recorded in `results/official/training_sources.json`. Canonical duplicate questions are grouped before splitting. Exact question matches and shared normalised 13-token question spans with any evaluation set are removed. A SHA256 question hash modulo 100 assigns buckets 0–89 to policy training, 90–94 to calibration and 95–99 to reference prompts. Hash parity divides calibration into initial scale/resource estimation and later selected-feedback audit calibration. Retained item manifests and exclusion counts are frozen before execution.

Mathematics retains signed-integer answers and uses exact final-answer matching. Code training uses Python solutions on deterministic exact-output problems, with all retained tests, a five-second per-test limit and 512 MiB memory. Tasks requiring custom output checkers are excluded. QA uses exact option-letter matching. Reference answers and test outputs never enter model prompts. The extra prompt rotates across the three domains in batches of 16. The reward-runner image and resolved trainable-module manifest accompany each run.

F.3 Token accounting

The continuation arms use $B = 16$, $G = 16$ and $N = 256$ under the settings in Table 10. Training stops at 64M generated tokens with a 600-update safeguard. Prefixes, selected continuations and calibration all count; the final selected batch completes. Initial calibration targets 4M tokens, completes groups at H, fixes the advantage scale and measures resource statistics. The head and risk windows then learn from selected feedback. Evaluations use completed batches nearest to $0, 16, 32, 48, 64$M tokens, retaining actual counts and consolidating duplicate checkpoint identifiers. Table 11 reports actual tokens, updates, prefix lengths and pre-selection censoring.

F.4 All-in computational costs

Table 12 uses matched hardware and software within each model. Training GPU-hours include generation, calibration, verification, residual gradients, risk learning, caching and policy/head work. All-in totals also include every checkpoint evaluation and independent audit; shared audits are

Table 11: **Training resources and censoring.** Training tokens include calibration. Prefix length is the rollout mean; censoring is measured at C. The n/a entries mark prefix-specific quantities for full-rollout methods. Complete-surrogate CE is reported in Table 2.

Arm	Tokens (M)	Updates	Mean prefix	Censor (%)
Qwen3.6-27B				
Native C	64.22	61	3856	16.2
Value bootstrap	64.31	59	3993	17.2
Fixed probability	64.13	48	4210	18.4
Length/cost	64.21	47	4246	18.7
Uncertainty	64.20	46	4288	19.3
PAIR (aligned)	64.25	47	4295	19.4
VIP (RLOO)	64.18	34	n/a	n/a
VIGOR (GRPO)	64.21	32	n/a	n/a
Learned joint	64.23	45	4330	19.8
Qwen3.6-35B-A3B				
Native C	64.21	60	3920	16.8
Value bootstrap	64.27	58	4059	17.8
Fixed probability	64.29	47	4300	19.1
Length/cost	64.34	46	4335	19.5
Uncertainty	64.38	45	4390	20.2
PAIR (aligned)	64.31	46	4410	20.4
VIP (RLOO)	64.26	33	n/a	n/a
VIGOR (GRPO)	64.33	31	n/a	n/a
Learned joint	64.38	44	4450	20.9

Table 12: **End-to-end cost of adaptive continuation.** Identical hardware is used within each model. Training GPU-hours include allocation-related calibration; all-in totals add every evaluation and audit. Relative all-in time divides the five-seed mean by Fixed within each model. GPU-hours are reported per run for eight devices.

Component	Fixed	M-IPW	J-SO	J-IPW
27B				
Training GPU-hours	400.0	419.0	424.0	432.0
All evaluations (GPU-h)	336.0	353.0	356.0	356.0
Audit GPU-hours	64.0	64.0	64.0	64.0
All-in GPU-hours	800.0	836.0	844.0	852.0
Relative all-in time	1.000	1.045	1.055	1.065
Peak device memory (GiB)	57.6	60.1	59.5	61.4
35B-A3B				
Training GPU-hours	280.0	299.0	302.0	307.0
All evaluations (GPU-h)	268.0	281.0	285.0	286.0
Audit GPU-hours	52.0	52.0	52.0	52.0
All-in GPU-hours	600.0	632.0	639.0	645.0
Relative all-in time	1.000	1.053	1.065	1.075
Peak device memory (GiB)	61.8	65.3	64.5	66.7
Established allocators: all-in GPU-hours				
Model		PAIR	VIP	VIGOR
27B		812.0	826.0	841.0
35B-A3B		612.0	625.0	639.0

charged equally. Gradient audits use the trainable coordinates in Table 10, with backward-pass, transfer and communication costs included. Relative time divides the five-seed mean by Fixed within each model.

F.5 MAIN COMPLETE AND PREFIX TARGETS

The main experiment uses a single-step, fixed-normalisation GRPO variant. Write θ_{old} for the policy used to sample the entire batch. Training samples the full softmax at temperature one; the score likelihood uses this same policy, with dropout disabled. Initial independent calibration fixes $s_0 = \max\{10^{-6}, [\mathrm{mean}_{\mathrm{cal\ groups}}\, G^{-1}\sum_i (Y_i - \bar{Y})^2]^{1/2}\}$ for the run. For a complete group, define

$$\begin{aligned} \bar{Y} &= G^{-1}\sum_i Y_i, \qquad a_i = (Y_i - \bar{Y})/s_0, \\ \rho_{it}(\theta) &= \frac{\pi_\theta(y_{it} \mid x, y_{i,<t})}{\pi_{\theta_{\mathrm{old}}}(y_{it} \mid x, y_{i,<t})}, \\ U_H &= \nabla_\theta \left[\frac{1}{GH} \sum_i a_i \sum_{t=1}^{\min(T_i,H)} \rho_{it}(\theta) \right] \Bigg|_{\theta=\theta_{\mathrm{old}}}. \end{aligned} \tag{18}$$

Rewards, advantages and the behaviour denominator are detached. There is one accumulated gradient update per rollout batch, so ratios are one at evaluation and PPO clipping is inactive; the main target has no KL term. The fixed $1/H$ denominator and calibration scale replace sequence-length and per-group standard-deviation normalisation. These define the variant used in the six continuation arms and the factorial ablation.

For h_U, replace Y_i by the observed/imputed $\tilde{Y}_i$, form $\tilde{a}_i = (\tilde{Y}_i - G^{-1}\sum_j \tilde{Y}_j)/s_0$, and restrict the inner sum in equation 18 to observed tokens. This supplies a prefix-computable gradient with the same denominator and scale. At the frozen evaluation point,

$$U_H = \frac{G-1}{GHs_0} g^\star, \qquad h_U = \frac{G-1}{GHs_0} h. \tag{19}$$

The complete rewards and advantages are computed after continuation; the policy, head and computation rule are frozen beforehand. Selection correction acts on these gradients. AdamW receives their negative for objective ascent after aggregate gradient-norm clipping; CE audits use the gradients before those optimiser operations.

Residual partition for the factorial ablation. Fix the full-group statistics and partition the GRPO token sum as $U_H = U_H^{\mathrm{P}} + U_H^{\mathrm{T}}$. With $R_U^{\mathrm{P}} = U_H^{\mathrm{P}} - h_U$ and $R_U^{\mathrm{T}} = U_H^{\mathrm{T}}$, $R_U^{\mathrm{P}} + R_U^{\mathrm{T}} = U_H - h_U$. Table 3 compares learning $\|R_U^{\mathrm{P}} + R_U^{\mathrm{T}}\|^2$ against $\|R_U^{\mathrm{P}}\|^2 + \|R_U^{\mathrm{T}}\|^2$. Full-group advantages and fixed normalisers are computed before partitioning, with shared parameter coordinates retained in both components. The unclipped probe follows Lemma 1.

G IMPLEMENTATION

Algorithm 1 gives one update. Prefix generation provides observed scores, rewards and head features; selected continuation supplies missing rewards and tail scores. The allocator and risk windows are updated after the batch.

Algorithm 1 Budgeted correction of censored group updates (one update)

input batch of B prompts, budget fraction β, memory length N
set $\alpha \leftarrow \min\{\beta, N^{-1/2}\}$; freeze policy, head and target rule; read frozen bin predictions $\hat{m}$
generate G rollouts per prompt up to C; form h_b, costs c_b, bins $k(b)$
$b_{\max} \leftarrow \beta BG(H - C)$; **assert** $\alpha \sum_b c_b \leq b_{\max}$
if all predictions are zero, allocate uniformly at the same budget; otherwise use equation 6
for $b = 1$ **to** B **do**
  if $c_b = 0$ **then**
    $\hat{g}_b \leftarrow h_b$ {already complete; no bin, no cost}
  else
    draw $Z_b \sim \mathrm{Bern}(\nu_b)$
    if $Z_b = 1$ **then**
      resume every censored member to H; form $g_b^\star$
      $\hat{g}_b \leftarrow h_b + \nu_b^{-1}(g_b^\star - h_b)$; queue revealed labels for the head
      stage record $(k(b), \nu_b, 1, \|g_b^\star - h_b\|^2)$
    else
      $\hat{g}_b \leftarrow h_b$; stage record $(k(b), \nu_b, 0, \cdot)$ {zero IPW contribution}
    end if
  end if
end for
apply $B^{-1} \sum_b \hat{g}_b$; commit staged records into the windows; update the head

For the main target, substitute $(U_H, h_U, \hat{U})$ for $(g^\star, h, \hat{g})$ throughout Algorithm 1, using equation 18. Optimiser state changes only after the batch gradient has been accumulated. All observed-token and tail-token terms use the same frozen policy and trainable coordinates.

Batch update order. Head and risk snapshots are read before selection. Residual records, zero IPW contributions and revealed censored-rollout labels are committed after the batch. Zero-cost groups take $\nu = 1$ and contribute their complete update; positive-cost arrivals populate the risk windows. The head is fitted on revealed labels from censored rollouts.

H LIMITATIONS

The guarantees concern a frozen-policy, unclipped finite-horizon gradient and an expected capacity budget. Population residual bounds, within-bin heterogeneity and window drift remain assumptions for language-model runs; optimiser dynamics and convergence require separate analysis.