DPR: FROM HIGH-ENTROPY TOKENS TO LEARNING CLAIMS
MATCHED EVALUATION FOR TOKEN-SELECTIVE LLM REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Decision-point reinforcement (DPR) is a verifier-free algorithmic framework for routing sequence-level reinforcement signals to informative response tokens. For each prompt, the policy generates on-policy rollouts and constructs a consensus pseudo-label from semantically equivalent parsed answers, yielding group-level reward signals. A frozen reference model ranks response positions by next-token entropy. DPR applies a clipped policy-gradient objective to the highest-entropy positions and anchors complementary tokens to the reference policy through a low-variance KL penalty. This dual-mask objective integrates entropy-guided credit assignment, reference-policy regularization, and total-token normalization within a unified training procedure. A count-matched random-routing counterpart and shared optimization contract enable precise evaluation of token-location quality. Controlled experiments show that DPR outperforms matched random routing in the anchor and data-shift evaluations, with favorable effects across learning checkpoints. Overall, DPR provides an effective and auditable approach to selective credit assignment for language-model reinforcement learning.

1 INTRODUCTION

Reinforcement learning can concentrate policy change at a sparse set of uncertain response positions. This observation motivates *decision-point reinforcement*: use a frozen reference model to identify uncertain tokens, apply the learning signal at those positions, and anchor the remaining response to the reference policy. The promise is direct—turn a coarse sequence reward into selective token-level credit without requiring a learned verifier.

That promise does not follow from localization alone. Knowing *where* a trained policy differs from its base model does not reveal *which direction* an untrained policy should move, or whether selecting those positions is better than selecting an equal number elsewhere. An intervention that already knows the desired replacement token combines location with privileged direction. A genuine learning claim must remove that privilege and compare the selector with a count-matched alternative.

We introduce DPR as both that comparison and an audit of the evidence it produces. The Decision arm selects frozen-reference high-entropy positions. The Random arm selects the same number of positions on its own live rollouts, so each policy retains its own trajectory and pseudo-label. Both arms share the reward, gradient denominator, clipping rule, anchoring rule, training schedule, and evaluation protocol. Global-signal controls diagnose optimizer and anchoring effects, but are not misrepresented as location-only baselines.

The retained localization summaries provide descriptive support for the premise. The available multi-arm training study, by contrast, cannot resolve an ordering and is reported only as a case study. We then apply the audit prospectively to thirteen normalization-matched seed-matched run pairs. At the prespecified endpoint, Decision exceeds Random by 2.12 percentage points (95% CI $[1.22, 3.02]$, two-sided $p = 0.00024$), while remaining a within-setting result. The contrast between these studies exposes reusable failure modes that can change an apparent token-selection result without changing the selector. Prospectively frozen scope tests make that boundary concrete: the

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

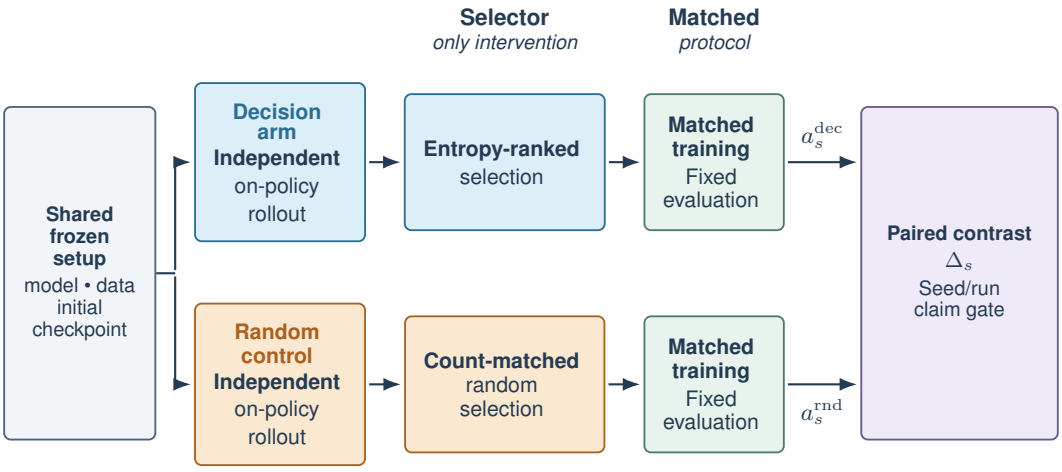


Figure 1: **DPR matched-control protocol.** Decision and Random share a frozen initialization and independently form on-policy rollouts and pseudo-labels. Only selection differs—entropy ranking versus count-matched random selection. Training configuration, nominal rollout/update budgets, and fixed evaluation are matched; the seed-matched run-pair contrast $\Delta_s = a_s^{\mathrm{dec}} - a_s^{\mathrm{rnd}}$ gates the learning claim.

advantage replicates under a training-data shift but reverses under a model-family shift. Figure 1 summarizes the DPR pipeline and the evidence gates between an intervention and a learning claim.

Optimizer comparability (Sec. 5.2). Selected-token normalization can move an arm into a different pre-clip gradient regime. Clipping then makes the comparison partly about optimizer scale rather than token location. Pre-clip norms and clipping incidence expose this confound without being mistaken for parameter-step measurements.

Counterfactual validity (App. P.3). Natural resampling and forced editing can bias token-leverage tests in opposite directions. At a nearly deterministic position, a mild alternative may not exist, leaving little common support. Perturbation strength and overlap must therefore precede a causal reading.

Statistical unit (App. P.4). Pooling evaluation problems across training seeds hides the run-to-run variation that a method claim is about. We retain pooled analysis only as a supplementary description and use the seed-matched run pair as the independent unit.

Scoring and provenance (App. P.5). Parsing, answer-format, and termination rules decide which rollouts can vote. A rule tuned to one model family can silently reject valid outputs from another; incomplete artifacts can also make the executed rule unrecoverable. These semantics must be executable and archived.

Contributions.

- An explicit separation of entropy localization, privileged edit direction, and token-selection learning advantage—claims that require different evidence.
- A count-matched Decision–Random protocol whose optimizer, reward, scoring, statistical, and provenance semantics are designed to be executable and auditable (§3).
- A raw-recomputed matched anchor, prospectively frozen model- and data-shift replications, and a practical checklist that turns a positive anchor, a data-shift replication, and a model-family reversal into an explicit scope boundary (Sec. 5 and App. S).

Scope. The prospective evidence supports Decision in the anchor and data-shift settings but reverses under the model-family shift. It therefore does not establish a universal entropy-gating ad-

vantage. The broader contribution is a controlled protocol and an evidence standard for deciding when such a claim is warranted.

2 Related Work

Policy optimization and credit assignment. Token-selective objectives inherit the credit-assignment problem of policy gradients, from REINFORCE, generalized advantage estimation, and PPO (Williams, 1992; Schulman et al., 2016; 2017) to language-model alignment and verifiable-reward training (Ouyang et al., 2022; Ahmadian et al., 2024; Shao et al., 2024; Guo et al., 2025; Liu et al., 2025; Yu et al., 2025a). Process supervision and token- or step-sensitive estimators instead seek denser credit (Cobbe et al., 2021; Lightman et al., 2024; Kazemnejad et al., 2025; Cui et al., 2026). DPR does not introduce a new optimizer or reward model. It holds the selected-token count and total-token denominator fixed so that the comparison isolates where an otherwise matched policy-gradient signal is applied.

Token localization and entropy intervention. Policy change in reasoning models can concentrate at sparse, high-entropy positions, motivating selective updates (Wang et al., 2025; Meng et al., 2026; Fu et al., 2026). Subsequent work studies entropy preservation or change, token-level gradient reweighting, exploration, and clipping (Petrenko et al., 2026; Lin et al., 2026; Cheng et al., 2026; Hao et al., 2026; Jin et al., 2026a; Chen et al., 2026). The closest contemporaneous proposals select high-entropy, surprising, or vision-anchored tokens and analyze where hindsight credit can reside (Wu et al., 2025; Jin et al., 2026b; He et al., 2026; Lv et al., 2026). A complementary RL-free study tests teacher-defined corrections and then applies entropy-gated contrastive training (Akgül et al., 2026). Thus, high-entropy selection itself is not our novelty. Our narrower contribution is a count-matched random-location estimand with a shared total-token denominator, independent paired training runs, and an auditable scoring and provenance protocol.

Verifier-free and self-generated supervision. Self-consistency, bootstrapped reasoning, self-reward, and self-training derive supervision from a model's own samples (Wang et al., 2023; Zelikman et al., 2022; Yuan et al., 2024; Singh et al., 2024). Verifier-free RL likewise uses sampled agreement, confidence, intrinsic signals, or entropy objectives (Zuo et al., 2025; Yu et al., 2025b; Zhao et al., 2026; Wen et al., 2026). These methods expose a shared evaluation risk: difficult prompts can yield confidently wrong pseudo-labels. We therefore separate reward construction, which defines the training signal, from entropy-based token selection, which defines where that signal is applied. Our case study does not identify entropy as the cause of an observed difference; it tests the matched intervention required to make that claim assessable.

Experimental evidence in reinforcement learning. RL conclusions are sensitive to random seeds, implementation details, and the statistical unit of analysis (Henderson et al., 2018; Engstrom et al., 2020; Andrychowicz et al., 2020; Colas et al., 2018). Aggregate point estimates can also obscure uncertainty and task heterogeneity (Agarwal et al., 2021; Patterson et al., 2024), while significance procedures must be declared with their comparison family (Dror et al., 2018). Accordingly, DPR uses independent seed-matched run pairs, paired intervals and tests, explicit multiplicity families, and separate environment-level conclusions rather than pooled problem rows.

Which groups are worth training on. Group-based objectives provide no policy-gradient signal when every sampled response receives the same reward. Replay-based work replaces such uninformative groups with off-policy examples (Mao et al., 2026). DPR addresses a different axis: conditional on an informative group, it controls which token positions receive the signal. We report uninformative groups as a diagnostic rather than presenting replay as part of the method.

Positioning. Efficiency, descriptive localization, privileged edit direction, and learning advantage are distinct claims. DPR contributes a controlled token-selection comparison and an evidence protocol spanning optimizer scale, counterfactual support, statistical independence, scoring semantics, and provenance. The diagnostic case study illustrates what a valid comparison must retain; its setting-dependent ordering is not interpreted as a universal gating result.

3 METHOD: A MATCHED TOKEN-SELECTION ESTIMAND

3.1 VERIFIER-FREE SIGNAL

For each prompt x, we sample a group of G rollouts $\{y^{(1)}, \ldots, y^{(G)}\}$ from the current policy. The prospective configurations select the `math_verify` parser and equivalence checker. Parseable answers are greedily assigned to the first equivalent existing cluster; the largest cluster defines the *consensus pseudo-label*. Stable sorting resolves a largest-cluster tie in favor of the cluster whose first member occurs earliest in rollout order. Rollout i receives $r_i = 1$ exactly when it belongs to that cluster and $r_i = 0$ otherwise; gold answers are used only for diagnostics, never for this reward.

Abstainers are excluded from clustering but receive zero reward and remain in the G-rollout normalization. Specifically,

$$\bar{r} = G^{-1}\sum_i r_i, \qquad \sigma_r = \sqrt{G^{-1}\sum_i (r_i - \bar{r})^2}, \qquad A_i = \frac{r_i - \bar{r}}{\sigma_r + 10^{-6}}. \tag{1}$$

Thus the implementation uses the population variance, including abstainers. With no valid voter it creates no pseudo-label and all rewards and advantages are zero; any other zero-variance group also contributes no policy-gradient signal.

A rollout *abstains*—is excluded from the vote though not from normalization—on parse failure or when the configured format requirement is unmet. Historical headline runs are conservatively interpreted as requiring a parseable `\boxed{}` answer: retained code does not establish that their trainer passed an EOS identifier, so historical no-EOS abstention is an open provenance question. The current prospective configurations expose the independent booleans `require_boxed` and `require_eos`; enabling the latter without a tokenizer EOS identifier is an error. This distinction matters because abstention changes who votes, not merely how unfinished rollouts score.

3.2 DECISION POINTS

Let H_t be the entropy of the frozen reference model's next-token distribution at position t of a rollout, computed in float32 over the full vocabulary. The decision-point set $\mathcal{D}$ is the top-p_{dp} fraction of response positions by H_t: for a non-empty response of length L, $k(L) = \min\{L, \max(m_{\min}, \lceil p_{\mathrm{dp}} L \rceil)\}$, with $p_{\mathrm{dp}} = 0.05$ and $m_{\min} = 1$. The executed `topk` call fixes the count but has no explicit secondary rule for exactly tied entropies; tied membership is therefore implementation defined.

3.3 MATCHED DECISION AND RANDOM ARMS

Let $m_t^{\mathrm{sig}}, m_t^{\mathrm{anc}} \in \{0, 1\}$ be per-token masks, $S = \sum_t m_t^{\mathrm{sig}}$, and T the total number of response tokens among non-empty completions in the batch. For rollout-level advantage A_i, define

$$\rho_t = \exp(\log \pi_\theta(a_t \mid h_t) - \mathrm{sg}[\log \pi_{\mathrm{old}}(a_t \mid h_t)]), \tag{2}$$

$$\ell_t^{\mathrm{PG}} = \max\{-A_i \rho_t, -A_i \,\mathrm{clip}(\rho_t, 0.8, 1.2)\}, \tag{3}$$

$$u_t = \log \pi_{\mathrm{ref}}(a_t \mid h_t) - \log \pi_\theta(a_t \mid h_t), \qquad k_{3,t} = e^{u_t} - u_t - 1. \tag{4}$$

The implemented loss, minimized by AdamW, is

$$\mathcal{L} = \underbrace{\frac{\sum_t m_t^{\mathrm{sig}} \ell_t^{\mathrm{PG}}}{D_{\mathrm{PG}}}}_{\text{policy gradient}} + \lambda \underbrace{\frac{\sum_t m_t^{\mathrm{anc}} k_{3,t}}{\sum_t m_t^{\mathrm{anc}}}}_{\text{anchoring}}, \tag{5}$$

where a term with an empty mask is omitted. This is the two-branch clipped GRPO/PPO surrogate, not a separate negative-advantage dual-clip rule. In the executed one-epoch configuration, $\log \pi_{\mathrm{old}}$ is the detached first forward value on the same completion; reference log-probabilities, advantages, masks, and gate entropies are stop-gradient quantities. The prospective estimand fixes $D_{\mathrm{PG}} = T$ (`pg_norm=total`), $\lambda = 0.05$, and changes only the signal mask:

Arm	m^{sig}	m^{anc}	isolates
DECISION	$\mathcal{D}$	$\overline{\mathcal{D}}$	selector under test
RANDOM	random, $\lvert\cdot\rvert = \lvert\mathcal{D}\rvert$	complement	*location*

For each RANDOM completion, the implementation recomputes $k(L)$ on that completion and draws a uniform $k(L)$-subset without replacement. It therefore matches the registered Decision count on its *own* live rollout, not a token count observed in the already-diverged DECISION arm. Conditional on a fixed rollout batch, reward vector and normalizer, the two definitions differ in where the signal lands. In on-policy training the arms subsequently produce different rollouts and pseudo-labels, so the empirical contrast is not a paired-token intervention. Each pair instead estimates the difference between two complete training procedures under a shared nominal rollout/update budget and optimizer contract.

Why an unmatched global baseline cannot be the headline. Under selected-token normalization, a sparse mask changes both location and per-token weight relative to an all-token baseline. Such a contrast can cross the clipping boundary before any token-location effect is identified. Count matching makes Decision–Random the appropriate selector contrast; fixing the total denominator also makes gradient mass auditable across environments. Historical Global and Global+KL arms are retained only as diagnostics (App. I).

3.4 A NORMALIZATION IDENTITY—AND ITS BOUNDARY

For a fixed non-empty mask and rollout batch, let g_S^{PG} and g_T^{PG} denote the policy-gradient component under the two denominators. Directly from Eq. 5,

$$g_S^{\mathrm{PG}} = \frac{T}{S}\, g_T^{\mathrm{PG}}. \tag{6}$$

Thus the pure-PG gradients have identical direction and a norm ratio exactly T/S whenever the gradient is non-zero. This is algebra about the implemented aggregator, not empirical evidence. It does *not* extend to the full gradient in general: with complement-mask KL, $g_S = (T/S)g_T^{\mathrm{PG}} + \lambda g^{\mathrm{KL}}$ whereas $g_T = g_T^{\mathrm{PG}} + \lambda g^{\mathrm{KL}}$. Nor does count matching make Decision and Random gradient vectors equal; it matches only S and therefore the scalar multiplier. Finally, global norm clipping maps g to $g\min(1, c/\|g\|)$: once both arms exceed c, their post-clip norms are equal even though their pre-clip magnitudes differ. A deterministic CPU test of the exact loss verifies Eq. 6, the empty-mask boundary, KL-induced direction changes and clip saturation (App. C).

3.5 ENFORCING THE ISOLATION

The claim that Decision and Random differ only in selector semantics is machine-checked. Resolved configs may differ only in arm and run identity; the validator pins model, tokenizer, ordered data, reward, sampling, optimizer, normalization, checkpoint and evaluation fields. Mask tests enforce equal selected-token counts under a common completion, complement anchoring, and deterministic Random selection. The Random generator seed is $1{,}000{,}003\, s + j$, where s is the training seed and j is the within-update prompt slot; it is not separately keyed by update or rollout. Rollout sampling instead consumes the globally seeded and checkpointed Torch/CUDA stream. Terminal admission additionally requires contiguous updates and hash-valid raw artifacts. These checks establish current implementation isolation; they do not repair missing historical source semantics. Details are in App. A.

4 EXPERIMENTAL SETUP

L-shaped scope design. E00 is the frozen anchor: Qwen2.5-Math-7B (Yang et al., 2024), a deterministic 512-record DeepScaleR Difficulty Level 2 training subset (Agentica Project, 2025), and frozen MATH-500 identifiers derived from the MATH benchmark (Hendrycks et al., 2021). E10 changes only the policy/reference model family to DeepSeekMath-7B-Base (Shao et al., 2024) while retaining E00 data. E01 changes only the training data to a deterministic 512-record English NuminaMath-CoT subset (AI-MO, 2024) while retaining the E00 model. We do not run the

joint model-and-data shift: the current claim requires two orthogonal replications, not a larger grid. Model, tokenizer, dataset, ordered train/evaluation IDs, source, analyzer and renderer are hash-bound before outcome access.

Frozen learning contract. Every prospective arm samples eight rollouts per prompt at temperature one, top-$p = 1$ and disabled top-k truncation, at most 1024 new tokens, eight prompts per update and 120 completed updates. Both arms use consensus reward, standardized group advantages, AdamW at learning rate 10^{-5}, the low-variance k_3 KL estimator with coefficient 0.05, gradient clipping at one, and `pg_norm=total`. A parseable boxed answer may vote; parse failure abstains and EOS is not required. Decision and Random each generate their own on-policy rollout and pseudo-label. The only registered semantic differences are the selector, run identifier and arm name.

Prospective seeds and endpoints. Each environment contains thirteen paired training seeds: 1000–1012 for E00, 2000–2012 for E10, and 3000–3012 for E01. The seed-matched run pair is the independent unit. The only primary endpoint in each environment is Decision minus Random accuracy after 81 completed updates. For the fixed 500 evaluation identifiers, we first compute one accuracy difference $d_s = a_{D,s} - a_{R,s}$ per paired seed, then report $\bar{d}$, a two-sided paired-t test, and $\bar{d} \pm t_{.975,12} s_d/\sqrt{13}$ with the sample standard deviation of the thirteen d_s. The exact sensitivity enumerates all 2^{13} sign assignments and compares absolute means. Within a setting, superiority requires a positive estimate whose two-sided 95% interval lies strictly above zero; non-significance is not equivalence. E00 is the immutable original anchor, while E10 and E01 are separately planned setting-specific replications. Their three update-81 p-values are unadjusted and are not an any-setting discovery family; the broader replication statement instead uses the registered conjunction requiring both E10 and E01 to pass the positive-interval rule. Updates 41 and 120 form exactly four extension secondary tests and receive Holm correction. Update 1 is a descriptive early checkpoint. No outcome-based stopping, seed exclusion, setting replacement or post-hoc primary is allowed. Intervals vary training seeds while conditioning on the fixed evaluation identifiers; they do not quantify problem-population uncertainty.

Technical admission and process deviation. The scheduler may release Random only after the same-seed Decision arm is terminal and technically valid. A pair is admitted only when process states, all 120 update keys, evaluations at updates 1/41/81/120, checkpoint markers and hashes, resolved identities, per-problem ID sets, finite-gradient indicators, attempt lineage, storage and GPU-hours pass the frozen validator. Accuracy, reward, arm ordering and partial contrasts are forbidden inputs to that validator. The registered plan deferred scientific outcome access until all extension pairs were complete. A terminal-debugging command nevertheless displayed intermediate fields after seeds, endpoints and analysis code were frozen. We preserve the incident and full registered sample; the extension is therefore analysis-frozen with a disclosed outcome-blindness process deviation, not outcome-blind admission (App. R).

Checkpoint and update semantics. Future runs use atomic checkpoint format v2, save every 20 completed updates and milestones 41/81/120, retain the latest two plus milestones, and fail closed on identity mismatch. The wrapper requests a signal ten minutes before preemption and serializes only after the current logical update. The loop updates before evaluating, so historical labels 0/40/80 denote 1/41/81 completed updates; all prospective artifacts use completed-update names directly. Detailed model/data revisions, selection statistics, and realized resource accounting are audited in App. N–O. The machine-readable preregistration and raw archive remain internal until a separately packaged artifact passes path and identity sanitization.

5 Controlled evidence

5.1 Localization motivates the intervention, but does not identify its value

Retained summaries reproduce the descriptive premise: positions where a trained policy departs from its base model have much higher base-model entropy. Their historical provenance is incomplete at the per-token level, so we use them only to motivate the selector (App. P). Localization

neither supplies a useful edit direction nor establishes that updating those positions beats updating an equal number elsewhere.

5.2 THE MATCHED ESTIMAND REMOVES THE DOMINANT SCALE AMBIGUITY

For a fixed rollout batch and non-empty signal mask, changing the policy-gradient denominator from selected-token count to total-token count multiplies the pure-PG gradient by exactly T/S (Eq. 6). Historical rows cross the clipping boundary under this choice (Table 12). The prospective comparison therefore fixes `pg_norm=total` for both arms and count-matches their selected positions. This controls the scalar gradient-mass difference; it does not force Decision and Random to have the same gradient direction or subsequent rollout.

5.3 E00: A MATCHED ANCHOR

The frozen E00 experiment comprises thirteen paired training seeds. At the prespecified evaluation after 81 completed updates, Decision minus Random is $+2.12$ percentage points (95% paired CI $[+1.22, +3.02]$; two-sided $p = 0.00024$). An exact sign-flip sensitivity gives $p = 0.00049$; every leave-one-seed-out interval remains above zero. These are run-level results independently recomputed from 26 technically admitted arms. They support Decision only for the E00 model, data, selector and budget—not a universal entropy-gating claim.

5.4 ORTHOGONAL SCOPE TESTS

E10 changes only the model family; E01 changes only the training data. For each environment, we prospectively fixed thirteen seed-matched run pairs, independent Decision/Random on-policy rollouts, the update-81 primary endpoint, and the same total-token normalization. The environments are analyzed separately and are never selectively pooled. The model-family shift (E10) is negative with its paired confidence interval below zero, whereas the training-data shift (E01) is positive. Environments remain separate and are not pooled.

Table 1 reports the prespecified run-pair contrasts, and Figure 2 shows their uncertainty, paired-seed distributions, and learning-budget trajectories without pooling environments. The setting-specific primary p-values are unadjusted; the broader scope statement is controlled by the registered E10-and-E01 conjunction, not by selecting any significant environment. The complete four-test Holm secondary family appears below in Table 2.

Table 1: Prospective normalization-matched Decision–Random comparisons across $n = 13$ seed-matched run pairs per environment. The seed-matched run pair is the inferential unit, update 81 is primary, and Δ = Decision minus Random. The final column gives the preregistered two-sided paired-test p-value and the exact sign-flip sensitivity p-value.

Environment	Δ accuracy (pp)	95% paired CI	Paired p / sign-flip p
Anchor (E00)	+2.12	[+1.22, +3.02]	0.0002 / 0.0005
Model shift (E10)	-2.18	[-2.45, -1.92]	5.51×10^{-10} / 0.0002
Data shift (E01)	+3.03	[+2.23, +3.83]	2.62×10^{-6} / 0.0002

E00 is the frozen anchor. E10 changes only the model family; E01 changes only the training data. Results are not selectively pooled.

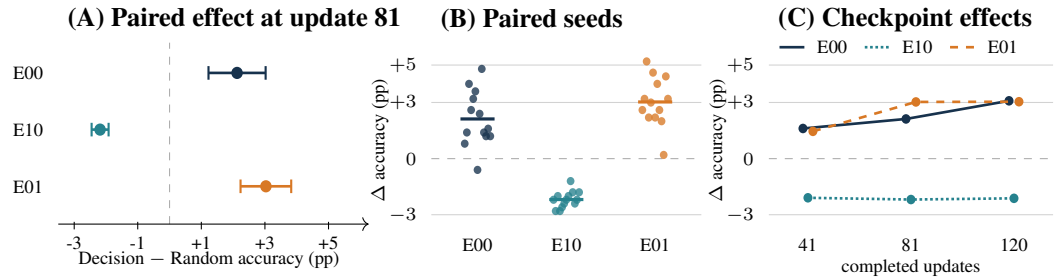


Figure 2: Orthogonal replication of the normalization-matched comparison. (A) Paired-run effects and 95% CIs; (B) all paired-seed differences with horizontal jitter and mean bars; (C) prespecified checkpoint effects with slight horizontal dodging for visibility. E00 is the anchor, E10 changes the model family, and E01 changes training data. Results are shown separately by environment; no pooled estimate is reported.

The registered secondary endpoints preserve the environment-specific directions beyond the primary checkpoint: E10 remains negative and E01 remains positive at both updates 41 and 120 (Table 2). Reporting the complete four-test family guards against selecting only the favorable data-shift result.

Table 2: Registered PHASE-10 extension secondary family. Effects are Decision minus Random over 13 seed-matched run pairs. Raw two-sided paired-test p-values are adjusted together over exactly these four tests by Holm's method; update 1 remains descriptive.

Environment	Update	Δ (pp)	95% paired CI	raw p	Holm p
Model shift (E10)	41	-2.09	[-2.66, -1.52]	3.79×10^{-6}	7.59×10^{-6}
Model shift (E10)	120	-2.12	[-2.52, -1.73]	6.05×10^{-8}	2.42×10^{-7}
Data shift (E01)	41	+1.46	[+0.60, +2.32]	0.0030	0.0030
Data shift (E01)	120	+3.05	[+2.36, +3.73]	4.81×10^{-7}	1.44×10^{-6}

These are raw-recomputed outputs of the frozen numerical analyzer. The disclosed outcome-blindness process deviation applies; environments are not pooled.

5.5 What E00 mechanism rows do and do not show

The raw E00 logs make several single-cause explanations less plausible. Decision responses are longer and have fewer abstentions; correlations with the primary effect are omitted because they are exploratory across only 13 run pairs. Per-sequence selected-token counts are matched by construction, while clipping-incidence and KL differences have intervals spanning zero. Gradient norms differ descriptively, as distinct selectors need not produce the same gradient direction even under a shared denominator. These diagnostics document optimizer quantities and training stability; none proves that entropy itself is the causal mechanism. Parse failure versus missing-box decomposition is not available, and historical EOS enforcement remains an open question. Consequently, the E10 reversal is scoped to the tested model/tokenizer, boxed-answer parser, reward, and training configuration; the available artifacts do not separate those interacting factors. Aggregate prospective voter, informative-group, pseudo-label-quality, and abstention diagnostics are reported descriptively in App. L; they do not identify a cause of either replication result.

Table 3 reports the complete preregistered E00 secondary family rather than a favorable subset. It confirms the final-accuracy and pseudo-label-quality differences while also retaining the null clipping-incidence and KL comparisons; these associations do not identify a mechanism.

Table 3: Complete preregistered secondary family for the E00 normalization-matched Decision–Random experiment. Each row is paired over 13 training seeds; Holm adjustment is across exactly these six tests. Diagnostic means aggregate 120 updates within each run before seed-level inference.

quantity	Decision	Random	Δ	raw p	Holm p
final accuracy (pp)	58.72	55.63	+3.09	$p = 8.91 \times 10^{-6}$	$p = 2.67 \times 10^{-5}$
mean pre-clip grad. norm	0.2577	0.1101	+0.1476	$p = 1.98 \times 10^{-15}$	$p = 1.19 \times 10^{-14}$
clipping incidence (pp)	0.06	0.00	+0.06	$p = 0.3370$	$p = 0.3370$
abstentions / group	1.4928	2.0708	-0.5780	$p = 1.84 \times 10^{-7}$	$p = 9.19 \times 10^{-7}$
pseudo-label accuracy (pp)	70.43	66.23	+4.20	$p = 5.08 \times 10^{-6}$	$p = 2.03 \times 10^{-5}$
mean KL loss	0.0012	0.0010	+0.0002	$p = 0.1003$	$p = 0.2006$

5.6 Failure modes retained as an evaluation checklist

Four checks determine whether a token-selection result can support a method claim: (i) match gradient normalization and report clipping; (ii) separate localization from privileged edit direction and verify counterfactual support; (iii) use independent seed-matched run pairs, not pooled problems, as the inferential units for paired comparisons; and (iv) preserve executable parsing, abstention and provenance semantics. The full historical diagnostics, including null and conflicting results, are retained in App. P; they are not promoted to headline evidence.

6 Discussion

Localization and learning claims come apart at evaluation. Retained summaries support concentration at high-entropy positions, but the causal evidence is oracle-conditional: the oracle supplies edit direction; entropy only location. The historical case study cannot resolve the selector's value, whereas the prospective E00 normalization-matched comparison supports Decision in its tested setting and the orthogonal replications delimit that effect's scope. That bounded ordering becomes interpretable only after controlling the denominator, retaining the seed-matched run pair as the independent unit, and preserving executable provenance. Equation 6, clipping, counterfactual support, statistical unit and abstention semantics can each alter the answer. The central result is therefore both empirical and methodological: location can help, but the claim is only as broad as the evidence gates it passes. Detailed limits are in App. R.

The model-family reversal is especially informative for that distinction. E10 retains the training records, evaluation identifiers, reward, selector definition, optimization contract, and analysis, yet changes the policy/reference model and tokenizer as one coupled family shift. It therefore rules out a family-invariant advantage, but does not identify whether entropy calibration, tokenization, base competence, or pseudo-label quality causes the reversal. E01, meanwhile, shows that the E00 ordering is not tied to one frozen training subset. The protocol thus delivers a falsifiable scope map rather than retrofitting one favorable setting into a universal mechanism claim.

The DPR implementation treats token selection as an explicit mask-producing interface. Rollout clustering and group normalization supply advantages, frozen-reference scoring supplies entropies, and the selector returns signal and complement-anchor masks to a shared clipped loss. The optimizer uses a fixed total-token denominator with common KL, clipping, and stop-gradient semantics. Resolved configurations bind reward parsing, sampling, model/tokenizer identities, ordered data, checkpoints, and evaluation, so each environment instantiates the same executable contract while changing only declared inputs.

7 Conclusion

High-entropy localization does not determine policy-gradient allocation. In matched selector-only runs, Decision favors the anchor and data shift but reverses for the model-family shift. DPR is an auditable protocol for bounded claims, not a universal gating method.

REPRODUCIBILITY STATEMENT

The reported prospective workflow binds every arm to content-addressed model and tokenizer identities (and an upstream revision where retained), ordered training and evaluation identifiers, resolved semantic configuration, source-tree hash, paired seed, and attempt ledger. Its validators accept a pair only after scheduler/process, identity, schema, checkpoint, finite-gradient and resource checks pass. During the extension replication, a debugging command exposed intermediate scientific fields after seeds, endpoints, and analysis code were frozen. We halted submission, preserved the incident, and completed only the unchanged registered work under an explicit author decision; we therefore describe the extension as analysis-frozen with a disclosed outcome-blindness process deviation, not as outcome-blind admission. A frozen analyzer maps raw artifacts to paired-run statistics, generated tables, and vector figures; the reported submission guard rejects partial experiments, identity drift, anonymous-template violations, or unresolved result markers. Historical claims that lack the same minimum provenance remain explicitly documented or open rather than being repaired by current code. The internal repository contains the prospective configs, raw rows, ledgers, frozen analyzers, and raw-to-PDF commands used here, but it is not itself an anonymous release bundle. Independent verification therefore requires a separately path-sanitized package, checked against this manuscript before submission; missing upstream revisions and licenses remain stated as limitations rather than silently inferred.

AI USE STATEMENT

Generative AI tools were used for repository and submission auditing, implementation and debugging support, literature discovery and reference formatting, LaTeX, table and figure preparation, and language editing. They were not treated as experimental evidence and did not autonomously generate or alter experimental results. The authors manually verified citations, code, numerical claims, and AI-assisted text and take responsibility for the final manuscript and artifacts.

REFERENCES

Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron C. Courville, and Marc G. Bellemare. Deep reinforcement learning at the edge of the statistical precipice. In *Advances in Neural Information Processing Systems*, volume 34, pp. 29304–29320, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/f514cec81cb148559cf475e7426eed5e-Abstract.html.

Agentica Project. DeepScaleR preview dataset. Hugging Face dataset, 2025. MIT license; https://huggingface.co/datasets/agentica-org/DeepScaleR-Preview-Dataset.

Arash Ahmadian, Chris Cremer, Matthias Gallé, Marzieh Fadaee, Julia Kreutzer, Olivier Pietquin, Ahmet Üstün, and Sara Hooker. Back to basics: Revisiting REINFORCE-style optimization for learning from human feedback in LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pp. 12248–12267. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.662. URL https://aclanthology.org/2024.acl-long.662/.

AI-MO. NuminaMath-CoT. Hugging Face dataset, 2024. Official dataset release; Apache-2.0 license.

Ömer Faruk Akgül, Rajgopal Kannan, Willie Neiswanger, and Viktor Prasanna. Rethinking RL for LLM reasoning: It's sparse policy selection, not capability learning. *arXiv preprint arXiv:2605.06241*, 2026.

Marcin Andrychowicz, Anton Raichuk, Piotr Stańczyk, et al. What matters in on-policy reinforcement learning? a large-scale empirical study. *arXiv preprint arXiv:2006.05990*, 2020. doi: 10.48550/arXiv.2006.05990. URL https://arxiv.org/abs/2006.05990.

Peter Chen, Xiaopeng Li, Ziniu Li, Wotao Yin, Xi Chen, and Tianyi Lin. Exploration vs exploitation: Rethinking RLVR through clipping, entropy, and spurious reward. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/d99e8e80a6c41e148db686918dd7eab3-Abstract-Conference.html. arXiv:2512.16912.

Daixuan Cheng, Shaohan Huang, Xuekai Zhu, Bo Dai, Xin Zhao, Zhenliang Zhang, and Furu Wei. Reasoning with exploration: An entropy perspective. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 30377–30385, 2026. doi: 10.1609/aaai.v40i36.40290. URL https://ojs.aaai.org/index.php/AAAI/article/view/40290.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021. doi: 10.48550/arXiv.2110.14168. URL https://arxiv.org/abs/2110.14168.

Cédric Colas, Olivier Sigaud, and Pierre-Yves Oudeyer. How many random seeds? statistical power analysis in deep reinforcement learning experiments. *arXiv preprint arXiv:1806.08295*, 2018. doi: 10.48550/arXiv.1806.08295. URL https://arxiv.org/abs/1806.08295.

Ganqu Cui, Lifan Yuan, Zefan Wang, et al. Process reinforcement through implicit rewards. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856. URL https://openreview.net/forum?id=9SkkifLopZ. arXiv:2502.01456.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pp. 1383–1392. Association for Computational Linguistics, 2018. doi: 10.18653/v1/P18-1128. URL https://aclanthology.org/P18-1128/.

Logan Engstrom, Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Firdaus Janoos, Larry Rudolph, and Aleksander Madry. Implementation matters in deep policy gradients: A case study on PPO and TRPO. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=r1etN1rtPB.

Yuqian Fu, Tinghong Chen, Jiajun Chai, Xihuai Wang, Songjun Tu, Guojun Yin, Wei Lin, Qichao Zhang, Yuanheng Zhu, and Dongbin Zhao. SRFT: A single-stage method with supervised and reinforcement fine-tuning for reasoning. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/a79699db176ed0efc04a9da171e52112-Abstract-Conference.html. arXiv:2506.19767.

Daya Guo et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645(8081):633–638, 2025. doi: 10.1038/s41586-025-09422-z. URL https://www.nature.com/articles/s41586-025-09422-z.

Zhezheng Hao, Hong Wang, Haoyang Liu, Jian Luo, Jiarui Yu, Hande Dong, Qiang Lin, Can Wang, and Jiawei Chen. Rethinking entropy interventions in RLVR: An entropy change perspective. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics*, pp. 31105–31133. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.1436. URL https://aclanthology.org/2026.acl-long.1436/. ACL 2026 Outstanding Paper; arXiv:2510.10150.

Yuhang He, Haodong Wu, Siyi Liu, Hongyu Ge, Hange Zhou, Keyi Wu, Zhuo Zheng, Qihong Lin, Zixin Zhong, and Yongqi Zhang. Where hindsight credit can reside: A signed-capacity view of token updates in RLVR. *arXiv preprint arXiv:2604.11056*, 2026. doi: 10.48550/arXiv.2604.11056. URL https://arxiv.org/abs/2604.11056.

Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. Deep reinforcement learning that matters. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, pp. 3207–3214, 2018. doi: 10.1609/aaai.v32i1.11694. URL https://ojs.aaai.org/index.php/AAAI/article/view/11694.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. arXiv:2103.03874.

Renren Jin, Pengzhi Gao, Yuqi Ren, Zhuowen Han, Tongxuan Zhang, Wuwei Huang, Wei Liu, Jian Luan, and Deyi Xiong. Revisiting entropy in reinforcement learning for large reasoning models. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 25300–25322. Association for Computational Linguistics, 2026a. doi: 10.18653/v1/2026.findings-acl.1266. URL https://aclanthology.org/2026.findings-acl.1266/. arXiv:2511.05993.

Senjie Jin, Peixin Wang, Boyang Liu, Xiaoran Fan, Shuo Li, Zhiheng Xi, Jiazheng Zhang, Yuhao Zhou, Tao Gui, Qi Zhang, and Xuanjing Huang. Entropy is not enough: Unlocking effective reinforcement learning for visual reasoning via vision-anchored token selection. *arXiv preprint arXiv:2606.03937*, 2026b. doi: 10.48550/arXiv.2606.03937. URL https://arxiv.org/abs/2606.03937.

Amirhossein Kazemnejad, Milad Aghajohari, Eva Portelance, Alessandro Sordoni, Siva Reddy, Aaron Courville, and Nicolas Le Roux. VinePPO: Refining credit assignment in RL training of LLMs. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 29557–29590. PMLR, 2025. URL https://proceedings.mlr.press/v267/kazemnejad25a.html. arXiv:2410.01679.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/aca97732e30bcf1303bc22ac3924fd16-Abstract-Conference.html. arXiv:2305.20050.

Zihan Lin, Xiaohan Wang, Jie Cao, Jiajun Chai, Guojun Yin, Wei Lin, and Ran He. ResT: Reshaping token-level policy gradients for tool-use large language models. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/8730837e119f6b5d1b25f6da1dc2c28a-Abstract-Conference.html. arXiv:2509.21826.

Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding R1-zero-like training: A critical perspective. *arXiv preprint arXiv:2503.20783*, 2025. doi: 10.48550/arXiv.2503.20783. URL https://arxiv.org/abs/2503.20783.

Outongyi Lv, Yanzhao Zheng, Yuanwei Zhang, Zhenghao Huang, Xingjun Wang, Baohua Dong, Hangcheng Zhu, and Yingda Chen. Which tokens matter? adaptive token selection for RLVR with the relative surprisal index. *arXiv preprint arXiv:2606.31575*, 2026. doi: 10.48550/arXiv.2606.31575. URL https://arxiv.org/abs/2606.31575.

Yixiu Mao, Yun Qu, Qi Wang, Heming Zou, and Xiangyang Ji. RLVR without ineffective samples: Group prioritized off-policy optimization for LLM reasoning. *arXiv preprint arXiv:2606.01281*, 2026.

Haoming Meng, Kexin Huang, Shaohang Wei, Chiyu Ma, Shuo Yang, Xue Wang, Guoyin Wang, Bolin Ding, and Jingren Zhou. Sparse but critical: A token-level analysis of distributional shifts in RLVR fine-tuning of LLMs. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/c9b7e6175f2bd3c2824f24aa7ac313d6-Abstract-Conference.html. arXiv:2603.22446.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, et al. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pp. 27730–27744, 2022. doi: 10.52202/068431-2011.

Andrew Patterson, Samuel Neumann, Martha White, and Adam White. Empirical design in reinforcement learning. *Journal of Machine Learning Research*, 25(318):1–63, 2024. URL https://www.jmlr.org/papers/v25/23-0183.html.

Aleksei Petrenko, Ben Lipkin, Kevin Chen, Erik Wijmans, Marco Cusumano-Towner, Raja Giryes, and Philipp Krähenbühl. Entropy-preserving reinforcement learning. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/124dde499d62b58e97e42a45b26d7369-Abstract-Conference.html. arXiv:2603.11682.

John Schulman, Philipp Moritz, Sergey Levine, Michael I. Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. In *International Conference on Learning Representations*, 2016. URL https://arxiv.org/abs/1506.02438.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. doi: 10.48550/arXiv.1707.06347. URL https://arxiv.org/abs/1707.06347.

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.

Avi Singh, John D. Co-Reyes, Rishabh Agarwal, et al. Beyond human data: Scaling self-training for problem-solving with language models. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=lNAyUngGFK. arXiv:2312.06585.

Shenzhi Wang, Le Yu, Chang Gao, Chujie Zheng, Shixuan Liu, Rui Lu, Kai Dang, Xiong-Hui Chen, Jianxin Yang, Zhenru Zhang, Yuqiong Liu, An Yang, Andrew Zhao, Yang Yue, Shiji Song, Bowen Yu, Gao Huang, and Junyang Lin. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for LLM reasoning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, 2025. doi: 10.52202/085713-3850. arXiv:2506.01939.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=1PL1NIMMrw. arXiv:2203.11171.

Xuexiang Wen, Hang Yu, Linchao Zhu, and Gaoang Wang. Verifier-free RL for LLMs via intrinsic gradient-norm reward. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 32089–32102. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.findings-acl.1606. arXiv:2605.09920; https://aclanthology.org/2026.findings-acl.1606/.

Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8:229–256, 1992. doi: 10.1007/BF00992696.

Jianghao Wu, Yasmeen George, Jin Ye, Yicheng Wu, Daniel F. Schmidt, and Jianfei Cai. SPINE: Token-selective test-time reinforcement learning with entropy-band regularization. *arXiv preprint arXiv:2511.17938*, 2025. doi: 10.48550/arXiv.2511.17938. URL https://arxiv.org/abs/2511.17938.

An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, Keming Lu, Mingfeng Xue, Runji Lin, Tianyu Liu, Xingzhang Ren, and Zhenru Zhang. Qwen2.5-Math technical report: Toward mathematical expert model via self-improvement. *arXiv preprint arXiv:2409.12122*, 2024.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, et al. DAPO: An open-source LLM reinforcement learning system at scale. In *Advances in Neural Information Processing Systems*, volume 38, 2025a. doi: 10.52202/085713-3775. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/a4277440d50f1f15d2cb4c14f7e0c0d2-Abstract-Conference.html. arXiv:2503.14476.

Tianyu Yu, Bo Ji, Shouli Wang, Shu Yao, Zefan Wang, Ganqu Cui, Lifan Yuan, Ning Ding, Yuan Yao, Zhiyuan Liu, Maosong Sun, and Tat-Seng Chua. RLPR: Extrapolating RLVR to general domains without verifiers. *arXiv preprint arXiv:2506.18254*, 2025b.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Xian Li, Sainbayar Sukhbaatar, Jing Xu, and Jason E. Weston. Self-rewarding language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 57905–57923. PMLR, 2024. URL https://proceedings.mlr.press/v235/yuan24d.html. arXiv:2401.10020.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah D. Goodman. STaR: Bootstrapping reasoning with reasoning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 15476–15488, 2022. doi: 10.52202/068431-1126. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/639a9a172c044fbb64175b5fad42e9a5-Abstract-Conference.html.

Xuandong Zhao, Zhewei Kang, Aosong Feng, Sergey Levine, and Dawn Song. Learning to reason without external rewards. In *International Conference on Learning Representations (ICLR)*, 2026. arXiv:2505.19590.

Yuxin Zuo, Kaiyan Zhang, Li Sheng, Shang Qu, Ganqu Cui, Xuekai Zhu, Haozhan Li, Yuchen Zhang, Xinwei Long, Ermo Hua, Biqing Qi, Youbang Sun, Zhiyuan Ma, Lifan Yuan, Ning Ding, and Bowen Zhou. TTRL: Test-time reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, 2025. doi: 10.52202/085713-4376. arXiv:2504.16084.

A ARM ISOLATION: WHAT IS MACHINE-CHECKED

The claim that non-arm configuration fields are matched and each arm's signal and anchor masks follow their declared definitions is a claim about code, so current configurations are tested directly. The targeted suite fails when run if any of the following does not hold. Historical process-dumped configs are now recovered, but without their source snapshot they do not prove which defaults or call paths the old processes consumed.

- **Configuration isolation.** All four arm configurations are loaded, flattened, and pairwise compared; the symmetric difference must be a subset of $\{\texttt{arm}, \texttt{exp_name}\}$. A second test pins the knobs whose divergence would silently invalidate the comparison — seed, learning rate, clip ranges, KL coefficient and type, group size, temperature, token budget, reward type, boxed/EOS requirements, dataset, step count, p_{dp} — and asserts they are identical across arms.
- **No gold leakage.** Every verifier-free arm must use the consensus reward; the gold reward is permitted only in the explicitly-labeled control arm.
- **Mask invariants.** On a common completion, the RANDOM mask has the same count as the hypothetical DECISION mask computed from that completion, under both the fraction and the threshold selector. This is not a claim that independently evolving B/C completions share a cross-arm token count. Signal and anchor masks are disjoint for GLOBAL/DECISION/RANDOM and both all-ones for GLOBAL+KL; RANDOM's mask is deterministic given training seed and within-update prompt slot; all arms return masks of identical shape for identical input; empty sequences are handled. The Random key omits update and rollout indices, so a slot can reuse a permutation for equal-length completions.

B IMPLEMENTATION PROVENANCE

The trainer is a compact re-implementation rather than a fork, so that the per-token loss is explicit and the mask isolation is auditable. The implementation records line-level correspondences to `verl` 0.4.0 for group-normalized advantage, the two-branch clipped policy loss, the $k_1/k_2/k_3$ KL estimators (we use the low-variance k_3), and the masked loss aggregator. There is no third negative-advantage dual-clip branch or constant in the executed function. Full-vocabulary token entropy is computed as $H = \mathrm{logsumexp}(\ell) - \mathbb{E}_p[\ell]$.

Claim provenance. A bounded search of only paths recorded in retained DPR configs and reports recovered 345 small artifacts (59.2 MB copied, all hash-verified and schema-valid), plus model metadata and dataset hashes. Stage-1 metrics and per-problem rows independently reproduce the arm table and statistical-unit example. The machine-readable manifest still marks a claim incomplete when its minimum set lacks source, rollouts, checkpoints, or semantic identity. Current corrected semantics are never used to infer historical semantics.

Table 4: Historical evidence disposition. A is raw-complete within its stated boundary; B is documented/partial and motivational; C is conflicting or administrative and is further downgraded. Prospective E00/E10/E01 evidence is tracked separately.

evidence item	class	permitted paper use / remaining limitation
Statistical unit	A	appendix statistical-unit evidence; does not establish historical arm ordering
Four-arm case study	B	underpowered historical case study in appendix
Counterfactual audit	B	motivating counterfactual-construction audit; does not support a causal-mechanism claim
Clipping audit	B	optimizer-scale diagnostic in appendix; historical logging conflict retained
Entropy localization	B	motivating localization premise; per-token replay unavailable
Format portability	B	same-rollout format audit in appendix; not cross-family method evidence
Power analysis	B	retrospective planning arithmetic only
Update naming	B	current-code/report-corroborated naming offset; historical source identity incomplete
Abstention semantics	C	historical EOS enforcement remains an open question; current semantics not projected backward
Config reconstruction	C	reconstruction only; never presented as historical resolved configuration
GPU ledger	C	administrative appendix/report evidence only; 73-job historical counting rule unresolved

Numerical precision. Entropy and log-probabilities are computed in float32 even when the model runs in bfloat16. H is a difference of two same-magnitude quantities. For $u = \log \pi_{\text{ref}} - \log \pi_\theta$, the k_3 estimator $\exp(u) - u - 1$ is nonnegative and is added to the minimized loss to anchor complement tokens toward the reference; it is sensitive to noise near $u = 0$. In bfloat16 both entropy and this estimator are dominated by rounding. Entropy separation is the measurement this paper's premise rests on, so the upcast is not optional.

C THE NORMALIZATION CONTROL, VERIFIED IN ISOLATION

Equation 6 follows by pulling the denominators outside the same PG numerator: for $S > 0$, $(1/S)\sum_t m_t \nabla \ell_t = (T/S)(1/T)\sum_t m_t \nabla \ell_t$. A deterministic CPU fixture using the actual `pg_per_token` and k_3 functions verifies ratio 4.000000 at $S/T = 0.25$ and cosine 1.000000 for the isolated PG term. It also verifies that an empty mask contributes zero PG without division, complement-mask KL can change the combined direction, and global clipping reports pre/post norms and saturation. These are mechanism checks, not training seeds or evidence that Decision locations outperform Random locations.

In the recovered headline rows, the original gated arm has mean pre-clip norm 5.0128 over four seeds, while the global-denominator control has 0.2510 over two seeds; the derived fraction above threshold changes from 100% to 0%. Because these are distinct on-policy runs, their ratio is an empirical diagnostic rather than Eq. 6 evaluated on one batch.

A deviation and why. The natural form of this control is to scale the *global* arm's learning rate up by $1/p_{\text{dp}}$ to match the gated arm's per-token weight. That is 2×10^{-4} here, and we had already measured that 10^{-4} caused training collapse (reward $\to 0.000$, gradient norm $\to 0.000$, every completion truncating). Scaling the gated arm's normalizer down achieves the same isolation without a learning rate known to be catastrophic.

D CHOOSING THE LEARNING RATE

The learning rate was measured, not inherited. At 10^{-6}, AdamW's total parameter movement over 40 steps is $\sim 4 \times 10^{-5}$ against a weight scale of $\sim 2 \times 10^{-2}$, and held-out accuracy does not move at all. A saturation test on four fixed problems drives reward $0.03 \to 0.75$ at 10^{-5} and causes training collapse at 10^{-4}. At our batch size (8 prompts/step), the 10^{-5} setting produced an observed $+7.8$-point held-out gain under a verifiable reward at step 80 ($p < 10^{-4}$, paired, $n_{\text{eval}} = 500$). These

diagnostics motivate the selected rate but do not substitute for a benchmark-disjoint hyperparameter-selection record.

E THE INFORMATIVE BAND

Verifier-free training is only meaningful where the consensus label is noisy but not garbage. We measured pseudo-label accuracy against gold on 200 problems per configuration and required $[0.40, 0.80]$.

base model / dataset	pseudo-label acc.	pass@1	informative groups
Qwen2.5-7B / DeepScaleR	0.335	0.200	0.985
Qwen2.5-Math-7B / DeepScaleR	0.430	0.241	0.975
Qwen2.5-Math-7B / Level 2 (used)	**0.685**	**0.403**	**0.970**
Qwen2.5-Math-7B / Level 1	0.820	0.574	0.965
Qwen2.5-Math-1.5B / Level 2	0.548	0.318	0.970

Bold marks the selected E00 model/data configuration. Pseudo-label accuracy and pass@1 move together across these rows, so this table does not isolate label quality from problem difficulty; it is descriptive rather than a clean noise intervention.

The base Qwen2.5-7B pass@1 conflict is now adjudicated by the recovered hash-verified `diagnostics.json`: it records 0.200 over 1,600 completions. The old 0.241 entry was the Qwen2.5-Math-7B value copied into the wrong row.

F ABSTENTION RULES AND FORMAT PORTABILITY

The historical training semantics are conservatively "parseable and boxed": trainer and pseudo-label diagnostic call sites did not pass an EOS identifier, whereas one full diagnostic did. Consequently, the previously stated EOS+boxed historical rule is unsupported. Recovered diagnostic rollouts and the archival reanalysis record that the stricter offline EOS+boxed rule changes pseudo-label accuracy from 0.430 to 0.428 while reducing mean voters from 7.90 to 4.12. This is a same-rollout offline diagnostic, not historical trainer behavior; its historical parser/source environment is still missing.

In the retained but incompletely identified DeepSeek-family sample, the boxed-format rule abstains on 81% of rollouts because the sample emits `\boxed` in only 19.1% of completions (Table 14). The exact checkpoint/tokenizer identity is unavailable, so this is a motivating historical audit rather than a reproducible family-level estimate. Abstention rates should be reported per model family before accuracy is reported.

For future runs, `reward.require_boxed` and `reward.require_eos` are independent, explicit fields consumed by trainer and diagnostics through the same labeling function. The historical-compatible default is `require_eos=false`; enabling it without a tokenizer EOS id fails loudly. Tests cover legal boxed completions, emitted and missing EOS, parse failures, missing boxed answers, and call-site consistency.

G CHECKPOINT NAMING, EXACT RESUME, AND RECONSTRUCTED CONFIGURATION

The loop updates before evaluating: historical labels 0/40/80 therefore denote 1/41/81 completed updates, while the final result is after 120. Current outputs additionally persist `updates_completed`. Future resumable runs commit state only after a full update and metric row, using a same-filesystem temporary directory, state/metadata hashes, a completion marker written last, and one atomic rename. Policy, optimizer, all Python/NumPy/PyTorch RNG states, sampler cursor/order hash and optional EMA state are restored; immutable reference/model/tokenizer, dataset, semantic-config and source hashes are validated. Attempts write separate JSONL segments, and duplicate/gapped $(\text{run}, \text{completed updates}, \text{phase})$ keys fail closed. CPU new-process tests produce bitwise-identical weights, optimizer state, prompt sequence and subsequent RNG draws. SIGUSR1

only sets a flag; serialization occurs after the current update and exits with a distinct interrupted code rather than success. This is current implementation evidence, not historical repair.

The file `configs/paper_case_study.yaml` reconstructs shared settings from retained reports and daemon overrides. Recovered per-run config dumps supersede part of that gap, but direct arm YAML files and this reconstruction are not historical source snapshots.

H GENERATION BUDGET

Raising the token budget from 1024 to 1536 halves truncation ($0.364 \rightarrow 0.191$) and raises pass@1 by 1.8 points, but leaves pseudo-label accuracy statistically unchanged ($0.430 \rightarrow 0.414$, s.e. 3.5 percentage points). Truncation is therefore not what limits consensus quality, and we keep 1024. The decoupling is itself informative: given more room the model also produces more varied *wrong* answers, fragmenting the vote.

I FULL FOUR-ARM RESULTS

Table 5: Historical case study at the preregistered evaluation after 81 completed updates. Means are recovered across four training seeds. The primary method-level contrast uses paired runs; the pooled-problem statistic is supplementary. No row supports an arm ordering.

arm	signal positions	anchor positions	mean acc.	role
GLOBAL	all	none	0.5910	status quo
DECISION	high entropy	complement	0.5745	decision selector
RANDOM	count-matched random	complement	0.5585	location control
GLOBAL+KL	all	all	0.5800	KL control
Primary DECISION−RANDOM		$+1.60$ pp; paired-run $p = .211$ (pooled-problem $p = .043$)		

Table 6: Held-out accuracy recomputed from hash-verified metrics, pooled descriptively over seeds. Label 80 follows 81 updates; label 120 is the final evaluation after 120 updates.

reward	label	A: global	B: decision	C: random	D: global+KL
consensus	80	0.5910	0.5745	0.5585	0.5800
	120	0.6050	0.5750	0.5510	0.5970
gold	80	0.5840	0.5750	0.5500	0.5750
	120	0.5955	0.5890	0.5540	0.5800

Table 7: Contrasts recomputed from matched per-problem rows; unadjusted two-sided p-values. No primary seed-level contrast is significant. The $p = .049$ secondary cell is unadjusted and does not establish an arm ordering; pooled tests are supplementary.

reward	contrast	label	effect (pp)	seed p	pooled p
consensus	GLOBAL−DECISION	80	$+1.65$	0.080	0.025
	DECISION−RANDOM	80	$+1.60$	0.211	0.043
	DECISION−GLOBAL+KL	80	-0.55	0.140	0.476
	GLOBAL−DECISION	120	$+3.00$	0.049	$< 10^{-4}$
	DECISION−RANDOM	120	$+2.40$	0.058	0.002
	DECISION−GLOBAL+KL	120	-2.20	0.088	0.002
gold	GLOBAL−DECISION	80	$+0.90$	0.593	0.245
	GLOBAL−GLOBAL+KL	80	$+0.90$	0.219	0.229
	GLOBAL−DECISION	120	$+0.65$	0.313	0.407
	GLOBAL−GLOBAL+KL	120	$+1.55$	0.110	0.033

J POWER TABLE

Table 8: Normal-approximation seeds required for 80% power at $\alpha = 0.05$ two-sided, using the documented effect and run-to-run SD. Retrospective inputs do not establish the effects.

contrast	effect (pp)	sd (pp)	seeds have	seeds need
consensus GLOBAL−DECISION, final	+3.00	1.87	4	4
consensus DECISION−RANDOM, final	+2.40	1.61	4	4
consensus GLOBAL−DECISION, label 80	+1.65	1.27	4	5
consensus DECISION−GLOBAL+KL, final	−2.20	1.77	4	6
consensus DECISION−RANDOM, label 80	+1.60	2.03	4	13
gold GLOBAL−DECISION, label 80	+0.90	3.00	4	88

K E00 SEED ROBUSTNESS AND DESCRIPTIVE MECHANISM CHECKS

Seed robustness. The exact 2^{13} sign-flip sensitivity gives $p = 0.000488$. Across leave-one-seed-out fits, the mean ranges from 1.90 to 2.35 percentage points and the lower 95% confidence limit from 1.05 to 1.53 points.

Table 9: Raw-recomputed E00 diagnostic contrasts, Decision minus Random, aggregated over 13 paired training seeds. Intervals are descriptive and unadjusted; these rows diagnose comparability and stability but do not identify why token location changes accuracy.

diagnostic	paired mean	95% descriptive CI
response length (tokens)	+17.86	[+11.87, +23.85]
abstentions per group	-0.578	[-0.696, -0.460]
selected-token fraction	-0.000023	[-0.000028, -0.000018]
pre-clip gradient norm	+0.148	[+0.141, +0.154]
post-clipping gradient norm	+0.147	[+0.141, +0.153]
clipping incidence	+0.0006	[-0.0008, +0.0020]
KL loss	+0.000228	[-0.000051, +0.000508]

Parse-failure versus missing-box decomposition is not present in the raw rows, and historical EOS enforcement remains an open question. Correlations with the primary effect are intentionally omitted from the paper table because they are exploratory across only 13 runs and cannot establish mediation.

L EXTENSION SECONDARY AND SCORING DIAGNOSTICS

The registered E10/E01 secondary family in Table 2 and the descriptive scoring audit below are regenerated from the same raw-recomputed PHASE-10 summary used for the primary results. The scoring table consumes every one of the 6,240 training-update rows from the 52 extension arms, averages updates within run, and then averages the thirteen run means. No diagnostic in the table is promoted to a mechanism test.

Table 10: Prospective scoring and aggregate-abstention diagnostics. Values first average 120 update-level records within each run and then average the 13 run means; they are descriptive, not additional registered tests.

Environment	Arm	Abstain/group	Voters/group	Informative/update	Pseudo-label acc.
Model shift (E10)	Decision	1.354	6.646	5.988/8 (74.85%)	0.17%
Model shift (E10)	Random	4.181	3.819	7.891/8 (98.64%)	1.16%
Data shift (E01)	Decision	1.410	6.590	6.481/8 (81.02%)	53.03%
Data shift (E01)	Random	1.786	6.214	6.897/8 (86.22%)	52.05%

Pseudo-label accuracy is the logged gold diagnostic among groups with a known majority label; gold does not enter the consensus training reward. Abstention combines parse failure and missing boxed answer (EOS was optional). The raw schema does not retain their separate counts, so reason-specific values are NOT AVAILABLE. These downstream diagnostics cannot identify the cause of either replication result.

M PROSPECTIVE INFERENCE SPECIFICATION

For environment e, arm $a \in \{D, R\}$, training seed s, and the fixed set of $J = 500$ evaluation identifiers, let

$$a_{eas} = J^{-1} \sum_{j=1}^{J} z_{easj}, \quad d_{es} = a_{eDs} - a_{eRs}, \quad \bar{d}_e = 13^{-1} \sum_{s=1}^{13} d_{es}. \tag{7}$$

The reported standard deviation is $s_{d,e}^2 = (13-1)^{-1} \sum_s (d_{es} - \bar{d}_e)^2$; the paired statistic is $t_e = \bar{d}_e / (s_{d,e}/\sqrt{13})$ with 12 degrees of freedom, and the two-sided 95% interval is $\bar{d}_e \pm t_{.975,12} s_{d,e}/\sqrt{13}$. The exact sensitivity is

$$p_{\text{sf}} = 2^{-13} \sum_{\epsilon \in \{-1,+1\}^{13}} \mathbf{1}\left\{ \left| 13^{-1} \sum_s \epsilon_s d_{es} \right| \geq |\bar{d}_e| \right\}, \tag{8}$$

using all sign assignments, the absolute mean, and no Monte Carlo or plus-one correction. E00, E10, and E01 are separate setting-specific primaries with unadjusted p-values; they are not pooled and do not form a post-hoc any-setting discovery family. The registered broad-scope gate is the conjunction of positive E10 and E01 estimates whose paired intervals both exclude zero. Holm adjustment is confined to the four extension tests at updates 41 and 120. These statements concern variation across training seeds conditional on the fixed evaluation identifiers.

N DEVELOPMENT REUSE AND OVERLAP AUDIT

The anchor setting is not an untouched-benchmark confirmation. Earlier development compared Qwen2.5 and Qwen2.5-Math models using pseudo-label diagnostics, compared DeepScaleR difficulty tiers, tested learning rates 10^{-6}, 10^{-5}, and 10^{-4}, and compared 1024- and 1536-token generation budgets. MATH-500 was inspected repeatedly during trainer and budget diagnostics, and the observed update-81 behavior informed the endpoint choice. The repository does not retain distinct, complete tuning-set ID registries for these decisions. We therefore call E00/E10/E01 *conditional benchmark replications at frozen MATH-500 identifiers*, not untouched-test population confirmation.

The E01 subset was selected before its B/C outcomes from a pinned official NuminaMath-CoT snapshot; the exact revision is retained in the machine-readable experiment manifest rather than printed here. Its recorded rule removes normalized exact duplicates and detected near-duplicates against DeepScaleR-L2, MATH-500, and already selected Numina records using five-token-shingle Jaccard ≥ 0.85 after a SimHash candidate lookup. Seven candidates were rejected by the combined reference-or-selected check before 512 records were frozen. This is a deterministic overlap screen, not a guarantee against every semantic duplicate. A separate execution-identity audit exhaustively compared the frozen 512 DeepScaleR-L2 records with the 500 MATH evaluation records under normalized exact matching and the same five-token-shingle Jaccard threshold. It found zero exact and zero thresholded near-duplicate pairs among 256,000 comparisons. This rules out overlap only under the registered text rules; semantic overlap below the threshold remains possible.

O PROSPECTIVE IDENTITIES AND REALIZED RESOURCES

The source, model, tokenizer, and dataset snapshots used by E00/E10/E01 were content-addressed before analysis. Exact revisions and digests are retained in the machine-readable experiment manifests and are omitted here for readability. E00/E01 use Qwen2.5-Math-7B under Apache-2.0; its exact executed bundle is recorded, although the upstream checkpoint revision was not retained. E10 uses a pinned DeepSeekMath-7B-Base checkpoint and safetensors conversion under DeepSeek License v1.0. NuminaMath-CoT is pinned under Apache-2.0. The exact DeepScaleR-L2 and MATH-500 files are likewise content-addressed, but their upstream revisions and the MATH-500 license were not bound in the execution manifest. The repository itself also lacks a root code license, so redistribution terms must be resolved before releasing an anonymous artifact.

Every prospective scientific arm requested two H800 GPUs, 32 CPUs, 220 GiB RAM, and at most eight hours. Hash-verified terminal-ledger recomputation records 384.980 GPU-hours for E00, 375.656 for E10 (172.951 Decision; 202.705 Random), and 376.447 for E01 (186.563 Decision; 189.884 Random); the E10 compatibility smoke adds 0.336 GPU-hours. The E10/E01 raw training rows contain 399,360 rollouts and 245,338,871 training-completion tokens: 57,416,789/54,848,251 for E10 Decision/Random and 66,850,556/66,223,275 for E01 Decision/Random. These counts exclude prompt and frozen-evaluation tokens, which were not logged comprehensively. Cumulative full-arm allocation times, including recovery allocations, range from 6.21–7.61 h (median 6.47) for E10 Decision, 7.63–9.20 h (median 7.67) for E10 Random, 6.89–7.75 h (median 7.04) for E01 Decision, and 7.02–8.54 h (median 7.11) for E01 Random. Preserved lineage records the failed-then-replaced E10 compatibility smoke and exactly two multi-attempt scientific arms: E10 Random seed 2006 and E01 Random seed 3004, each with two recovery allocations after its initial allocation.

The scheduler always released Decision before same-seed Random. Both requested GPUs showed memory and compute in bounded runtime samples, but no complete driver/container snapshot or paired system-load drift audit was retained; the jobs used the documented `knowrl` conda environment rather than a container. The project environment records Python 3.10, CUDA 12.4, PyTorch 2.6.0, Transformers 4.51.1, and `math-verify` 0.7.0, but that file is not a per-job package proof. Consequently, we claim matched nominal configuration and resource envelopes, not matched realized token count, wall-clock load, or system state.

P HISTORICAL DIAGNOSTICS AND EVIDENCE BOUNDARIES

These retained studies motivated the controls used in the prospective experiments. Their numerical rows are preserved for auditability, but incomplete historical source, rollout, or semantic identity keeps them below the raw-complete standard except where stated explicitly.

P.1 LOCALIZATION SUMMARY

The reported analysis forces a verifiable-reward "teacher" rollout through the base model and marks positions where the base-model argmax differs from the teacher token.

Table 11: Documented localization replication. Hash-verified per-problem summaries were recovered; historical per-token logits and masks required to replay exclusions were not.

	T=0.6, top-p 0.95	T=1.0, top-p 1.0	reported originally
reranked fraction	5.83%	7.35%	1–4%
teacher token in base top-5	95.7%	93.6%	"always"
entropy at reranked (nats)	0.9015	0.9426	higher
entropy at unchanged (nats)	0.0987	0.1134	—
ratio / Cohen's d	9.13× / 1.909	8.31× / 1.824	—
paired t over problems	48.72	45.47	—

The summaries support entropy localization, not a token-selection learning advantage.

P.2 OPTIMIZER-SCALE ROWS

Table 12: Recovered pre-clip norms and derived clipping incidence in the historical four-arm study (threshold 1.0). The historical `clipfrac` field was incorrectly constant zero.

arm	mean pre-clip norm	updates above clip	post-clip norm
GLOBAL	0.406	0.0%	0.406
DECISION	5.013	100.0%	1.0
RANDOM	2.176	100.0%	1.0
GLOBAL+KL	0.431	0.0%	0.431

These rows diagnose a nonlinear optimizer-scale confound; they do not measure AdamW parameter-step norms. The later total-denominator Decision control moved to the unclipped regime without resolving the historical selector ordering, so the confound was real but is not a post-hoc explanation of that ordering.

P.3 COUNTERFACTUAL CONSTRUCTION

One operationalization of token leverage asks whether perturbing gated positions changes the final answer more than perturbing others. We built four successive intervention designs. Their documented effects disagree in *sign*, showing that the estimand changes with the intervention.

Table 13: Four intervention designs. "Effect" is the high-minus-low difference in each design's own outcome; a positive value would indicate more leverage under that intervention. These are reported summaries rather than raw-recomputed estimates.

design	downstream	effect	diagnostic
v1 truncate-and-resample	T=1.0	$+0.039$ ($t = 2.26$)	downstream noise dominates
v2 free-vs-fixed	T=1.0	-0.004 ($t = -0.21$)	MDE 0.053: underpowered
v3 forced counterfactual	greedy	-0.039 ($t = -1.84$)	injects 22.6 nats more at low-H
v4 surprisal-controlled	greedy	-0.101 ($t = -3.06$)	only 8.6% common support

The two biases run in opposite directions. Natural resampling (v1, v2) favors high-entropy positions: a low-entropy token resamples to itself with probability near one, so "the answer did not change" is trivially true there, and the design measures token *variability* rather than leverage. Forcing a counterfactual (v3, v4) removes that and introduces its mirror image: where the reference token has probability 1.000, any substitute is a $\sim$23-nat surprise, while at a gated position ($P(\text{ref}) = 0.27$) the alternative is a natural continuation. Measured directly, the forced substitution injects 22.99 nats at low-entropy positions against 0.38 at high-entropy ones.

Controlling for it is necessary, and barely possible. Regressing the answer-flip indicator on injected surprisal and group gives a group effect of -0.077 over all rows and -0.101 restricted to the overlapping surprisal window. But the overlap is thin: the gated group's surprisal concentrates at 2.30 nats ($p_{90} = 4.30$) while the ungated group's begins at 15.19 (p_{10}), so only 423 of 4926 gated interventions (8.6%) fall in the shared range. The adjusted estimate describes an atypical slice of the treatment group, and we do not use it to support an arm ordering or to falsify localization.

Why this is close to ill-posed. For the observed discrete forced-token substitutions, a near-deterministic position offered no mild perturbation. Any edit there is far off-distribution by construction. So high- and low-entropy positions cannot be compared at matched perturbation strength over most of the range, and the question "does this position matter under an intervention?" has no perturbation-independent answer. Anyone running such a comparison should report the injected-surprisal distributions of both groups and the fraction of common support; without those numbers a positive or negative effect cannot be interpreted.

P.4 STATISTICAL UNIT

We preregistered the comparison point, the primary and secondary contrasts, the multiplicity correction, and a mapping from outcomes to conclusions. We did not preregister the *unit of analysis*, and the two defensible choices disagree.

Pooling every (problem, seed) pair and running McNemar over the pooled vector gives our primary contrast as $+1.60$ points, $p = 0.043$. Treating each seed-matched run pair as one observation — four seeds contributing $+0.40$, $+4.60$, $+1.00$, $+0.40$ — gives $t = 1.58$, $p = 0.211$ two-sided. The pooled test multiplies n by the number of seeds and treats runs as exchangeable, discarding exactly the run-to-run variance a claim about a *method* is about.

We report the historically planned pooled statistic as supplementary and state that it does not support a method-level claim. There are three reasons to disclose both analyses rather than select one post hoc. The choice was not made in advance, so making it after seeing results is the degree of freedom preregistration exists to remove. The disagreement is not marginal — $p = 0.043$ against $p = 0.211$. This example shows that pooling can overstate precision.

A second degree of freedom we declined. The primary contrast is larger at the final evaluation after 120 updates ($+2.40$ points, seed-level $t = 2.98$, two-sided $p = 0.058$) than at the preregistered historical label step 80 (81 updates: $t = 1.58$, $p = 0.211$). Neither is significant at conventional two-sided $\alpha = 0.05$. We did not move the primary point, and we have no mechanism to explain the checkpoint divergence.

Recommendation. Preregistrations for arm comparisons should fix the unit explicitly, and seed-level statistics should be the primary result. Two-sided tests should be the default: our own earlier working notes used one-sided critical values and labeled results as significant even though the corresponding two-sided tests were not.

P.5 FORMAT PORTABILITY

The retained portability analysis scores the same DeepSeek-family rollouts with a Qwen-tuned boxed rule and a format-agnostic parser. Historical parser and EOS identity remain incomplete, so the comparison is a motivating audit rather than headline empirical support.

Table 14: Documented same-rollout format audit. A family-specific boxed rule rejects most answers and changes the apparent self-consistency signal.

scoring rule	pseudo-label acc.	majority size	abstentions / group
with `\boxed` requirement	0.219	0.98 / 8	6.47
without it	0.275	3.50 / 8	0.00

The appropriate conclusion is that scoring conventions require family-specific validation, not that one family is intrinsically incapable of consensus.

P.6 HISTORICAL CASE STUDY AND THE PROSPECTIVE E00 ANCHOR

We report the experiment in which the failure modes co-occurred. Four mask-defined arms (§3) share the pseudo-label rule and hyperparameters. They train Qwen2.5-Math-7B for 120 updates on DeepScaleR Level 2 ($G = 8$), with 500 matched evaluation problems. Consensus uses four seeds; the gold control has four seeds for A/B/D and two for C. Equal seed identifiers do not imply equal on-policy rollouts or pseudo-labels after policies diverge.

The seed/run analysis is the method-level result and is unresolved. The later DECISION–RANDOM comparison is also not conventionally significant. The historical primary table, full checkpoints, gold control, and contrasts are in App. I; the power analysis is in Table 8. Together they show that the historical study is a diagnostic case, not a basis for ranking arms. The exploratory gating-fraction sweep remains supplementary (App. Q).

Preregistered normalization-matched test. The confirmatory design fixes the global policy-gradient denominator and uses the seed-matched run pair as its independent unit. All paired runs passed outcome-blind technical admission, and the result below is generated only from the resulting hash-validated raw artifacts. Table 15 reports the registered endpoints, and Figure 3 shows the corresponding checkpoint trajectory.

Table 15: Preregistered normalization-matched Decision–Random result. The independent unit is the seed-matched run pair; positive values favor Decision. Final-checkpoint inference belongs to the declared six-test Holm family.

completed updates	Decision	Random	Δ (pp)	95% CI	two-sided inference
81 (primary)	0.5757	0.5545	+2.12	[+1.22, +3.02]	$p = 0.0002$
120 (secondary)	0.5872	0.5563	+3.09	[+2.18, +4.01]	Holm $p = 2.67 \times 10^{-5}$

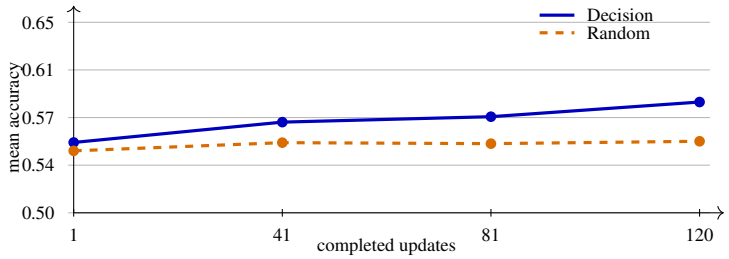


Figure 3: Normalization-matched Decision and Random trajectories from 13 paired training seeds. Updates 1 and 41 are descriptive; update 81 is the preregistered primary, and update 120 belongs to the Holm-adjusted secondary family. The lines show seed means, not additional independent observations; Table 15 reports the paired-run uncertainty.

Confirmatory result. At the preregistered primary evaluation after 81 completed updates, the seed-matched run-pair estimate for Decision minus Random is +2.12 percentage points (95% CI [+1.22, +3.02], two-sided $p = 0.0002$). The paired interval lies above zero. This supports the Decision selector for this model, dataset, and training budget; it does not establish a model-family-wide advantage.

Q Exploratory gating-fraction sweep

Table 16: Observation, not a claim. Two seeds per gated fraction; no individual seed-level comparison is significant. GLOBAL uses four seeds, so this is not a balanced arm expansion.

label	fraction	DECISION acc	GLOBAL acc	GLOBAL−DECISION (pp)	seed p
80 (81 updates)	1%	0.5540	0.5860	+3.20	0.193
	5%	0.5745	0.5910	+1.65	0.080
	10%	0.5950	0.5860	−0.90	0.323
final (120)	1%	0.5650	0.6080	+4.30	0.074
	5%	0.5750	0.6050	+3.00	0.049
	10%	0.6040	0.6080	+0.40	0.795

R Detailed limitations

- The underpowered case study estimates 13 seeds for its primary contrast and 88 for the perfect-signal contrast under retrospective normal-approximation planning, but ran four.
- E00 uses Qwen2.5-Math-7B, DeepScaleR Level 2, one headline gating fraction and 120 updates. E10 and E01 change one axis at a time; they do not cover the joint shift, other

model scales, tasks, selectors, or budgets. E01 replicates the ordering while E10 reverses it, so the prospective evidence is explicitly setting-dependent. The two-seed fraction sweep remains exploratory, and the historical step-80/final divergence has no mechanism.

- Recovered metrics and per-problem rows reproduce the arm statistics, but historical source snapshots, training rollouts and checkpoints remain absent. Current corrected EOS and resume semantics therefore cannot be projected backward.
- The clipping confound was found during diagnosis rather than by an exhaustive audit; other optimizer or evaluation interactions may remain.
- In the extension replication, a terminal-debugging command inadvertently displayed intermediate scientific fields after the seed sets, endpoints, analyzer, and continuation rules were frozen. No displayed value changed scheduling, inclusion, retry, or analysis, but outcome blindness was not restored. We retain the full registered sample and report the extension as an analysis-frozen replication with a disclosed process deviation.
- Paired arms were not order-randomized: the Random arm was released only after its same-seed Decision arm became terminal. Actual hardware, runtime, retry, and generated-token accounting therefore must be checked in the anonymous artifact bundle before interpreting the nominally matched training budget as matched realized compute.

S EVALUATION CHECKLIST

The seven checks are cheap to apply once named. For anyone comparing a token-selection method against a baseline in LLM RL:

1. **Log pre-clip gradient norms and clipping incidence per arm.** If arms differ in the fraction of steps that hit the clip, they are not running the same optimizer and any difference between them is confounded with pre-clip gradient scale and the post-clipping gradient-norm budget. Where a gated and a global arm must be compared, run the gated arm additionally with the global normalizer.
2. **Preregister the unit of analysis, and default to an independent training run or prespecified matched run pair.** Pooling evaluation items across seeds inflates n by the seed count and discards the variance that a method-level claim is about. Report both if they disagree and interpret the prespecified analysis as primary. Use two-sided tests.
3. **In perturbation studies, report the injected-surprisal distribution of each group and the fraction of common support.** Natural resampling and forced counterfactuals are biased in opposite directions; without surprisal control the sign of the effect is a design choice. Where common support is thin, say so rather than reporting an adjusted estimate as if it generalized.
4. **Make abstention semantics executable and model-specific.** Configure parse, answer format and EOS requirements independently; fail if an enabled EOS rule lacks a tokenizer EOS identifier. Report each abstention reason per model family before reporting accuracy.
5. **Run the power analysis before the experiment, not after.** Retrospective planning showed that the historical primary contrast needed 13 seeds, but that case study ran four; the prospective comparison then used the declared paired-run count. Publishing the power basis distinguishes "no effect" from "no measurement".
6. **Archive a claim-to-artifact bundle.** For every headline number retain resolved configuration and code, dataset and model identifiers, per-problem evaluation, metrics, logs, Python exit status and checkpoint metadata. Scheduler "COMPLETED" alone is not Python success, and a consolidated JSON is not a raw run.
7. **Name evaluations by completed updates.** If the loop updates before evaluating, a zero-based log label is offset by one. Persist both the historical label and an explicit `updates_completed` field so checkpoint comparisons cannot drift semantically.