WHICH UPDATES ERASE WHAT THE BASE MODEL KNEW? OPTIMIZER-CONDITIONED ATTRIBUTION OF REASONING-SUPPORT LOSS IN RLVR

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement learning with verifiable rewards (RLVR) can improve short-horizon accuracy while erasing solution routes already supported by the base model. Entropy decline reveals this contraction but does not identify which updates cause it. We introduce CAGE (Cross-context Attribution of Gradient Effects), an optimizer-conditioned cross-context attribution score that estimates how each current update changes entropy at held-out correct states and selectively regularizes the predicted high-risk tail. Accumulated CAGE exposure predicts subsequent support loss ($\hat{\beta}_{25} = -.113$) and remains predictive when recent entropy decline is included on the same rows. CAGE improves macro Pass@1 by +2.06 pp over OPEFO. Under the construction-matched discrete intervention, CAGE-D exceeds matched random targeting by +2.59 pp in macro Pass@1 and +.037 in correct-family mass, and exceeds OPEFO by +0.93 pp in macro Pass@1; uniform KL, source-strength matching, and mean-probe controls recover only parts of the gain. Selective KL recovers most of the high-k loss, with CAGE-D above matched random selection in point estimate. It costs $1.055\times$ GRPO wall-clock and yields positive transfer point estimates on Qwen-7B, Llama-8B, and code, with weaker accuracy evidence under the larger shifts. Together, these results support targeting update-level risk rather than treating entropy preservation alone as sufficient for intervention design.

1 INTRODUCTION

RLVR can make a reasoning model better at finding a correct answer quickly while making some correct routes harder to reach at all. On Qwen2.5-3B, within the same nested 512-generation pool per item, GRPO raises Pass@64 from 80.2 to 83.2, yet at $k = 512$ the ordering reverses: the frozen base model reaches 91.8, versus 89.5 after GRPO. Large-k evaluations report the same broad phenomenon in reasoning RL (Yue et al., 2025). Which training updates erased what the base model already knew?

Entropy is a symptom, not an attribution. Entropy decline reveals distributional contraction, but it does not identify which current update damages which held-out correct state. Existing attribution tools do not directly answer this question either. Influence functions and TracIn estimate how past training examples affect predictions or losses (Koh & Liang, 2017; Pruthi et al., 2020), while forward-only last-layer methods such as For-Value exploit hidden-state and prediction-error structure for data valuation (Deng et al., 2026). RLVR requires a different estimand: the unit being removed is a sampled token inside the current policy-gradient step, the realized update passes through the optimizer rather than raw Euclidean gradient geometry, and the quantity at risk is held-out reasoning support rather than current training loss, making attribution update- and context-specific within RLVR.

CAGE asks a local counterfactual question: if a current token contribution were removed from the optimizer update, how would uncertainty change at held-out states along correct base-model solutions? We call the resulting quantity CAGE (Cross-context Attribution of Gradient Effects), an optimizer-conditioned cross-context estimator. Its direct LM-head term is forward-computable, making the score practical enough to use online. CAGE uses the score continuously; CAGE-D isolates its ranking by targeting the top 5% positive-score tokens. In the primary study, CAGE reaches Pass@512 of 92.7, within the base-model range (base 91.8; GRPO 89.5), with selective KL accounting for most of the recovery (matched random 91.5) under the same token budget.

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.

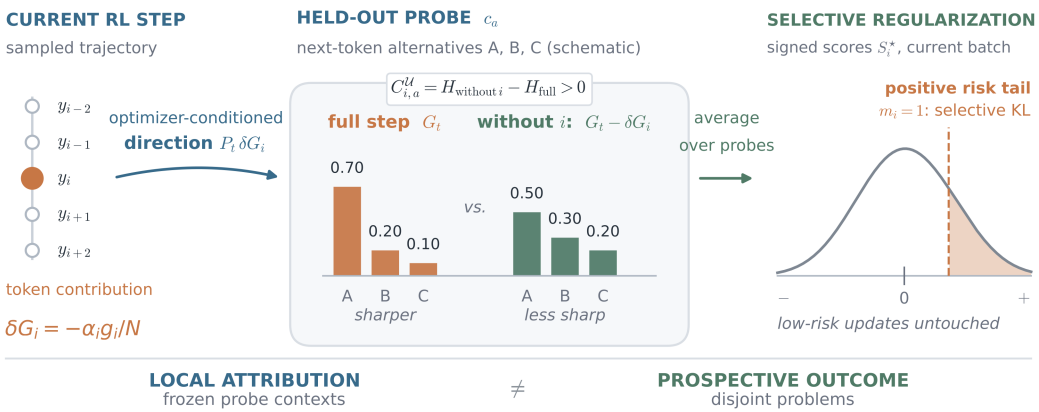


Figure 1: CAGE method schematic. Removing a sampled token contribution δG_i from the optimizer-conditioned step changes the next-token distribution on frozen held-out probe contexts; averaging this local effect gives the signed attribution score $S_i^\star$, whose positive risk tail receives selective KL. Prospective outcome problems are disjoint from the frozen probes and are used solely to test whether accumulated attribution exposure predicts later support loss; they never enter score construction.

Our contributions are: (i) **Attribution**: an optimizer-conditioned cross-context estimand with a forward-computable LM-head closure to rank individual updates by effects on held-out reasoning support; (ii) **Prediction**: prospective evidence that accumulated exposure predicts later support loss beyond entropy and simpler comparators; and (iii) **Intervention**: published-baseline comparisons, construction-matched controls, high-k recovery, and transfer tests of whether the ranking is actionable.

2 Related Work

Entropy contraction in RLVR. GRPO and large-scale RLVR made reinforcement learning practical for mathematical reasoning (Shao et al., 2024; Guo et al., 2025; Yu et al., 2025; Hu et al., 2025), while entropy dynamics during reasoning RL have been studied directly (Jin et al., 2026). Large-k evaluations show that RLVR can improve low-sample accuracy while narrowing exploration relative to the base model (Yue et al., 2025). Methods respond to this narrowing by controlling exploration or entropy-related quantities during training: Clip-Higher relaxes the upper clipping boundary (Yu et al., 2025); Forking-20 retains policy-gradient updates only on the highest-entropy 20% of tokens (Wang et al., 2025); Clip-Cov and KL-Cov act through the covariance mechanism (Cui et al., 2025); STEER reweights by entropy change (Hao et al., 2026); and OPEFO balances on-policy entropy flow (Xu et al., 2026). We evaluate all six under the shared primary protocol. These methods use batch-local or training-context signals such as entropy, clipping geometry, covariance, or entropy flow; CAGE instead ranks tokens by their optimizer-conditioned effect on held-out contexts. Meng et al. (2026) show that RLVR induces sparse token-level distributional shifts and use cross-sampling interventions to establish that a small subset of token decisions can be functionally critical. CAGE addresses a complementary prospective question: rather than retrospectively identifying consequential output shifts, it attributes how individual current updates affect base-supported behavior on held-out reasoning states, enabling targeted regularization before those updates are applied to the policy.

Cross-example learning dynamics and attribution. Influence functions relate predictions to training points via gradients and inverse curvature (Koh & Liang, 2017), with scalable/LLM variants (Schioppa et al., 2022; Grosse et al., 2023); TracIn, datamodels, and TRAK provide local, trajectory, or predictive views (Pruthi et al., 2020; Ilyas et al., 2022; Park et al., 2023). Gradient-inner-product similarity gives a related interaction view (Charpiat et al., 2019), while representer points and For-Value exploit output-layer structure for cheaper valuation (Yeh et al., 2018; Deng et al., 2026). Most directly, Ren & Sutherland (2025) decompose how a finetuning example changes another prediction through an empirical-NTK interaction. At the LM head under Euclidean geometry, their cross-context kernel reduces to hidden-state alignment; Eq. (6) is our entropy-target, sampled-token instance. CAGE instead targets held-out correct-state entropy under realized optimizer state and uses the token counterfactual online for selective regularization without probe gradients during training.

Table 1: Relationship to adjacent attribution and entropy-dynamics formulations. "Extra item backward" means item-specific backward beyond training; CAGE needs probe forwards, not probe backward. Columns distinguish attribution unit, target, geometry, context, and learning role; they compare estimands rather than rank methods. $D\mathcal{U}_t[G_t]$: optimizer Jacobian (Section 3.2).

Method	Unit	Target	Geometry	Context	Use / extra item backward
Influence / TracIn	example	loss / prediction	curvature / SGD	cross-example	analysis / yes
LESS	example	validation relevance	Adam-aware	cross-example	data selection / yes
Ren & Sutherland	example/response	prediction dynamics	Euclidean eNTK	cross-context	dynamics / –
Cui et al.	update/token	entropy dynamics	PG/NPG covariance	same-context	intervention / no
CAGE	step token	held-out entropy	$D\mathcal{U}_t[G_t]$	cross-context	online / no probe backward

Optimizer-aware influence. LESS is an optimizer-aware method for targeted instruction-data selection and constructs Adam-aware gradient features for example-level influence (Xia et al., 2024). CAGE asks a different question. If $\mathcal{U}_t$ is nonlinear, the marginal effect of removing a contribution from the current aggregate gradient is locally $D\mathcal{U}_t[G_t]\,\delta G_i$; it is not in general obtained by applying an optimizer transformation to the isolated contribution alone. We therefore use the Jacobian at the current aggregate gradient. This is an estimand distinction, not an empirical claim that CAGE dominates LESS: LESS operates at example level before training-data selection, whereas CAGE operates at token level inside the current RL step. Our finite-LOO audit establishes only improved fidelity over the Euclidean CAGE score, not a comparison of these two optimizer-aware estimands.

Entropy covariance as a limiting connection. Cui et al. (2025) show that first-order entropy change can be written through a covariance between log probability and a logit perturbation, motivating Clip-Cov and KL-Cov. CAGE contains the same covariance structure in a narrower limit: when source and probe contexts coincide and the geometry is Euclidean in logit space, $\phi^\top r = \mathrm{Cov}_{v\sim\pi}(\log \pi_v, r_v)$, so the CAGE score reduces to a signed advantage-weighted same-context covariance (the same-context covariance proposition in Section 3.3). The deployed method instead evaluates cross-context interactions at held-out correct states under optimizer geometry. We do not infer from the present experiments that cross-context probing is intrinsically superior to same-context covariance; the relation locates the estimands and clarifies what the existing controls do and do not identify.

Optimizer geometry. Adaptive and structured optimizers transform gradients into parameter updates, from coordinate-wise Adam (Kingma & Ba, 2015) to block-structured K-FAC and Shampoo (Martens & Grosse, 2015; Gupta et al., 2018). A Euclidean gradient inner product therefore may not reflect the realized AdamW geometry. CAGE defines its counterfactual through the Jacobian of the optimizer map; AdamW is a diagonal case of the vocabulary-row-block geometry in Section 3.3.

3 CAGE: Optimizer-Conditioned Cross-Context Influence

CAGE proceeds in three conceptual steps. First, it defines the finite counterfactual effect of a current token contribution on entropy at a held-out correct state. Second, it linearizes this effect through the actual optimizer map, so attribution follows the geometry of the realized update rather than the raw gradient alone. Third, it exploits LM-head structure to compute the resulting cross-context score online without probe-specific backward passes. The final probe-population score aggregates these interactions into the risk signal used by CAGE and CAGE-D during online selection. Eqs. (1)–(2) define the counterfactual target, Eqs. (3)–(4) its optimizer-aware linearization, Proposition 1 its forward-computable online form, and Section 3.4 the probe-population score used for selection.

3.1 Finite counterfactual estimand

Let c_i denote the prefix at sampled training token y_i. A held-out probe context c_a is a frozen prefix of a correct base-model solution to a held-out problem, taken at one of four normalized reasoning positions (0, .25, .50, or .75 of the solution trace); probes are never outcome problems in the prospective panel.

$$\begin{aligned} g_i &= \nabla_\theta \log \pi_\theta(y_i \mid c_i) \qquad \delta G_i = -\frac{\alpha_i}{N} g_i, \\ \theta_t^+(G) &= \theta_t - \eta_t \mathcal{U}_t(G) \qquad H_a(\theta) = -\sum_v \pi_\theta(v \mid c_a) \log \pi_\theta(v \mid c_a). \end{aligned} \tag{1}$$

where α_i is the signed GRPO advantage weight multiplying token i's log-policy gradient; because G_t is the gradient of the minimized loss, the corresponding contribution is $-\alpha_i g_i/N$, N is the number of valid response tokens, and $\mathcal{U}_t$ is the conditional update map with inherited optimizer state, schedule, reference policy, weight-decay state, sampled batch, and selector mask m held fixed. If $G_t(m)$ is the complete current-step gradient, our realized-batch conditional estimand for token i is

$$\boxed{C^{\mathcal{U}}_{i,a}(m) = H_a\big(\theta_t^+(G_t(m) - \delta G_i)\big) - H_a\big(\theta_t^+(G_t(m))\big)\,.} \tag{2}$$

Removing i does not recompute m: otherwise the selector would become part of the definition of the quantity that defines the selector. Positive $C^{\mathcal{U}}_{i,a}$ means that including i sharpened the held-out state; we therefore orient positive first-order scores as harmful entropy-reducing exposure. This direction is checked independently against exact finite leave-one-out updates: at the tied 3B audit checkpoint, the deployed optimizer-aware score has .914 sign agreement with the finite counterfactual. Full notation and sign conventions used throughout the main analysis are collected in Appendix A.2.

3.2 Optimizer-conditioned approximation

Optimizer state changes the attribution geometry. The optimizer map, not the raw gradient alone, determines the realized update. For $\mathcal{U}_t$ Fréchet differentiable at $G_t(m)$, define $P_t(m) = D\mathcal{U}_t[G_t(m)]$.

Optimizer-aware first-order score. With the selector and inherited optimizer state held fixed,

$$\boxed{C^{\mathcal{U}}_{i,a}(m) = -\frac{\eta_t \alpha_i}{N}\,\langle \nabla_\theta H_a(\theta_t), P_t(m) g_i\rangle + R^{\mathcal{U}}_{i,a}.} \tag{3}$$

Thus the ranking score is

$$\boxed{s^{\mathcal{U}}_{i,a}(m) = -\alpha_i\,\langle \nabla_\theta H_a, P_t(m) g_i\rangle\,.} \tag{4}$$

The Euclidean special case has $P_t = I$. For AdamW, the induced bilinear form is generally sign-indefinite, so we treat it as an optimizer-conditioned interaction kernel rather than a similarity metric. The same source token can therefore receive a different attribution under different optimizer states. The approximation is first order in the leave-one perturbation; the remainder bound and AdamW differential are in Appendices A.3 and A.6. The remaining challenge is computational: evaluating this interaction naively for every source-token/probe pair would preclude use in training.

3.3 Forward-computable head closure

At the LM head, the expensive interaction collapses. Both source-token and probe gradients are rank one; under a vocabulary-row-block optimizer operator, output rows remain separated, leaving only forward-computable row-wise scalar products. For vocabulary head $z = Wh$, define

$$\phi_{a,v} := -\frac{\partial H_a}{\partial z_{a,v}} = \pi_{a,v}(\log \pi_{a,v} + H_a), \qquad r_{i,v} := \frac{\partial \log \pi(y_i \mid c_i)}{\partial z_{i,v}} = \mathbf{1}[v = y_i] - \pi_{i,v}. \tag{5}$$

In the Euclidean case, the rank-one direct-head structure gives the forward-only factorization

$$s^H_{i,a} = +\alpha_i(\phi_a^\top r_i)(h_a^\top h_i), \tag{6}$$

so no per-instance backward pass is required: rollout/policy forwards provide π_i, h_i, and a batched probe forward provides π_a, h_a. This Euclidean case exposes the factorization; CAGE preserves it while replacing Euclidean geometry with the optimizer-induced operator for online scoring.

Conceptually, the head-level pairwise score combines a signed RL update weight, the sampled-token residual, a held-out probe entropy factor, and an optimizer-conditioned hidden-state kernel.

The following result states when this row-wise computation is exact.

Proposition 1 (forward closure under vocabulary-row-block geometry). *Let the output embedding be untied and the local head operator be block diagonal across vocabulary rows, $P_H = \mathrm{blockdiag}(D_1, \ldots, D_{|\mathcal{V}|})$, with $D_v \in \mathbb{R}^{d\times d}$. Then the exact direct-head score is*

$$\boxed{s^{H,P}_{i,a} = +\alpha_i \sum_v \phi_{a,v} r_{i,v}\, h_a^\top D_v h_i.} \tag{7}$$

Coordinate-wise Adam-type preconditioners correspond to diagonal D_v, giving the directly computed kernel $\kappa_v^P(a,i) = h_a^\top D_v h_i$. The closure is exact for an untied output path under row-block geometry. In the primary model the input and output embeddings share parameters, so those parameters also receive an input-side gradient absent from the isolated output-head derivation. The primary tied model therefore uses this closure as a direct-head approximation rather than an exact identity; Section 7.5 empirically audits its finite-counterfactual fidelity. Appendix A.5 quantifies the shared-parameter boundary; Appendix A.4 proves the result and quantifies row mixing beyond the direct head.

Proposition 2 (same-context covariance limit). Let the probe equal the source context, use Euclidean logit-space geometry, and set $h_a^\top h_i = 1$ (logit parameterization). With $\phi_v = \pi_v(\log \pi_v + H)$ and any logit direction r, the entropy interaction is exactly a covariance:

$$\phi^\top r = \sum_v \pi_v(\log \pi_v + H) r_v = \mathrm{Cov}_{v\sim\pi}(\log \pi_v, r_v). \tag{8}$$

Hence Eq. (6) reduces to $s_i = +\alpha_i \,\mathrm{Cov}_{v\sim\pi}(\log \pi_v, r_{i,v})$ in this limit. This identifies the covariance structure underlying prior same-context entropy-dynamics analyses as a limiting case of the interaction used here; it does not make Clip-Cov or KL-Cov algebraically identical to CAGE, whose deployed estimand combines held-out contexts with optimizer geometry for online token ranking.

3.4 Deployed CAGE score and online procedure

CAGE aggregates these optimizer-conditioned interactions into a source-token risk score. The held-out target is the frozen-population expectation, not a fixed anchor score.

$$\boxed{S_i^\star = \mathbb{E}_{(c,\zeta)\sim Q_{\mathrm{probe}}}[s_{i,(c,\zeta)}].} \tag{9}$$

Here Q_{probe} is the fixed stratified distribution over held-out probe contexts, and ζ records the frozen reasoning-position stratum attached to context c. Online, we estimate it with $K = 16$ stratified probe samples and contract only the $M_{\mathrm{vocab}} = 128$ largest-magnitude entropy-side vocabulary entries plus the sampled token. The support is fixed before scoring. Q_{probe} defines the behavior being protected, while the frozen stratified $K = 16$ approximation tracks the $K = 64$ reference ($\rho = .956$, top-5% Jaccard .827; Table 10). The support rule, retained-mass/ranking validation, Monte-Carlo convergence, and Bernstein concentration bound are in Appendices A.16–A.17. Under the common wall-clock accounting this deployed scoring path costs 1.055× GRPO (Table 30); finite-counterfactual fidelity under the same scoring rule used online is validated in Table 13.

Online procedure. Scoring requires only frozen-probe forward passes, with no attribution-specific probe backward pass; the complete eight-step implementation sequence is given in Appendix A.14.

4 Linking CAGE to Reasoning-Support Loss

Attribution is useful only if predicted damage corresponds to reasoning-support loss. We define base-referenced retention and test whether CAGE exposure prospectively predicts its decline.

4.1 Correct-family retention

Entropy reduction need not be harmful: a policy may sharpen around a better solution. R^{mass} measures the fraction of probability mass on correct solution families supported by the base model that survives training; newly discovered correct solutions do not increase the score. Family construction, independent discovery/reference/evaluation streams, and the conditional companion R^{cond} are defined in Appendix A.20. Accordingly, the primary endpoint isolates retained base-supported mass:

$$\boxed{R_x^{\mathrm{mass}}(t) = \frac{\sum_{j\in\mathcal{F}_x^{\mathrm{base}}} \min\{\hat{p}_{x,j}^{\mathrm{mass}}(t), p_{x,j}^{\mathrm{mass,ref}}(0)\}}{\sum_{j\in\mathcal{F}_x^{\mathrm{base}}} p_{x,j}^{\mathrm{mass,ref}}(0)}.} \tag{10}$$

Unless marked otherwise, R below denotes R^{mass} throughout the main text and tables.

Because R^{mass} is referenced to base-supported mass, reference-KL methods can be structurally advantaged over unconstrained GRPO on this metric; this is why the construction-matched retention contrast is CAGE-D versus random matched KL, which shares the same selective-KL construction.

4.2 EXPOSURE AND PROSPECTIVE PANEL

Prospective prediction links CAGE exposure to subsequent support loss. To avoid using the same problem as probe and outcome, each exposure group g has disjoint probe and outcome sets P_g and O_g. Groups are topic crossed with a two-bin (early/late) frozen reasoning position ($G = 10$); difficulty, base accuracy, response length, and advantage variance enter as controls Z. Let $a_{g,u}$ be the token-normalized positive CAGE exposure at optimizer step u, averaged over the frozen probes in P_g; the primary predictor is its mean over the preceding 25 optimizer steps. For outcome problem $x \in O_g$, define $Y_{x,t,\Delta} = R_x(t + \Delta) - R_x(t)$. All regressors use only steps $\leq t$; a future-exposure lead placebo is null ($-.001$; Table 22). The temporally ordered prospective panel is

$$\boxed{Y_{x,t,\Delta,s} = \beta_\Delta A^{(25)}_{g(x),t,s} + \gamma^\top Z_{x,t,s} + \mu_x + \lambda_{t,s} + \epsilon_{x,t,s}.} \tag{11}$$

The primary horizon is $\Delta = 25$ and the secondary horizon is $\Delta = 50$; a 10-step estimate is confined to dense diagnostic windows. On identical rows, M_1 uses preceding-window entropy decline only, M_2 uses CAGE exposure only, and M_3 includes both, testing whether CAGE contributes incremental prospective information beyond entropy. Inference uses problem and seed-specific time fixed effects and a wild cluster bootstrap over seed×group trajectories (Stock & Watson, 2008; Cameron et al., 2008); full construction and robustness checks for these specifications are in Appendix A.22.

5 INTERVENTIONS

Prediction alone does not establish that CAGE identifies useful intervention targets. The discrete intervention tests the ranking; the continuous intervention additionally uses score magnitude. We therefore use a common fixed-reference selective-KL construction for CAGE causal controls:

$$\mathcal{L} = \mathcal{L}_{\mathrm{RL}} + \lambda \frac{\sum_i m_i D_{\mathrm{KL}}(\pi_i \| \pi_{\mathrm{ref},i})}{\sum_i m_i + \varepsilon}. \tag{12}$$

The discrete targeted and matched controls use the same eligible-token budget and realized selective-KL target. CAGE-D/CAGE denote discrete/continuous targeting. Appendix Table 32 gives the canonical reading logic; exact constructions and secondary factorial diagnostics are in Appendix A.29.

Random (ranking), source-strength matched (update magnitude), anti-targeted (direction), mean-probe (individual probe information), and $|s|$ (sign) controls test the corresponding alternative explanations under matched constructions where applicable to each contrast (Appendix A.29).

The discrete arm regularizes the top 5% positive-score eligible tokens (realized primary fraction 5.04%). The continuous arm preserves relative positive CAGE magnitudes via clipped EMA normalization and matches mean KL weight to the discrete arm over the same window; weighting, source-strength matching, and normalization for both arms are detailed in Appendix A.27.

6 EXPERIMENTAL DESIGN

Reading rules. We label evidence as primary, secondary, or descriptive per the claim map (Appendix A.1). Published baselines retain source-method semantics; they are not selector-only controls. Selector-specific conclusions use construction-matched controls with the same selective-KL construction and budget. The 30-item AIME sets are reported for completeness, not as standalone support for headline effects. These rules apply throughout unless a local caveat states otherwise.

Table 2 fixes each main block's model, evidential role, and paired training-seed count where defined; diagnostic audits use Qwen2.5-Math-1.5B (Yang et al., 2024) and are reported separately in the supplement as mechanism checks rather than primary evidence for headline intervention effects.

Math RL uses DAPO-Math-17K (Yu et al., 2025); code RL uses TACO (Li et al., 2023), with HumanEval+/MBPP+ (Liu et al., 2023) as external tests. Primary 3B comparison uses 500 AdamW steps, group 8, and prompt batch 64. Evaluation uses MATH-500 (Hendrycks et al., 2021; Lightman et al., 2024), AMC23, AIME24/25, Pass@k, and MATH500/AMC family retention; the 30-item AIME sets are too small for reliable retention (75–82% still have at most one discovered family at $M = 128$; Appendix Table 17). Probe/evaluation problems are decontaminated from training data; Appendix A.32 reports paired-seed operating characteristics and uncertainty accounting.

Table 2: Claim-to-experiment map. Seed counts refer to paired training seeds for intervention/transfer blocks and GRPO training seeds for the prospective panel. The untied companion combines fidelity and intervention audits with distinct sampling units, so no single seed count is asserted here.

Setting	Model	Evidential role	Seeds	Main endpoint
Primary	Qwen2.5-3B (Qwen et al., 2024)	Efficacy + identification	10	Accuracy, retention, Pass@k
Prospective	Qwen2.5-3B	Prediction	5	Future retention
Untied companion	Qwen2.5-3B	Closure validation	–	Fidelity + intervention
Scale transfer	Qwen2.5-Math-7B (Yang et al., 2024)	Transfer	10	Accuracy + retention
Family transfer	Llama-3.1-8B-Instruct (Grattafiori et al., 2024)	Transfer	5	Accuracy + retention
Domain transfer	Qwen2.5-Coder-7B (Hui et al., 2024)	Transfer	5	Code accuracy + retention

Because published gains are sensitive to model scale, decoding, update count, and metric, we do not normalize the shared-protocol results by each paper's reported gain. Appendix A.26 instead documents implementation provenance, source hyperparameters, protocol differences, and absolute shared-protocol outcomes; the strongest published-method baselines improve materially over GRPO rather than serving as weak reference points. Family-resolution diagnostics are reported in Appendix A.20, and hyperparameters and evaluation rules in Appendices A.30 and A.32.

For architecture validation, the 3B untied-from-initialization companion copies the tied input/output embedding matrix into separate input and output parameters with identical initial values before RL, after which the parameters are trained independently under the same recipe. This creates a closure-exact output-head path without changing the starting numerical embedding. Qwen-7B tests scale, Llama cross-family math transfer, and Qwen-Coder executable-reward transfer with deterministic program-strategy signatures. Retention is the primary transfer endpoint and accuracy is secondary. Transfer protocols, family audits, and inference rules for transfer are in Appendices A.24–A.25.

7 RESULTS

We test this evidence chain: prediction → intervention → ranking specificity → transfer → approximation fidelity, with each stage testing a distinct claim under its corresponding control. Macro Pass@1 is the unweighted mean over benchmarks within each seed, then averaged across seeds.

7.1 CAGE EXPOSURE PREDICTS FUTURE SUPPORT LOSS BEYOND ENTROPY

CAGE exposure predicts later loss of base-supported correct-family mass after problem/time controls. At the primary 25-step horizon, $\hat{\beta}_{25} = -.113$. With recent entropy decline on the same rows, the CAGE coefficient remains -.110, while the entropy coefficient attenuates to -.005 (Figure 2a; Table 3).

The predictive signal is not reproduced by simpler attribution geometry. On identical rows, CAGE separates more strongly than the Euclidean and family-balanced For-Value comparators, with paired coefficient differences of -.060 and -.082, respectively (Table 3); both are more negative for CAGE.

The association is stable across seeds: all five seed-specific coefficients are negative with varying magnitudes, and lead and group-shuffle placebo intervals include zero. It also persists at the secondary 50-step horizon and under leave-one-seed-out, correctness-conditional, tail-enrichment, and code-domain checks. Details and robustness results are in Appendices A.22, A.31, and A.25.

Table 3: Prospective evidence. Primary 25-step specifications use identical rows; coefficients give retention change per predictor SD and R^2 is incremental over fixed effects and controls. The final row gives the secondary 50-step horizon; the comparator panel is Appendix Table 14.

Specification	Focal predictor coeff.	95% CI	R^2
Entropy only	-.057	$[-.073, -.041]$	0.012
CAGE exposure only	-.113	$[-.128, -.096]$	0.047
CAGE + entropy	-.110	$[-.124, -.095]$	0.047
CAGE − Euclidean (paired Δ)	-.060	$[-.074, -.046]$	+.026
CAGE − family-balanced For-Value (paired Δ)	-.082	$[-.099, -.064]$	+.032
CAGE exposure ($\Delta = 50$)	$-.102$	$[-.131, -.074]$	0.073

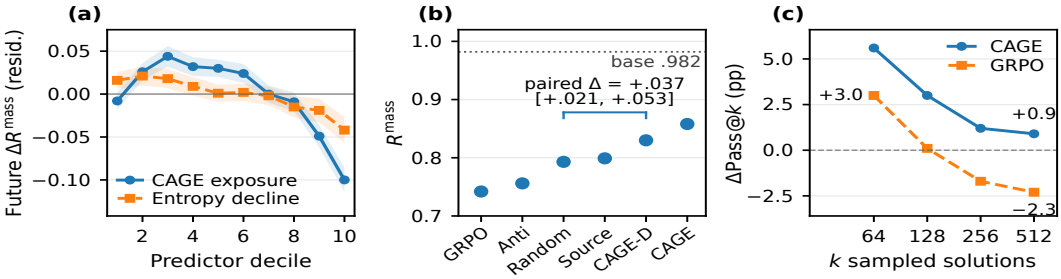


Figure 2: From prediction to intervention. **(a) Prospective signal beyond entropy:** future retention change by decile of CAGE exposure and entropy decline (95% bands). **(b) Selector-specific intervention:** correct-family retention under construction-matched targeting; dotted line is the base reference, bracket the paired CAGE-D–random contrast. **(c) Early gain, long-tail reversal:** Pass@k relative to base; at $k = 512$, CAGE is +0.9 pp, but its CAGE–base interval includes zero.

Table 4: Primary 3B results (Pass@1, ten-seed mean (SD)). Published baselines shown are the strongest on accuracy (OPEFO) and retention (Clip-Cov); random is the construction-matched control for CAGE-D. Full baseline comparison is in Appendix Table 24; paired contrasts are in Table 5.

Method	MATH500	AMC23	AIME24	AIME25	Avg.	R^{mass}	Cost
GRPO	65.00 (1.28)	53.00 (2.30)	6.67 (2.72)	9.33 (3.06)	33.50 (1.48)	.742	1.000×
Uniform KL	65.00 (0.68)	53.25 (3.74)	7.00 (2.46)	9.33 (2.63)	33.65 (1.27)	.798	1.008×
Clip-Cov	67.46 (1.46)	55.00 (2.04)	7.67 (3.53)	8.67 (2.81)	34.70 (1.27)	.810	1.010×
OPEFO	70.08 (1.18)	54.25 (1.21)	8.00 (2.33)	9.67 (2.92)	35.50 (1.17)	.809	1.004×
Random matched-mass KL	64.94 (2.26)	53.75 (2.43)	7.33 (2.11)	9.33 (3.44)	33.84 (1.48)	.793	1.015×
CAGE-D	70.54 (1.30)	54.50 (1.97)	8.67 (3.22)	12.00 (2.81)	36.43 (1.29)	.830	1.045×
CAGE	73.06 (1.61)	57.50 (2.04)	9.00 (2.25)	10.67 (2.63)	**37.56 (1.28)**	**.858**	1.055×

7.2 CAGE IMPROVES ACCURACY, RETENTION, AND HIGH-k COVERAGE

Continuous CAGE improves accuracy; CAGE-D isolates selector-specific retention. Against OPEFO, CAGE gains +2.06 pp in macro Pass@1. With selective-KL construction and token budget matched to random, CAGE-D improves R^{mass} by +.037; it also exceeds OPEFO by +0.93 pp in macro Pass@1 (Table 5), so the accuracy gain extends to discrete selection. MATH500 and AMC23 support the gain; AIME estimates are noisier at $n = 30$ (Appendix A.32) and are not used as standalone evidence here.

CAGE restores high-k performance to the base-model range, and this diagnostic separates selective KL from the incremental ranking signal. GRPO shifts from +3.0 pp over base at $k = 64$ to -2.3 pp at $k = 512$. At $k = 512$, base/GRPO/CAGE reach 91.8/89.5/92.7; the CAGE–base interval includes zero (Table 35). In the construction-matched discrete comparison, CAGE-D exceeds matched random in point estimate (92.1 vs. 91.5); this increment is not separately tested. Thus selective KL accounts for substantial recovery, while the available point estimates are consistent with an additional ranking contribution; evidence supports recovery to, rather than beyond, the base-model range (Figure 2c). Pass@1 and Pass@512 therefore probe distinct failure modes: single-sample accuracy can rise while high-k success falls, so low-k gains alone do not establish preserved coverage.

7.3 RANKING SPECIFICITY AND ALTERNATIVE EXPLANATIONS

With matched budgets, discrete controls isolate CAGE-D ranking from generic selective regularization and source-update magnitude. CAGE-D exceeds matched random by +2.59 pp in macro Pass@1 and +.037 in R^{mass}. Source-strength matching recovers about 12% of this accuracy gain, with CAGE-D still +2.28 pp higher. Anti-targeting underperforms matched random (Pass@1/R^{mass}: 32.35/.756 vs. 33.84/.793), supporting directional information beyond matched KL mass.

Against uniform KL (global regularization), CAGE raises retention .798→.858 and macro Pass@1 33.65→37.56. Mean-probe CAGE recovers 54% of the matched-random-to-CAGE gain; full CAGE remains +1.70 pp higher in macro Pass@1 and +.033 in R^{mass}. Because aggregation order changes, this contrast identifies the full construction, not individualized probe information alone. See Appendix A.29 (specificity/cost), Appendix A.28 (optimizer/intervention), and Table 13 (fidelity).

Table 5: Primary paired contrasts. Macro contrasts use unrounded seed-wise values. Published-baseline intervals are Bonferroni-adjusted across six baselines; uniform-KL and matched-control intervals are paired-t over ten seeds. CAGE–Clip-Cov is a secondary published-method comparison.

	CAGE − OPEFO	CAGE − Uniform KL	CAGE − Clip-Cov	CAGE − mean-probe control	CAGE-D − OPEFO	CAGE-D − random	CAGE-D − source-matched
Macro Pass@1 Δ [95% CI]	+2.06 [+1.30, +2.81]	+3.91 [+2.60, +5.23]	*not a target contrast*	+1.70 [+1.20, +2.20]	+0.93 [+0.37, +1.49]	+2.59 [+1.89, +3.29]	+2.28 [+1.73, +2.83]
R^{mass} Δ [95% CI]	*not a target contrast*	+.060 [+.031, +.089]	+.048 [+.017, +.079]	+.033 [+.015, +.051]	*not a target contrast*	+.037 [+.021, +.053]	+.031 [+.006, +.056]

7.4 Transfer across scale, model family, and code

Transfer stays positive but attenuates with distribution shift (Table 6). Qwen-7B/Llama show positive continuous-CAGE accuracy/retention contrasts over OPEFO and positive CAGE-D point estimates; Coder-7B retains positive retention but continuous-CAGE accuracy CI includes zero. Continuous-CAGE gains over GRPO are +4.06/+3.33/+2.32/+1.69 pp across primary/Qwen-7B/Llama/code. The Llama gain spans four benchmarks (+0.66 to +1.00 pp each), not one 30-item AIME cell.

Table 6: Transfer endpoints for both CAGE variants. Continuous-CAGE contrasts are versus OPEFO with paired-t 95% CIs over matched seeds ($n = 10/5/5$); retention is primary. Benchmark-level values are in Appendices A.24–A.25; measured cost is in Table 30 (Appendix A.27).

Cell	Variant	R^{mass}	Macro	OPEFO R^{mass}	OPEFO Macro	Macro Pass@1 Δ [95% CI]	ΔR [95% CI]
Qwen-7B	CAGE	.874	47.59	.846	45.63	+1.96 $[+1.05, +2.87]$	+.028 $[+.014, +.042]$
	CAGE-D	.867	46.99				
Llama-8B	CAGE	.831	28.40	.807	27.62	+0.78 $[+0.28, +1.28]$	+.024 $[+.006, +.042]$
	CAGE-D	.823	28.02				
Coder-7B	CAGE	.823	55.69	.795	55.19	+0.50 $[-0.18, +1.18]$	+.028 $[+.006, +.050]$
	CAGE-D	.814	55.48				

7.5 Counterfactual fidelity of the deployed score

Optimizer conditioning improves agreement with the finite leave-one-out counterfactual: the Jaccard overlap J_5 between the score-selected and explicitly recomputed harmful top-5% rises from .716 to .791 on tied 3B and from .754 to .861 on the closure-exact untied companion. This fidelity difference is behaviorally consequential under the same discrete intervention: replacing the optimizer-aware score with its Euclidean counterpart lowers macro Pass@1 from 36.43 to 35.54 and retention from .830 to .812 (Appendix A.28, Table 31). The direct-head approximation preserves the harmful tail (head-only versus full first-order $J_5 = .927$), and on the closure-exact untied companion CAGE remains +2.40 pp above matched random targeting; fidelity and interaction audits are in Appendices A.19 and A.28.

8 Limitations

CAGE requires held-out correct probes unused directly by reproduced baselines; matched controls isolate ranking value given probe access, not probe value itself. CAGE does not equate predictive entropy with reasoning support; frozen correct-prefix entropy operationalizes local support loss. CAGE is conditional on Q_{probe}: changing the protected probe population could change token rankings even with the training batch fixed. Sensitivity to such population shifts remains untested. Code-transfer accuracy CI vs. OPEFO includes zero, though retention is positive; see Appendix A.33.

9 Conclusion

Which updates erase what the base model knew? CAGE ranks current updates by predicted cross-context damage to base-supported reasoning routes. Across prediction and intervention, this ranking identifies update-level risk beyond entropy and simpler attribution comparators while improving accuracy and preserving base-supported coverage, subject to the reported uncertainty. Probe-resource value and same- versus cross-context attribution remain open questions for controlled follow-up.

REPRODUCIBILITY STATEMENT

The analysis protocol specifies data partitions, independent discovery/reference/evaluation streams, family construction, primary/secondary horizons, the endpoint hierarchy including the $k = 512$ high-k endpoint, and inference procedures. Section 6 and the supplementary experimental protocols record the corresponding configuration, power analyses, robustness checks, and reporting rules. The supplementary source records the complete training recipe and result-value manifest used by the manuscript. Code is provided to support reproducibility.

AI USE STATEMENT

Generative AI tools were used to aid and polish manuscript writing, including language editing, organization, and consistency checking. All AI-assisted changes were reviewed by the authors, who take responsibility for the final content of the submission.

REFERENCES

Daya Guo et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645:633–638, 2025.

Yi Ren and Danica J. Sutherland. Learning Dynamics of LLM Finetuning. *ICLR*, 2025.

Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. LESS: Selecting Influential Data for Targeted Instruction Tuning. *ICML*, 2024.

Ganqu Cui et al. The entropy mechanism of reinforcement learning for reasoning language models. arXiv:2505.22617, 2025.

Jingcheng Hu et al. Open-Reasoner-Zero: An Open Source Approach to Scaling Up Reinforcement Learning on the Base Model. *NeurIPS*, 2025.

Renren Jin et al. Revisiting entropy in reinforcement learning for large reasoning models. *Findings of ACL*, 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. *ICLR*, 2026.

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. *ICML*, 2017.

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. *ACL*, 2026.

Zhihong Shao et al. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. arXiv:2402.03300, 2024.

Huimin Xu et al. Understanding and preventing entropy collapse in RLVR with on-policy entropy flow optimization. *Findings of ACL*, 2026.

Qiying Yu et al. DAPO: An open-source LLM reinforcement learning system at scale. *NeurIPS*, 2025.

Shenzhi Wang et al. Beyond the 80/20 Rule: High-Entropy Minority Tokens Drive Effective Reinforcement Learning for LLM Reasoning. *NeurIPS*, 2025.

Garima Pruthi, Frederick Liu, Satyen Kale, and Mukund Sundararajan. Estimating training data influence by tracing gradient descent. *NeurIPS*, 2020.

Andrew Ilyas, Sung Min Park, Logan Engstrom, Guillaume Leclerc, and Aleksander Madry. Datamodels: Understanding predictions with data and data with predictions. *ICML*, 2022.

Sung Min Park, Kristian Georgiev, Andrew Ilyas, Guillaume Leclerc, and Aleksander Madry. TRAK: Attributing Model Behavior at Scale. *ICML*, 2023.

Andrea Schioppa, Polina Zablotskaia, David Vilar, and Artem Sokolov. Scaling up influence functions. *AAAI*, 2022.

Chih-Kuan Yeh, Joon Sik Kim, Ian E. H. Yen, and Pradeep Ravikumar. Representer Point Selection for Explaining Deep Neural Networks. *NeurIPS*, 2018.

Guillaume Charpiat, Nicolas Girard, Loris Felardos, and Yuliya Tarabalka. Input Similarity from the Neural Network Perspective. *NeurIPS*, 2019.

Roger Grosse et al. Studying Large Language Model Generalization with Influence Functions. arXiv:2308.03296, 2023.

Wenlong Deng, Qi Zeng, Jiaming Zhang, Minghui Chen, Zixin Ding, Christos Thrampoulidis, Boying Gong, and Xiaoxiao Li. For-Value: Efficient Forward-Only Data Valuation for finetuning LLMs and VLMs. *ACL*, 2026.

Yang Yue et al. Does Reinforcement Learning Really Incentivize Reasoning Capacity in LLMs Beyond the Base Model? *NeurIPS*, 2025.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *ICLR*, 2015.

James Martens and Roger Grosse. Optimizing neural networks with Kronecker-factored approximate curvature. *ICML*, 2015.

Vineet Gupta, Tomer Koren, and Yoram Singer. Shampoo: Preconditioned stochastic tensor optimization. *ICML*, 2018.

James H. Stock and Mark W. Watson. Heteroskedasticity-robust standard errors for fixed effects panel data regression. *Econometrica*, 76(1):155–174, 2008.

A. Colin Cameron, Jonah B. Gelbach, and Douglas L. Miller. Bootstrap-based improvements for inference with clustered errors. *Review of Economics and Statistics*, 90(3):414–427, 2008.

Qwen et al. Qwen2.5 Technical Report. arXiv:2412.15115, 2024.

An Yang et al. Qwen2.5-Math Technical Report: Toward Mathematical Expert Model via Self-Improvement. arXiv:2409.12122, 2024.

Dan Hendrycks et al. Measuring Mathematical Problem Solving With the MATH Dataset. *NeurIPS Datasets and Benchmarks*, 2021.

Hunter Lightman et al. Let's Verify Step by Step. *ICLR*, 2024.

Aaron Grattafiori et al. The Llama 3 Herd of Models. arXiv:2407.21783, 2024.

Binyuan Hui et al. Qwen2.5-Coder Technical Report. arXiv:2409.12186, 2024.

Rongao Li et al. TACO: Topics in Algorithmic Code Generation Dataset. arXiv:2312.14852, 2023.

Jiawei Liu, Chunqiu Steven Xia, Yuyao Wang, and Lingming Zhang. Is Your Code Generated by ChatGPT Really Correct? Rigorous Evaluation of Large Language Models for Code Generation. *NeurIPS*, 2023.

A Supplementary Theory, Diagnostics, and Experimental Protocols

A.1 Audit map and canonical evidence locations

Table 7 maps each headline claim to its canonical evidence and the corresponding interpretive scope; this is the recommended entry point to the supplement.

Table 7: Endpoint hierarchy and audit map. The last column states the interpretation supported by each evidence block.

Question / claim	Role	Canonical evidence	Interpretive scope
Prospective prediction beyond entropy	Primary	Table 3; App. A.22	Supports prospective attribution beyond entropy.
Horizon/comparator persistence	Secondary	Apps. A.22, A.31	Assesses persistence and localization across horizons/comparators.
Targeted vs matched controls	Primary	Tables 4, 5; App. A.27	Supports selector-specific intervention effects.
Accuracy / retention effects	Primary	Table 5; App. A.32	Quantifies behavioral accuracy and retention effects.
High-k reversal ($k = 512$ primary)	Primary	Fig. 2c; Table 35	Tests the motivating high-k reversal.
Qwen-7B targeted retention transfer	Primary	App. A.24	Tests targeted-retention transfer at 7B.
Finite-counterfactual score fidelity	Audit	Table 13; App. A.19	Quantifies fidelity of the deployed approximation.
Correct-family retention definition	Audit	App. A.20; Tables 15, 16	Validates the correct-family retention construction.
Optimizer-aware calibration	Audit	App. A.28; Table 31	Isolates the contribution of optimizer conditioning.
Cross-family / code transfer	Secondary	Apps. A.24–A.25	Assesses generalization beyond the primary setting.

PART I: COUNTERFACTUAL FOUNDATIONS AND CAUSAL ESTIMANDS

Supports Sections 3–4: this Part defines the token-removal estimand and the exact scope of the direct-head closure; the key boundary is the tied/body decomposition.

A.2 NOTATION AND SIGN CONVENTIONS

Table 8 fixes the conventions used throughout the derivations. In particular, $r_i = e_{y_i} - \pi_i$ is the log-probability residual, whereas $\phi_a = \pi_a \odot (\log \pi_a + H_a)$ is defined so that $-\nabla_W H_a = \phi_a h_a^\top$. With the counterfactual convention in Eq. (2), a positive score means that including the training token lowers probe entropy; positive therefore means entropy sharpening and is the harmful tail selected by CAGE. The finite-LOO audit reports sign agreement explicitly so the implementation direction is independently checked.

A.3 COMPLETE FIRST-ORDER DERIVATION

Let $F_a(\theta) = H_a(\theta)$ and write the full Euclidean post-update state as $x = \theta + \Delta\theta$. With $d = \delta_i = -(\eta/N)\alpha_i g_i$, removing instance i gives the state $x + d$, so the counterfactual convention in Eq. (2) is

$$C_{i,a} = F_a(x + d) - F_a(x) = \int_0^1 \nabla F_a(x + \tau d)^\top d \, d\tau. \tag{13}$$

Adding and subtracting $\nabla F_a(\theta)$ gives

$$C_{i,a} = \nabla H_a(\theta)^\top \delta_i + R_{i,a}, \tag{14}$$

$$R_{i,a} = \int_0^1 [\nabla H_a(\theta + \Delta\theta + \tau\delta_i) - \nabla H_a(\theta)]^\top \delta_i \, d\tau. \tag{15}$$

If ∇H_a is L_H-Lipschitz on the update segment,

$$|R_{i,a}| \le L_H \int_0^1 \|\Delta\theta + \tau\delta_i\| \, \|\delta_i\| d\tau \tag{16}$$

$$\le L_H \|\Delta\theta\| \|\delta_i\| + \frac{L_H}{2} \|\delta_i\|^2. \tag{17}$$

Table 8: Core notation and sign conventions.

Symbol	Meaning
c_i, y_i	training prefix and sampled next token
c_a	frozen held-out probe context
α_i	signed GRPO advantage weight multiplying $\nabla \log \pi(y_i \mid c_i)$
$r_i = e_{y_i} - \pi_i$	log-probability residual in vocabulary space
$\phi_a = \pi_a \odot (\log \pi_a + H_a)$	entropy-side vocabulary vector
$s_{i,a}$	first-order contribution; positive means entropy sharpening
$S^{\mathrm{harm}}_{i,a} = \max\{s^{\mathcal{U}}_{i,a}, 0\}$	nonnegative harmful exposure
K	number of online probes
M_{vocab}	sparse-vocabulary truncation size
$R_x(t)$	correct-family retention on outcome problem x
$A^{(25)}_{g,t}$	preceding-25-step cross-fitted harmful exposure for group g at step t
J_5	harmful top-5% Jaccard agreement with the finite counterfactual
$J_5^{\mathrm{opt}}, J_5^{\mathrm{probe}}, J_5^{\mathrm{trunc}}, J_5^{\mathrm{body}}, J^{\mathrm{tie}}_{5,+/-}$	optimizer/probe/truncation/body/tied-tail variants of J_5
$J^{\mathrm{opt}}_{5,\mathrm{diag}}$	optimizer-aware J_5 in the diagnostic fidelity audit
J_5^{target}	harmful-top-5% Jaccard against the finite target named by the fidelity row
$J_{\mathrm{KL}}, J_{\mathrm{body}}, J_{\mathrm{total}}$	selector-feedback Jaccards for KL feedback, body augmentation, and the resulting total mask change
$R^{(c)}_{M_{\mathrm{vocab}}}, R^{(s)}_{M_{\mathrm{vocab}}}$	retained-mass ratios for entropy-side and scored sparse support (not family-retention R)

With $\|\Delta\theta\| = O(\eta)$ and $\|\delta_i\| = O(\eta/N)$, the leading leave-one remainder is $O(\eta^2/N)$; the $O(\eta^2/N^2)$ term is secondary.

Substituting $\delta_i = -(\eta/N)\alpha_i g_i$ yields

$$C_{i,a} = -\frac{\eta}{N}\alpha_i \langle \nabla H_a, g_i \rangle + R_{i,a}. \tag{18}$$

The normalized first-order score used for ranking is therefore $s_{i,a} = -\alpha_i \langle \nabla H_a, g_i \rangle$; multiplying all instances in a batch by the common positive factor η/N leaves the selector unchanged.

General conditional optimizer calculus. This section separates the counterfactual definition from any particular optimizer. Let the current batch objective, after fixing all selector-side quantities, induce an aggregate gradient $G_t \in \mathbb{R}^p$. Write a generic optimizer step as

$$\theta_t^+(G) = \theta_t - \eta\, \mathcal{U}_t(G), \tag{19}$$

where the subscript t means that all optimizer state inherited from previous steps is held fixed. For instance i, let $\delta G_i = -N^{-1}\alpha_i g_i$ denote only the RL-driving contribution that is removed. The selector mask m, reference policy, previous moments, learning-rate schedule, and weight-decay state are part of the conditioning set and are not recomputed under the removal. The conditional finite counterfactual is

$$C^{\mathcal{U}}_{i,a}(m) = H_a\big(\theta_t^+(G_t(m) - \delta G_i)\big) - H_a\big(\theta_t^+(G_t(m))\big)\,. \tag{20}$$

Lemma 1 (optimizer marginalization). Assume $\mathcal{U}_t$ is Fréchet differentiable in a neighborhood of G_t. With

$$P_t := D\mathcal{U}_t[G_t], \tag{21}$$

we have

$$\theta_t^+(G_t - \delta G_i) - \theta_t^+(G_t) = \eta P_t \delta G_i + \varepsilon^{\mathcal{U}}_{i,t}, \qquad \|\varepsilon^{\mathcal{U}}_{i,t}\| \le \frac{\eta L_{\mathcal{U}}}{2}\|\delta G_i\|^2, \tag{22}$$

whenever $D\mathcal{U}_t$ is $L_{\mathcal{U}}$-Lipschitz on the connecting segment.

Proof. Apply the fundamental theorem of calculus in gradient space:

$$\mathcal{U}_t(G_t) - \mathcal{U}_t(G_t - \delta G_i) = \int_0^1 D\mathcal{U}_t[G_t - \tau\delta G_i]\,\delta G_i\, d\tau \tag{23}$$

$$= P_t \delta G_i + \int_0^1 \{D\mathcal{U}_t[G_t - \tau\delta G_i] - P_t\}\delta G_i\, d\tau. \tag{24}$$

The second term is bounded by $L_{\mathcal{U}}\|\delta G_i\|^2/2$. Multiplying by η proves the claim. □

Proposition 3 (conditional optimizer-aware first order; full remainder). Suppose H_a has L_H-Lipschitz gradient on the relevant parameter segment and the assumptions of Lemma 1 hold. Then

$$C^{\mathcal{U}}_{i,a}(m) = -\frac{\eta\alpha_i}{N}\,\langle \nabla H_a(\theta_t), P_t g_i\rangle + R^{\mathcal{U}}_{i,a}, \tag{25}$$

where $\Delta_t = \theta_t^+(G_t) - \theta_t$, $d_{i,t} = \eta P_t \delta G_i$, and $z_{i,t} = d_{i,t} + \varepsilon^{\mathcal{U}}_{i,t}$. The non-asymptotic remainder obeys

$$|R^{\mathcal{U}}_{i,a}| \le \|\nabla H_a(\theta_t)\|\,\|\varepsilon^{\mathcal{U}}_{i,t}\| + L_H\|\Delta_t\|\,\|z_{i,t}\| + \frac{L_H}{2}\|z_{i,t}\|^2. \tag{26}$$

Consequently, under the stated boundedness assumptions, $\|d_{i,t}\| = O(\eta/N)$ and $\|\varepsilon^{\mathcal{U}}_{i,t}\| = O(\eta/N^2)$, giving $R^{\mathcal{U}}_{i,a} = O(\eta^2/N + \eta/N^2)$. For fixed positive η, the $O(\eta^2/N)$ cross-update term dominates as $N \to \infty$.

Proof. Let $x = \theta_t + \Delta_t$. By Lemma 1 the removed state is $x + z_{i,t}$. Applying the fundamental theorem of calculus to $H_a(x + z_{i,t}) - H_a(x)$ and subtracting $\nabla H_a(\theta_t)^\top d_{i,t}$ yields the term $\nabla H_a(\theta_t)^\top \varepsilon^{\mathcal{U}}_{i,t}$ plus a path integral of $[\nabla H_a(x + \tau z_{i,t}) - \nabla H_a(\theta_t)]^\top z_{i,t}$. Lipschitz continuity gives the stated bound: the $\Delta_t \times d_{i,t}$ term is $O(\eta^2/N)$, the explicit optimizer-nonlinearity term is $O(\eta/N^2)$, and the remaining cross/quadratic terms are no larger under the stated bounds. □

Corollary 1 (Euclidean special case). For SGD, $\mathcal{U}_t(G) = G$, hence $P_t = I$ and $L_{\mathcal{U}} = 0$. Eq. (25) reduces exactly to the Euclidean first-order score derived above; the Euclidean cross-context score is the $P_t = I$ special case.

A.4 Direct LM-head closure and tied-embedding scope

Let the vocabulary head be $z = Wh$, $\pi = \mathrm{softmax}(z)$. The softmax Jacobian gives

$$\frac{\partial H}{\partial z_v} = -\pi_v(\log \pi_v + H) = -\phi_v, \qquad \frac{\partial \log \pi_y}{\partial z_v} = \mathbf{1}[v = y] - \pi_v = r_v. \tag{27}$$

For an untied output block, $\partial z_v/\partial w_v = h$ and $\partial z_v/\partial w_u = 0$ for $u \neq v$, so

$$-\nabla_W H_a = \phi_a h_a^\top, \qquad \nabla_W \log \pi(y_i \mid c_i) = r_i h_i^\top. \tag{28}$$

Using $\langle ab^\top, cd^\top\rangle_F = (a^\top c)(b^\top d)$,

$$s^H_{i,a} = +\alpha_i(\phi_a^\top r_i)(h_a^\top h_i). \tag{29}$$

This identity is exact for the direct output path and requires only quantities already available after policy/probe forward passes. It equals the full output-head-block contribution only when the output matrix is untied; Appendix A.5 gives the tied decomposition. The kernel $h_a^\top h_i$ arises from the shared linear head.

Vocabulary-row-block closure.

Proof of Proposition 1 (forward closure under vocabulary-row-block geometry). Assume the output embedding is not shared with the input embedding. Let P_H be any linear preconditioner on the LM head that is block diagonal across vocabulary rows,

$$P_H = \mathrm{blockdiag}(D_1, \ldots, D_V), \qquad D_v \in \mathbb{R}^{d\times d}. \tag{30}$$

Then the preconditioned head influence satisfies

$$s^{H,P}_{i,a} = +\alpha_i \sum_{v=1}^{V} \phi_{a,v} r_{i,v}\, h_a^\top D_v h_i. \tag{31}$$

Proof. The vth row of $-\nabla_W H_a$ is $\phi_{a,v} h_a^\top$ and the vth row of $\nabla_W \log \pi_i$ is $r_{i,v} h_i^\top$. Row-block diagonality prevents cross-vocabulary mixing, so the bilinear form decomposes as a sum of within-row forms. Each row contributes $+\alpha_i \phi_{a,v} r_{i,v} h_a^\top D_v h_i$. Summing proves the result. □

The architectural requirement is therefore weaker than coordinate-wise diagonality. Euclidean SGD corresponds to $D_v = cI$; diagonal Adam-type local metrics correspond to diagonal D_v; row-wise second-order or low-rank metrics are also covered whenever they do not mix vocabulary rows. The closure applies locally to preconditioners that preserve vocabulary-row blocks.

Equivalent stacked form and invariances. Vectorizing W by rows gives

$$-\nabla_W H_a = \mathrm{stack}_{v=1}^{V}(\phi_{a,v} h_a^\top), \qquad \nabla_W \log \pi_i(y_i) = \mathrm{stack}_{v=1}^{V}(r_{i,v} h_i^\top) \tag{32}$$

$$-\alpha_i \langle \nabla_W H_a, P_H \nabla_W \log \pi_i(y_i) \rangle = +\alpha_i \sum_{v=1}^{V} \phi_{a,v} r_{i,v}\, h_a^\top D_v h_i. \tag{33}$$

No orthogonality assumption on the output embeddings is required. The result is invariant to within-row rotations and permits dense or low-rank within-row metrics; only cross-vocabulary mixing breaks the closure.

Corollary 2 (diagonal adaptive optimizer). If $D_v = \mathrm{diag}(d_{t,v})$, then

$$\kappa_{t,v}^{P}(a,i) = h_a^\top D_v h_i = (h_a \odot h_i)^\top d_{t,v}. \tag{34}$$

All $d_{t,v}$ are read from optimizer state and the current forward-computable head gradient; no backward pass through a probe is required online.

Corollary 3 (Euclidean reduction). If $D_v = c_t I$ for every v, then

$$s_{i,a}^{H,P} = +c_t \alpha_i (\phi_a^\top r_i)(h_a^\top h_i). \tag{35}$$

The common positive c_t does not alter a within-batch top-k selector. This is the precise sense in which the simpler cross-context kernel is a special case.

Boundary of the closure. If the preconditioner contains blocks P_{uv} with $u \neq v$, Eq. (33) acquires terms

$$+\alpha_i \sum_{u \neq v} \phi_{a,u} r_{i,v}\, h_a^\top P_{uv} h_i, \tag{36}$$

so vocabulary-diagonal collapse no longer holds. The closure therefore does not extend to preconditioners that mix vocabulary rows.

A.5 Tied embeddings: exact decomposition of the shared-block discrepancy

When input and output embeddings are shared, write $\nabla_W H_a = d_a + t_a$ and $g_{i,W} = d_i + t_i$, with d the direct output path and t the input-mediated path. AdamW's local operator also depends on the current aggregate gradient. Holding inherited state fixed, let $G_W^{\mathrm{full}} = G_W^{\mathrm{dir}} + G_W^{\mathrm{input}}$ and $P_W^{\mathrm{full}} = D\mathcal{U}_t[G_W^{\mathrm{full}}]$, $P_W^{\mathrm{dir}} = D\mathcal{U}_t[G_W^{\mathrm{dir}}]$. The full shared-block term minus the deployed direct-head term is

$$\begin{aligned} R_{i,a}^{\mathrm{tied}} = & \underbrace{\langle d_a, (P_W^{\mathrm{full}} - P_W^{\mathrm{dir}}) d_i \rangle}_{R_{i,a}^{\mathrm{op}}} \\ & + \underbrace{\langle d_a, P_W^{\mathrm{full}} t_i \rangle + \langle t_a, P_W^{\mathrm{full}} d_i \rangle + \langle t_a, P_W^{\mathrm{full}} t_i \rangle}_{R_{i,a}^{\mathrm{mixed}}}. \end{aligned} \tag{37}$$

Thus tied error contains the three mixed paths plus an operator mismatch; $R^{\mathrm{op}} = 0$ for gradient-independent Euclidean geometry but not generally for AdamW. G_W^{input} includes current input-mediated contributions from the active objective, including selective regularization. AdamW also has one inherited shared m_W, v_W, so no unique post-hoc input/output moment split exists. We use artificial untying only for the Euclidean support audit; the tied 3B finite-counterfactual audit applies the actual full shared update and therefore absorbs both R^{mixed} and R^{op}. In the tied 3B offline decomposition, the target-mixed, source-mixed, and tie–tie terms carry .12/.10/.03 of the direct–direct magnitude, with $J_{5,+}^{\mathrm{tie}} = .79$ and $J_{5,-}^{\mathrm{tie}} = .67$.

A.6 Conditional AdamW marginal operator

For completeness we derive the exact current-step differential underlying the optimizer-aware audit. Fix the pre-update optimizer state (m_{t-1}, v_{t-1}) and intervention mask. For a batch gradient G_t,

$$m_t = \beta_1 m_{t-1} + (1-\beta_1) G_t, \qquad v_t = \beta_2 v_{t-1} + (1-\beta_2) G_t^{\odot 2}, \tag{38}$$

$$\hat{m}_t = m_t/(1-\beta_1^t), \qquad \hat{v}_t = v_t/(1-\beta_2^t). \tag{39}$$

The conditional AdamW update at the current step is

$$\theta_{t+1} = (1 - \eta\lambda_{\mathrm{wd}})\theta_t - \eta \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon}, \tag{40}$$

where all operations in the adaptive term are coordinate-wise. The decoupled weight-decay term does not depend on the current instance contribution.

AdamW differential from first principles. Fix coordinate j, write $g = G_{t,j}$, and define

$$\hat{m}(g) = \frac{\beta_1 m_{t-1,j} + (1-\beta_1)g}{1-\beta_1^t}, \qquad \hat{v}(g) = \frac{\beta_2 v_{t-1,j} + (1-\beta_2)g^2}{1-\beta_2^t}, \tag{41}$$

$$a_t := \frac{1-\beta_1}{1-\beta_1^t}, \qquad b_t := \frac{1-\beta_2}{1-\beta_2^t}. \tag{42}$$

With $s(g) = \sqrt{\hat{v}(g)}$, $d\hat{m}/dg = a_t$, $d\hat{v}/dg = 2b_t g$, and $ds/dg = b_t g/s$. Hence for $u(g) = \hat{m}(g)/(s(g)+\epsilon)$,

$$(P_t)_{jj} := \frac{du}{dg} = \frac{a_t}{s_j + \epsilon} - \frac{b_t \hat{m}_{t,j} G_{t,j}}{s_j (s_j + \epsilon)^2}. \tag{43}$$

Thus a small removal δG_i changes the current AdamW step by $\eta P_t \delta G_i + O(\|\delta G_i\|^2)$. Decoupled weight decay is independent of the current instance contribution and therefore drops from this marginal. The second-moment term is first order and is retained.

Deployed stabilization. The exact differential in Eq. (43) is unfloored. For deployed scoring we stabilize only its denominators, defining a surrogate rather than the Jacobian of the hard-floored map $\hat{m}/(\max\{s, \sqrt{\nu}\} + \epsilon)$, whose floor branch has a different derivative. Let $\tilde{s}_j = \max\{s_j, \sqrt{\nu}\}$ and define

$$(P_t^{(\nu)})_{jj} := \frac{a_t}{\tilde{s}_j + \epsilon} - \frac{b_t \hat{m}_{t,j} G_{t,j}}{\tilde{s}_j (\tilde{s}_j + \epsilon)^2}. \tag{44}$$

The reported rank/Jaccard audits test this stabilized surrogate against the unfloored conditional target and show high agreement (Appendix A.19); the stabilized surrogate is therefore used in the reported implementation.

Remark (sign indefiniteness). The first term in Eq. (43) is positive, but the second has the sign of $\hat{m}G$. Thus P_t need not be positive semidefinite coordinate-wise; the optimizer-aware kernel is an interaction kernel rather than a metric similarity in the strict geometric sense.

Lemma 2 (stabilized-surrogate regularity). Let $P_t^{(\nu)}$ be the surrogate in Eq. (44). Since

$$\hat{v}_{t,j} \geq b_t G_{t,j}^2, \tag{45}$$

we have $|G_{t,j}| \leq s_j/\sqrt{b_t} \leq \tilde{s}_j/\sqrt{b_t}$. If $|\hat{m}_{t,j}| \leq G_{\infty,j} := \sup_{s \leq t} |G_{s,j}|$, then

$$|P_{t,j}^{(\nu)}| \leq \frac{a_t}{\sqrt{\nu} + \epsilon} + \frac{\sqrt{b_t} G_{\infty,j}}{(\sqrt{\nu} + \epsilon)^2} =: C_{P,j}(\nu). \tag{46}$$

The moment bound follows because bias-corrected $\hat{m}_t$ is a convex combination of past gradients. Let $C_P(\nu) = \max_j C_{P,j}(\nu)$.

Corollary 4 (error from stabilized AdamW scoring). Proposition 3 controls the exact unfloored marginal P_t. For $\Delta P_t^{(\nu)} = P_t^{(\nu)} - P_t$,

$$\left| C_{i,a}^{\mathcal{U}} + \frac{\eta \alpha_i}{N} \langle \nabla H_a, P_t^{(\nu)} g_i \rangle \right| \leq |R_{i,a}^{\mathcal{U}}| + \frac{\eta |\alpha_i|}{N} \|\nabla H_a\| \, \|\Delta P_t^{(\nu)}\|_{\mathrm{op}} \, \|g_i\|. \tag{47}$$

Thus stabilization trades numerical regularity against an explicit surrogate-bias term; the ν sweep and finite-counterfactual rank/Jaccard audit test whether it changes the deployed ranking materially.

Current-step versus future momentum. Eq. (43) is conditional on the current optimizer state. Removing an instance also perturbs future first moments by factors proportional to $(1-\beta_1)\beta_1^k$. If subsequent gradients and preconditioners are frozen, these factors sum to one across an infinite horizon, while the immediate first-moment contribution is approximately $1-\beta_1$. In the scalar-preconditioner limit this rescales every current-batch instance by the same positive factor and leaves ranking invariant. With coordinate-dependent future preconditioners the equivalence is approximate, so we make only the conditional one-step claim and audit finite-step ranking directly.

A.7 Forward-computable current head gradient with selective KL

Optimizer awareness requires the current aggregate head gradient. For the RL term its direct-output component is available from the policy forward pass: Throughout this section G denotes the gradient of the minimized training loss, so the policy-gradient contribution is $-\alpha_i \nabla \log \pi_i / N$; the KL contribution below already has the ordinary loss-gradient sign.

$$G_H^{\mathrm{RL}} = -\frac{1}{N}\sum_i \alpha_i r_i h_i^\top. \tag{48}$$

For $D_i = \mathrm{KL}(\pi_i \| \pi_{\mathrm{ref,i}})$, define

$$u_i = \pi_i \odot \left(\log \frac{\pi_i}{\pi_{\mathrm{ref,i}}} - D_i \mathbf{1}\right). \tag{49}$$

Direct differentiation of $\sum_v \pi_v \log(\pi_v/\pi_{\mathrm{ref},v})$ through the softmax gives $\partial D_i/\partial z_i = u_i$. Therefore, for a fixed binary mask m and normalization $Z_m = \sum_i m_i + \varepsilon_m$,

$$G_H(m) = G_H^{\mathrm{RL}} + \frac{\lambda}{Z_m}\sum_i m_i u_i h_i^\top. \tag{50}$$

Every term in Eq. (50) is available from policy/reference forward passes. On an untied output head this is the relevant current head gradient. On a tied shared matrix Eq. (50) constructs P_W^{dir}, omitting the current input-mediated component of G_W^{full}; Appendix A.5 includes this operator mismatch in the deployed approximation.

One-step self-consistency. Because the top-k map is discontinuous, we use a single explicit update:

$$m^{(0)} = \mathcal{M}(S_H(m=0)), \tag{51}$$

$$G_H^{(1)} = G_H(m^{(0)}), \tag{52}$$

$$m^{(1)} = \mathcal{M}(S_H(m^{(0)}) + S_B). \tag{53}$$

The counterfactual estimand conditions on the chosen mask; the iteration constructs the deployed selector online.

Continuous one-step conditioning. For $Z_w = \sum_i w_{i,t}^{\mathrm{raw}} + \varepsilon_w$, the final continuous head gradient corresponding to Eq. (94) is

$$G_H^{\mathrm{cont}} = G_H^{\mathrm{RL}} + \frac{\lambda\gamma_t}{Z_w}\sum_i w_{i,t}^{\mathrm{raw}} u_i h_i^\top. \tag{54}$$

The deployed arm computes $m^{(0)}$, recomputes the score under $G_H(m^{(0)})$, forms w^{raw} and the EMA scalar γ_t by Eqs. (92)–(94), and executes the final loss once. Attribution is conditioned on $G_H(m^{(0)})$ while the final update uses G_H^{cont}; the J_{KL} audit below applies to binary selector feedback.

Perturbation audit. Because the KL contribution in Eq. (50) is proportional to λ, we report the observed rank-boundary sensitivity

$$J_{\mathrm{KL}} = J\{\mathcal{M}(S_H(0)), \mathcal{M}(S_H(m^{(0)}))\}, \tag{55}$$

$$J_{\mathrm{body}} = J\{\mathcal{M}(S_H(m^{(0)})), \mathcal{M}(S_H(m^{(0)}) + S_B)\}, \tag{56}$$

$$J_{\mathrm{total}} = J(m^{(0)}, m^{(1)}). \tag{57}$$

Only J_{KL} is interpreted as the empirical counterpart of an $O(\lambda)$ feedback perturbation. For the deployed head-only configuration, $J_{\mathrm{KL}} = .964$ and $J_{\mathrm{total}} = .964$; the separate offline body diagnostic gives $J_{\mathrm{body}} = .927$.

A.8 TRUNCATION BOUNDS AND RANKING PRESERVATION

Let the exact row-block head score be $S_i = \sum_v a_{i,v}$ with

$$a_{i,v} = +\alpha_i \phi_{a,v} r_{i,v} \kappa^P_{t,v}(a, i). \tag{58}$$

For a retained set $\mathcal{T}_{M_{\text{vocab}}}$, define $\widetilde{S}_i = \sum_{v \in \mathcal{T}_{M_{\text{vocab}}}} a_{i,v}$ and tail radius

$$\epsilon_i^{(M_{\text{vocab}})} := \sum_{v \notin \mathcal{T}_{M_{\text{vocab}}}} |a_{i,v}|. \tag{59}$$

Then, deterministically,

$$|S_i - \widetilde{S}_i| \le \epsilon_i^{(M_{\text{vocab}})}. \tag{60}$$

For two instances i, j,

$$\widetilde{S}_i - \widetilde{S}_j > \epsilon_i^{(M_{\text{vocab}})} + \epsilon_j^{(M_{\text{vocab}})} \quad \Longrightarrow \quad S_i > S_j. \tag{61}$$

Thus the sparse score exactly certifies any pair whose approximate margin exceeds the sum of tail radii.

Top-k certificate. Let τ_{sel} be the kth selection threshold under $\widetilde{S}$. If every selected i satisfies $\widetilde{S}_i - \epsilon_i^{(M_{\text{vocab}})} > \tau^+_{\text{sel}}$ and every unselected j satisfies $\widetilde{S}_j + \epsilon_j^{(M_{\text{vocab}})} < \tau^-_{\text{sel}}$ for some $\tau^-_{\text{sel}} > \tau^+_{\text{sel}}$ arranged around the exact boundary, the selected set is unchanged by restoring the vocabulary tail. The deterministic radii are conservative; Appendix A.16 therefore reports empirical full-vocabulary retained mass and top-tail Jaccard alongside Eq. (61).

A.9 RANKING STABILITY UNDER SCORE ERROR

The intervention consumes a ranking, so calibration error matters only insofar as it crosses the selection boundary. Let exact aggregate scores be S_i and approximations $\widehat{S}_i$ with deterministic or audited error radii $|S_i - \widehat{S}_i| \le \epsilon_i$. Let $\widehat{\tau}$ be the approximate top-k threshold and define the ambiguous set

$$\mathcal{A} = \{i : |\widehat{S}_i - \widehat{\tau}| \le \epsilon_i + \epsilon_\tau\}. \tag{62}$$

All membership changes are confined to $\mathcal{A}$: if an instance lies outside the band, its interval cannot cross the threshold interval. Hence

$$|\mathcal{M}(S) \triangle \mathcal{M}(\widehat{S})| \le 2|\mathcal{A}|, \tag{63}$$

with a sharper one-sided count available when the number selected is fixed. This observation motivates reporting top-5% Jaccard and the empirical boundary-margin distribution in addition to global Spearman correlation.

If the error sources admit radii $\epsilon_i^{\text{vocab}}$, $\epsilon_i^{\text{probe}}$, ϵ_i^{body}, and $\epsilon_i^{\text{feedback}}$, the triangle inequality gives a transparent combined radius

$$\epsilon_i \le \epsilon_i^{\text{vocab}} + \epsilon_i^{\text{probe}} + \epsilon_i^{\text{body}} + \epsilon_i^{\text{feedback}}. \tag{64}$$

A.10 WHY THE TRANSFORMER BODY DOES NOT ADMIT THE SAME EXACT CLOSURE

The exact head result relies on a single shared linear map at the evaluated position. A body matrix W_ℓ is reused at every sequence position. For a scalar sequence loss,

$$\nabla_{W_\ell} L_i = \sum_t \delta_{i,\ell,t} x_{i,\ell,t}^\top, \qquad \nabla_{W_\ell} H_a = \sum_s \delta_{a,\ell,s} x_{a,\ell,s}^\top. \tag{65}$$

Their Frobenius interaction is therefore exactly

$$\langle \nabla_{W_\ell} H_a, \nabla_{W_\ell} L_i \rangle = \sum_{s,t} \langle \delta_{a,\ell,s}, \delta_{i,\ell,t} \rangle \langle x_{a,\ell,s}, x_{i,\ell,t} \rangle. \tag{66}$$

The backward factors δ cannot in general be reconstructed from forward activations alone. Last-position hidden similarity keeps only one term of Eq. (66); it is not an exact body analogue of the head closure.

Low-rank positional factorization. Let the backward-position interaction matrix be $B_{st} = \langle \delta_{a,s}, \delta_{i,t} \rangle$. If

$$B \approx \sum_{r=1}^{R} \lambda_r u_r v_r^\top + E, \tag{67}$$

then

$$\sum_{s,t} B_{st} \langle x_{a,s}, x_{i,t} \rangle \approx \sum_{r=1}^{R} \lambda_r \left\langle \sum_s u_r(s) x_{a,s}, \sum_t v_r(t) x_{i,t} \right\rangle + \mathcal{E}, \tag{68}$$

with

$$|\mathcal{E}| \le \|E\|_{\mathrm{op}} \, \|X_a\|_F \, \|X_i\|_F. \tag{69}$$

Constant u, v gives sum pooling; a terminal delta gives last-position pooling. Mean pooling divides the sums by sequence lengths and absorbs the corresponding $L_a L_i$ factor into the fitted scalar coefficient. This derivation motivates the forward feature family but also makes its approximation explicit.

Probe-span reduction. For a fixed set of A probes, only the action of the body interaction operator on the span generated by probe-side backward vectors can affect the measured probe scores. If Π_Ψ projects onto this span, then $\psi_a = \Pi_\Psi \psi_a$ and

$$\psi_a^\top M \xi_i = \psi_a^\top \Pi_\Psi M \xi_i. \tag{70}$$

Thus an A-dimensional probe span is a principled candidate low-rank subspace for tightening a residual bound; its usefulness is evaluated against exact body labels.

Body contribution: diagnostic result. On untied checkpoints we decompose

$$s_{i,a} = s^H_{i,a} + s^B_{i,a} \tag{71}$$

and report $r_{i,a} = |s^B_{i,a}| / (|s^H_{i,a}| + \epsilon)$, together with the change in top-tail ranking when s^B is included. On tied checkpoints, the analogous non-direct residual contains both tied-input cross paths and remaining body contributions, so it is not labeled a pure body residual. On untied checkpoints, head-only and full first-order harmful tails agree at $J_5^{\mathrm{body}} = .927$; the accompanying magnitude ratio has median $r = .24$, confirming that the body contribution is nonzero while rarely changing the selected tail. Optimizer awareness remains part of the analytic score because it introduces no fitted calibration model, whereas the body closure is reported as an offline reference unless the realized ranking audit shows a materially larger effect.

The head-vs-full Jaccard is high, so the reported implementation uses the direct-head score without a learned body closure; the probe-span reduction above remains an offline diagnostic.

Exact body diagnostic. The body audit uses the exact first-order quantity under the same fixed mask and optimizer state,

$$y^B_{i,a} = -\alpha_i \left\langle \nabla_B H_a, P_{t,B} g_{i,B} \right\rangle. \tag{72}$$

Offline, $\nabla_B H_a$ is computed once per probe/checkpoint and contracted with instance gradients or an equivalent directional derivative. Backward/JVP cost is therefore confined to the audit and never enters online CAGE. This exact audit supplies the head-vs-full quantities summarized above; the realized agreement is high ($J_5 = .927$), and all reported online runs use the forward-only direct-head score; the body-aware calculation remains an offline fidelity audit rather than an online component.

A.11 Finite-difference and exact-LOO audit hierarchy

We audit three targets:

1. **Differential target:** $s^{(1)}_{i,a} = -\alpha_i \langle \nabla H_a, P_t g_i \rangle$.
2. **Finite optimizer-local target:** recompute the conditional optimizer step with $G_t - \delta G_i$ while holding the mask and inherited optimizer state fixed.

3. **Finite training LOO target:** compare the complete post-step probe entropy with and without the instance under the fixed conditional definition.

The first tests algebra, the second tests optimizer linearization, and the third adds the entropy Taylor remainder. We report Spearman, sign agreement, relative error on a robust non-near-zero subset, and top-5% Jaccard at each transition.

Finite-difference noise determines the measurement ceiling. For a direction d, compare forward and central estimates

$$D_\epsilon^+ H = \frac{H(\theta + \epsilon d) - H(\theta)}{\epsilon}, \tag{73}$$

$$D_\epsilon^0 H = \frac{H(\theta + \epsilon d) - H(\theta - \epsilon d)}{2\epsilon} \tag{74}$$

across at least two ϵ values. If these estimators disagree materially, their mutual agreement is reported as the measurement ceiling and no method is claimed to exceed it.

The canonical local-fidelity values for this diagnostic setting are reported once in Table 13; the finite-LOO hierarchy above defines their targets.

A.12 Intervention estimands and causal controls

The intervention stage tests whether selector specificity matters beyond merely adding regularization. The CAGE-D (targeted), random matched-mass, source-strength-matched, and anti-targeted selective-KL causal-control arms match regularized token mass; the published baselines retain their original intervention semantics under the common training recipe. The primary selector-specific intervention contrast is targeted versus random matched-mass in support retention. Anti-targeting is a directional negative control, and source-strength/mean-probe controls may recover substantial signal. Separately, the headline CAGE method is compared with GRPO and the strongest published-method baseline on the benchmark-balanced macro under the shared primary protocol; count-weighted aggregation is reported separately as sensitivity.

The discrete intervention uses a top-k mask. The continuous intervention uses a monotone weight $w_i = f(S_i)$ calibrated so that its expected regularization mass matches the discrete condition. The continuous condition tests whether improved score calibration matters when magnitude, rather than only rank, is consumed. We therefore cross intervention type with Euclidean versus optimizer-aware scoring in a 2×2 design and report fidelity and behavioral effects separately.

Part II: Estimator implementation and finite-counterfactual audits

Supports Section 7.5: this Part tests whether the deployed approximation preserves the harmful tail, covering complexity, online realization, sparse/probe estimation, and local-fidelity validation; Table 13 is the main entry point.

A.13 Complexity accounting

The online direct-head method requires no probe backward pass and no JVP. Policy rollouts already provide h_i, π_i, and r_i; reference-policy evaluation provides the KL quantities; frozen probe forwards provide h_a and sparse ϕ_a. The dominant additional score contraction after truncation is $O(ANKM_{\text{vocab}}d)$ for a naive implementation, but batching shares matrix operations across instances and probes. Dense $V \times d$ materialization per (a, i) pair is never performed.

We report wall-clock overhead decomposed into probe forward, sparse score contraction, selector construction, and any extra reference computation. Offline costs (full-vocabulary audits, exact body teachers, family generation, power simulation) are one-time per model and excluded from the reported training ratios. The primary $K = 16, M_{\text{vocab}} = 128$ configuration costs $1.055\times$ GRPO at 3B; neighboring K and M_{vocab} values are reported below as convergence diagnostics.

A.14 ALGORITHMIC REALIZATION WITHOUT PROBE BACKWARD PASSES

The online computation is organized so that every probe-dependent quantity is produced by forward evaluation. The following sequence is normative for the implementation.

1. Run the policy rollout batch and retain, for every scored token, the final hidden state h_i, sampled token y_i, policy probabilities needed for the sparse vocabulary set, and the RL scalar α_i.
2. Run the reference policy quantities required by the KL objective. These are already required by the baseline objective and are not attribution-specific backward computations.
3. Evaluate the frozen probe bank in a batched forward pass. For each probe retain h_a, H_a, and the top-M_{vocab} entries by $|\phi_{a,v}|$ for $\phi_a = \pi_a \odot (\log \pi_a + H_a)$, then union the sampled token y_i exactly as in Eq. (75).
4. Form G_H^{RL} from loss-gradient outer products $-\alpha_i r_i h_i^\top / N$. In practice this is accumulated only on vocabulary rows required by the sparse contraction plus any optimizer-state audit rows.
5. Read the inherited AdamW moments for those head rows, evaluate the exact unfloored differential in Eq. (43), and apply the stabilized scoring surrogate $P_t^{(\nu)}$ in Eq. (44).
6. Compute the provisional head score with $m = 0$, aggregate over probes using the frozen stratified weights, and obtain $m^{(0)}$.
7. Use Eq. (50) to add the selective-KL head gradient under $m^{(0)}$, update the local head operator, and recompute the head score. The deployed method uses the head-only score, $S_B = 0$; the offline body diagnostic is reported separately.
8. Produce $m^{(1)}$ for the discrete intervention or continuous weights for the continuous intervention. The training backward pass is then executed once on the final training loss, exactly as in ordinary RL training.

Attribution adds no backward pass beyond the training loss's own.

A.15 ANALYSIS PROTOCOL

The analysis protocol fixes the data partitions, symbolic canonicalizer, probe strata, M-sensitivity grid, primary 25-step horizon, exposure construction, optimizer-state conditioning, matched-mass and source-strength-matched intervention controls, seed pairing, and inference procedures. The prospective analysis reports both R^{mass} and correctness-conditioned R^{cond} outcomes under the same fixed-effects and bootstrap specification. Family discovery, baseline-mass estimation, and evaluation use independent generation streams. The family budget is selected from late-checkpoint replicate stability, and unassigned correct-trajectory rates are reported by arm and checkpoint. Panel inference uses problem and seed-specific time fixed effects with wild cluster bootstrap inference and seed-wise coefficient sensitivity. Hyperparameters shared across methods are fixed globally, while method-specific regularization strengths receive matched development budgets.

A.16 SPARSE VOCABULARY CONTRACTION AND ERROR AUDIT

For each token–probe pair, the deployed support is fixed before scoring as

$$\mathcal{T}_{M_{\mathrm{vocab}},i,a} = \mathrm{Top}_{M_{\mathrm{vocab}}}(|\phi_a|) \cup \{y_i\}. \tag{75}$$

For the Euclidean head score, define $c_v = \phi_{a,v} r_{i,v}$; for a row-block metric the contribution is $c_v \kappa_v^P$ with $\kappa_v^P = h_a^\top D_v h_i$. The audit uses exactly the deployed support from Eq. (75). Two retained-mass diagnostics are reported:

$$R^{(c)}_{M_{\mathrm{vocab}}} = \frac{\sum_{v \in \mathcal{T}_{M_{\mathrm{vocab}},i,a}} |c_v|}{\sum_v |c_v|}, \tag{76}$$

$$R^{(s)}_{M_{\mathrm{vocab}}} = \frac{\sum_{v \in \mathcal{T}_{M_{\mathrm{vocab}},i,a}} |c_v \kappa_v^P|}{\sum_v |c_v \kappa_v^P|}. \tag{77}$$

The second is the key quantity for optimizer-aware scoring because a large metric weight can amplify a token with small $|c_v|$. We additionally report effective vocabulary support

$$N_{\text{eff}}^{v} = \frac{(\sum_v |c_v \kappa_v^P|)^2}{\sum_v |c_v \kappa_v^P|^2} \tag{78}$$

and a rare-row diagnostic that correlates top contribution rank with a row-wise summary of $\hat{v}$. We report $R^{(s)}_{M_{\text{vocab}}}$ together with the contribution-rank/second-moment diagnostic to show whether the sparse contraction is concentrated on common or near-zero-second-moment rows.

Table 9: Sparse-contraction audit.

M_{vocab}	$R^{(c)}_{M_{\text{vocab}}}$	$R^{(s)}_{M_{\text{vocab}}}$	Spearman	J_5^{trunc}
32	.934	.906	.941	.731
64	.968	.951	.967	.792
128	.986	.978	.982	.842
256	.995	.991	.992	.901

We use $M_{\text{vocab}} = 128$ in the primary configuration. Table 9 reports retained contribution mass and ranking agreement across neighboring M_{vocab} values; at $M_{\text{vocab}} = 128$, $R^{(s)}_{M_{\text{vocab}}} = .978$ and top-5% Jaccard is .842.

A.17 Probe population, stratification, and concentration

Estimator, Bernstein bound, and allocation. For strata $q = 1, \ldots, Q$ with weights w_q, independent samples X_{iqj}, means μ_{iq}, and variances σ_{iq}^2,

$$S_i^\star = \sum_q w_q \mu_{iq}, \qquad \widehat{S}_i = \sum_q \frac{w_q}{n_q} \sum_{j=1}^{n_q} X_{iqj}, \tag{79}$$

so $\mathbb{E}\widehat{S}_i = S_i^\star$ and

$$\text{Var}(\widehat{S}_i) = \sum_q \frac{w_q^2 \sigma_{iq}^2}{n_q}. \tag{80}$$

If $|X_{iqj} - \mu_{iq}| \le M_{iq}$, a weighted Bernstein bound gives

$$\Pr\{|\widehat{S}_i - S_i^\star| \ge \epsilon\} \le 2 \exp\left[-\frac{\epsilon^2/2}{\sum_q w_q^2 \sigma_{iq}^2 / n_q + \epsilon \max_q (w_q M_{iq}/n_q)/3} \right]. \tag{81}$$

A worst-case Hoeffding bound does not improve merely from proportional stratified allocation; the gain from stratification comes through reduced within-stratum variance. We therefore retain the variance-sensitive Bernstein control above when motivating allocation. For fixed $K = \sum_q n_q$, Neyman allocation is $n_q \propto w_q \sigma_{iq}$; because σ_{iq} varies by token, a pilot bank estimates a global stratum scale $\bar{\sigma}_q$ and fixes $n_q \propto w_q \bar{\sigma}_q$ before primary scoring. The Bernstein expression is used as a theoretical concentration control; the empirical rank/Jaccard behavior across probe counts is reported below.

The reference bank contains up to 64 problems per topic-position cell offline; logits are not retained densely. We immediately store the hidden state, entropy-side top-M_{vocab} vocabulary entries, and the metadata needed to reconstruct the score. We use $K = 16$ in the primary configuration. Against the frozen $K = 64$ reference, it yields $\rho = .956$ and $J_5^{\text{probe}} = .827$; Table 10 reports the convergence pattern across neighboring probe counts.

A.18 Diagnostic grouping selection

Candidate grouping families are topic, frozen reasoning position, response-length bin, and topic crossed with a two-bin (early/late) frozen position. Grouping diagnostics use trajectories disjoint

Table 10: Probe convergence against the frozen $K = 64$ reference. Jaccard is computed on the harmful top-5% set.

Online K	Spearman ρ	J_5^{probe}	Probe SD / seed SD	Combined overhead
5	.872	.646	.71	1.027×
8	.914	.712	.53	1.036×
16	.956	.827	.34	1.055×
32	.978	.884	.24	1.086×

Table 11: Candidate exposure groupings on diagnostic checkpoints. Grouping is selected by simulated power under the fixed generation budget; "Topic×pos." gives $G = 10$.

Grouping	Group share	Time share	Residual $u_{g,t}$ share	Power at $\lvert\beta\rvert = .10$
Topic only ($G = 5$)	.17	.64	.19	.58
Frozen position ($G = 4$)	.10	.65	.25	.64
Response length ($G = 4$)	.08	.79	.13	.47
Topic×pos. ($G = 10$)	.16	.56	.28	**.83**

from the prospective panel. The descriptive variance shares and simulated powers from the diagnostic design are reported in Table 11.

The topic×coarse-position grouping leaves sufficient within-group, within-time variation for the prospective test.

A.19 Exact-counterfactual and local-fidelity audits

Local fidelity is audited at four finite-counterfactual scopes: a small tied diagnostic checkpoint for development, the tied 3B primary checkpoint for deployed-score fidelity, an untied-from-initialization 3B companion, and a naturally untied 7B checkpoint for optimizer-aware fidelity on closure-exact architectures. A fifth, artificial-untie 1.5B setting is a separate architecture-support audit and uses Euclidean finite differences only. The shared matrix is copied into separate input and output parameters with identical numerical values; the analytic score is computed once and compared with Euclidean finite differences before and after untying. At 4–6 checkpoints and five seeds, we sample eight rollout batches and at most 200 valid response positions per batch. For each selected instance/probe pair we compute: (i) exact finite-step leave-one-contribution $C_{i,a}$ by applying the full and removed updates to a copied model state; (ii) the full first-order gradient inner product offline; (iii) the analytic direct-head score (full first-order output-head block only on untied checkpoints); and (iv) the sparse online approximation. On tied checkpoints, (i) uses the full shared update while (iv) uses the deployed direct-gradient operator, so their comparison includes both terms in Eq. (37). We report Spearman, sign agreement, top-5% Jaccard, robust relative error, and calibration slope.

On the closure-exact untied-from-initialization 3B companion, matched-random scores 30.05 versus 32.45 for CAGE, a paired difference of +2.40 pp (95% CI $[+1.28, +3.52]$). Because splitting the tied embedding changes the trained parameterization, these absolute Pass@1 levels are companion-run quantities and are not compared numerically with the primary tied-model means.

Optimizer conditioning improves finite-counterfactual agreement without degrading the rare-row audit. The realized tied-3B Jaccard gain is $+.075$; in untied 7B, J_5 rises from .751 (Euclidean) to .846 (optimizer-aware), a gain of $+.095$. Together with the matched calibration experiment, these fidelity gains motivate retaining the optimizer-aware variant.

Part III: Behavioral endpoints and prospective inference

Supports Section 7.1: this Part defines retention and gives the full prospective panel, exposure construction, sensitivity, family-resolution audits, endpoint-specific uncertainty, and prospective power.

Table 12: Metric and ranking audits in the 1.5B diagnostic setting. J_5^{probe} compares $K = 16$ with the frozen reference bank; $J_{5,\mathrm{diag}}^{\mathrm{opt}}$ measures direct-score agreement with finite conditional LOO; J_5^{body} compares optimizer-aware head-only with full first-order ranking.

Audit	Reported	Interpretation
Family notation-invariance agreement	.96	diagnostic agreement
Correct-route family separation	.88	route separation
Independent base-replicate $R_x(0)$	$.986 \pm .028$	centered near 1
$J_5^{\mathrm{probe}}(K{=}16, K_{ref})$	.827	ranking agreement
Euclidean $J_{5,\mathrm{diag}}^{\mathrm{opt}}$ vs finite LOO	.713	diagnostic
Optimizer-aware $J_{5,\mathrm{diag}}^{\mathrm{opt}}$ vs finite LOO	.762	+.049 vs Euclidean
Untied body/head magnitude ratio	.24	magnitude diagnostic
Untied J_5^{body} (head vs full)	.927	head/full ranking agreement

Table 13: Local-fidelity audits.

Audit	Metric	Reported	Interpretation
Direct-head algebra	Spearman ρ	.997	algebra check
1.5B tied Euclidean direct	Spearman ρ	.861	architecture-limited
1.5B tied optimizer-aware direct	Spearman ρ	.902	higher than Euclidean
3B tied Euclidean direct	Spearman ρ / J_5	.864 / .716	deployed finite target
3B tied optimizer-aware direct	ρ / sign / J_5	.909 / .914 / .791	deployed finite target
3B untied Euclidean direct	Spearman ρ / J_5	.889 / .754	closure-exact companion
3B untied optimizer-aware direct	Spearman ρ / J_5	.946 / .861	closure-exact companion
7B untied Euclidean direct	Spearman ρ / J_5	.887 / .751	diagnostic
7B untied optimizer-aware direct	Spearman ρ / J_5	.938 / .846	closure-exact
Full first-order offline	Spearman ρ	.946	offline reference
$10\times$ smaller step	Spearman ρ	.968	higher local fidelity
Sparse $M_{\mathrm{vocab}} = 128$	$R_{128}^{(s)}$	.978	retained contribution mass
Sparse vs full score	J_5^{trunc}	.842	ranking agreement
Online $K{=}16$, $M_{\mathrm{vocab}}{=}128$	train multiplier	$1.055\times$	measured overhead

Table 14: Complete prospective comparator panel corresponding to Table 3. Coefficients use identical outcome rows, controls, and seed×group wild-bootstrap draws.

Specification	Focal predictor coeff.	95% CI	R^2
Entropy only	-.057	$[-.073, -.041]$	0.012
For-Value-style + entropy	-.020	$[-.031, -.009]$	0.014
Family-balanced For-Value + entropy	-.028	$[-.041, -.016]$	0.015
Euclidean cross-context + entropy	-.050	$[-.063, -.038]$	0.021
CAGE exposure only	-.113	$[-.128, -.096]$	0.047
CAGE + entropy	-.110	$[-.124, -.095]$	0.047
CAGE − Euclidean (paired Δ)	-.060	$[-.074, -.046]$	+.026
CAGE − family-balanced For-Value (paired Δ)	-.082	$[-.099, -.064]$	+.032

A.20 Correct-family retention: complete specification

The behavioral metric is defined independently of training-method comparison. Each generated trace is first judged for final-answer correctness. The family pipeline has two explicit layers. At the trace level, we remove formatting-only tokens, parse displayed/inline equations and explicit equality or inequality statements into mathematical states when possible, and remove duplicate canonical states while preserving first occurrence. At the expression level, parsed states are canonicalized by exact rationalization where safe, commutative operand ordering for addition and multiplication, normalized equality orientation (including zero-form normalization), bounded symbolic simplification with polynomial expansion/factor normalization when semantics are preserved, cancellation of common rational factors, and normalization of equivalent exact constants. Strings that fail symbolic parsing

Table 15: Behavioral-support results on the family-resolved MATH500/AMC analysis set.

Method	R^{mass}	N_{eff}	Pass@64	Unassigned (%)
Base model (independent replicate)	.982	3.84	80.2	4.2
GRPO	.742	2.31	83.2	4.9
Uniform KL	.798	2.66	82.6	4.7
Clip-Higher	.775	2.50	81.4	4.9
Forking-20	.808	2.74	83.1	4.8
Clip-Cov	.810	2.76	83.3	4.6
STEER	.807	2.83	84.2	4.5
OPEFO	.809	2.86	84.4	4.6
KL-Cov	.804	2.71	83.0	4.6
Random matched-mass KL	.793	2.60	83.8	4.8
Source-strength matched random	.799	2.75	83.2	4.7
Anti-targeted KL	.756	2.08	77.5	5.3
CAGE-D	.830	2.84	84.9	4.5
CAGE	**.858**	**2.92**	**85.8**	4.3

Table 16: Correctness-conditional family retention on the primary family-resolved MATH500/AMC set. Family frequencies are normalized within correct generations, so this companion removes changes in total correctness probability.

Method	R^{cond}
Base model	.985
GRPO	.806
Clip-Higher	.814
Forking-20	.841
Clip-Cov	.847
STEER	.843
OPEFO	.846
KL-Cov	.844
Random matched-mass KL	.832
Source-strength matched random	.825
Anti-targeted KL	.808
CAGE-D	.854
CAGE	.882

are retained only as auxiliary lexical features and cannot alone define a family. No learned embedding is used.

For trace τ, let $\Sigma(\tau)$ be its canonical intermediate-state set and

$$d_J(\tau, \tau') = 1 - \frac{|\Sigma(\tau) \cap \Sigma(\tau')|}{|\Sigma(\tau) \cup \Sigma(\tau')|}. \tag{82}$$

Because threshold proximity is not transitive, discovery uses deterministic complete-linkage clustering: cluster distance is the maximum pairwise d_J; the closest pair is merged while its distance is $\leq \tau_{\mathrm{fam}}$; exact ties are broken lexicographically by serialized canonical signatures. Family IDs are then frozen. For a later correct trace τ and frozen discovery family $F_{x,j}$, define $D_j(\tau) = \max_{\tau' \in F_{x,j}} d_J(\tau, \tau')$. The trace is assigned to the qualifying family with minimum D_j (ties by frozen ID); if none qualifies, it is labeled novel. Novel traces are clustered separately by the same rule for N_{eff} but never enter frozen-base retention. Only canonicalization/assignment failures are unassigned, and reference/evaluation traces never alter discovery families. Both τ_{fam} and δ_{disc} are selected using only base-model traces. Invariance pairs created by notation-only transformations must remain in the same family; correct-route negative controls must separate; and repeated base sampling must produce a nondegenerate family-frequency distribution rather than singleton explosion. The ordered-overlap distance $d_{\mathrm{alt}} = 0.75 d_J + 0.25 d_{\mathrm{ord}}$ is used only for robustness.

For outcome problem x, baseline sampling is split into three independent streams. A discovery stream defines correctness-conditional family frequencies $\hat{\pi}^{\mathrm{disc}}_{x,j} = C^{\mathrm{disc}}_{x,j} / C^{\mathrm{disc}}_{x}$ and freezes

$$\mathcal{F}^{\mathrm{base}}_{x} = \{j : \hat{\pi}^{\mathrm{disc}}_{x,j} \geq \delta_{\mathrm{disc}}\}, \tag{83}$$

Table 17: Family-resolution audit. $\bar{F}$ is mean discovered base-family count; $P_{\leq 1}$ is the fraction of problems with at most one discovered correct family. $\mathrm{SD}_{\mathrm{late}}(R)$ is the SD between independent evaluation replicates at late diagnostic checkpoints (steps 250 and 500).

Benchmark	Metric	$M = 16$	32	64	128
MATH500	$\bar{F}$	1.46	2.12	2.74	2.91
	$P_{\leq 1}$	.55	.35	.22	.19
	$\mathrm{SD}_{\mathrm{late}}(R)$	.087	.058	.036	.027
AMC	$\bar{F}$	1.28	1.83	2.34	2.53
	$P_{\leq 1}$	.63	.46	.31	.27
	$\mathrm{SD}_{\mathrm{late}}(R)$	.098	.066	.039	.030
AIME24	$\bar{F}$	.18	.34	.61	.89
	$P_{\leq 1}$	.94	.89	.82	.75
AIME25	$\bar{F}$	.09	.18	.39	.63
	$P_{\leq 1}$	.97	.94	.88	.82

then is discarded for probability estimation. An independent reference stream estimates $p_{x,j}^{\mathrm{mass,ref}}(0) = C_{x,j}^{\mathrm{ref}}/M_{\mathrm{ref}}$ and $p_{x,j}^{\mathrm{cond,ref}}(0) = C_{x,j}^{\mathrm{ref}}/C_x^{\mathrm{ref}}$. At every behavioral checkpoint, including a separate $t = 0$ replicate, an evaluation stream estimates the analogous $\hat{p}_{x,j}^{\mathrm{mass}}(t)$ and $\hat{p}_{x,j}^{\mathrm{cond}}(t)$. The two retention estimands are

$$R_x^{\mathrm{mass}}(t) = \frac{\sum_{j \in \mathcal{F}_x^{\mathrm{base}}} \min\{\hat{p}_{x,j}^{\mathrm{mass}}(t), p_{x,j}^{\mathrm{mass,ref}}(0)\}}{\sum_{j \in \mathcal{F}_x^{\mathrm{base}}} p_{x,j}^{\mathrm{mass,ref}}(0)}, \tag{84}$$

$$R_x^{\mathrm{cond}}(t) = \frac{\sum_{j \in \mathcal{F}_x^{\mathrm{base}}} \min\{\hat{p}_{x,j}^{\mathrm{cond}}(t), p_{x,j}^{\mathrm{cond,ref}}(0)\}}{\sum_{j \in \mathcal{F}_x^{\mathrm{base}}} p_{x,j}^{\mathrm{cond,ref}}(0)}. \tag{85}$$

The discovery event and both baseline estimates are independent. A no-training sanity check requires both estimands at $t = 0$ to be centered near one. R^{mass} measures survival of base-supported correct probability mass; R^{cond} conditions away overall correctness and tests preservation of the family mixture among correct generations. A checkpoint with zero correct evaluation traces is missing for R^{cond} rather than assigned zero; family-resolved MATH500/AMC checkpoints almost never trigger this rule. Unless otherwise marked, R in the panel denotes R^{mass}.

R^{mass} and Pass@64 are distinct endpoints: the former measures how probability mass over base-supported solution families is retained, whereas the latter records whether at least one sampled solution is correct. Their method ordering therefore need not coincide; in particular, the anti-targeted arm has $R^{\mathrm{mass}} = .756$ while its Pass@64 remains 77.5.

Effective family count is

$$N_{\mathrm{eff},x}(t) = 1/\sum_{j \in \mathcal{F}_x^{\mathrm{obs}}(t)} [\tilde{p}_{x,j}(t)]^2, \qquad \tilde{p}_{x,j}(t) = \frac{C_{x,j}(t)}{\sum_{\ell \in \mathcal{F}_x^{\mathrm{obs}}(t)} C_{x,\ell}(t)}, \tag{86}$$

computed over all successfully assigned correct families observed at the checkpoint, including stable novel families. It is a threshold-free secondary summary rather than a replacement for validation of the frozen base-family definition.

Table 18: Family-metric validation audit.

Test	Observed value	Diagnostic role
Notation invariance agreement	.96	canonicalization stability
Correct-route separation	.88	family discriminability
MATH/AMC balanced $P(\lvert\mathcal{F}_x^{\mathrm{base}}\rvert \leq 1)$, $M = 64$	.27	family resolution
MATH/AMC balanced families/problem ($M^\star = 64$)	2.54	family multiplicity
Late-checkpoint replicate retention reliability	.91	replicate stability

Family-resolution adequacy is evaluated benchmark-by-benchmark using the M-sensitivity audit in Table 17. Varying τ_{fam} by ± 0.05 around its base-trace-selected value leaves the CAGE–KL-Cov

retention gap at $+.050/+.054/+.058$ and the CAGE–random matched-mass gap at $+.061/+.065/+.069$. The frozen central values are $\tau_{\text{fam}} = .35$ (Jaccard distance) and $\delta_{\text{disc}} = .05$. Replacing pure Jaccard with the ordered-overlap robustness distance leaves the CAGE–KL-Cov and discrete–random mass-retention gaps at $+.039$ and $+.037$, respectively.

A.21 Cross-fitted exposure construction

Held-out problems are partitioned into exposure groups $g = 1, \ldots, G$. The grouping variable is chosen to induce cross-context geometry: topic is primary; frozen reasoning position and response-length bin are candidate secondary axes. Difficulty and base correctness are not grouping variables and instead enter $Z_{x,t}$. Within every group, probe problems $\mathcal{P}_g$ and outcome problems $\mathcal{O}_g$ are disjoint.

At optimizer step u, define instantaneous exposure

$$a_{g,u} = \frac{1}{|\mathcal{P}_g|} \sum_{a \in \mathcal{P}_g} \frac{1}{|B_u|} \sum_{i \in B_u} \max\{\hat{s}_{i,a,u}, 0\}. \tag{87}$$

We normalize by valid-token count so that exposure does not mechanically scale with batch length. The primary prospective predictor aggregates the same preceding 25-step interval used by the entropy comparator,

$$A_{g,t}^{(25)} = \frac{1}{25} \sum_{u=t-24}^{t} a_{g,u}. \tag{88}$$

For interpretability, $A_{g,t}^{(25)}$ is standardized using calibration-run mean and variance only. The retention outcome is not standardized: the primary outcome is $Y_{x,t,25} = R_x(t+25) - R_x(t)$, so reported β values are retention-change units per SD of exposure. For the secondary $\Delta = 50$ outcome, the exposure window remains 25 steps; this fixes the predictor definition across horizons.

No outcome problem in $\mathcal{O}_g$ appears in its group's exposure. Behavioral checkpoints are stored at 25-step intervals; $\Delta = 25$ is primary and $\Delta = 50$ secondary, while a 10-step horizon is used only in dense diagnostic windows. Difficulty remains an outcome-side covariate rather than an exposure stratifier. Identification comes from relative group-by-time exposure variation within the same seed/checkpoint after problem and seed-specific time effects, rather than from global training drift.

For panel-design diagnostics, exposure variation is decomposed as

$$A_{g,t,s}^{(25)} = \bar{A} + \alpha_{g,s} + \lambda_{t,s} + u_{g,t,s}. \tag{89}$$

The residual $u_{g,t,s}$ is the variation that survives group-within-seed and seed-specific time effects and therefore identifies β from relative exposure differences within the same seed and checkpoint. The empirical covariance of $u_{g,t}$ enters the power simulation in Appendix A.23; for the $G = 10$ topic×position grouping, the simulated power for standardized $|\beta| = .10$ is .83.

A.22 Panel estimator and inference

For the regularly sampled panel, $\Delta = 25$ is primary and $\Delta = 50$ is secondary. A $\Delta = 10$ estimate is reported only inside dense diagnostic windows and is not pooled with the main panel. Retention is bounded in $[0, 1]$; we use the linear fixed-effects model as the primary estimand for interpretable changes. For $\Delta \in \{25, 50\}$,

$$Y_{x,t,\Delta,s} = \beta_\Delta A_{g(x),t,s}^{(25)} + \gamma^\top Z_{x,t,s} + \mu_x + \lambda_{t,s} + \epsilon_{x,t,s}. \tag{90}$$

$Z_{x,t}$ contains base accuracy, current accuracy, group advantage variance, response length, and current retention. Problem fixed effects absorb time-invariant problem difficulty, while seed-specific time fixed effects $\lambda_{t,s}$ absorb arbitrary common training drift separately within each seed. Because exposure and outcome noise are serially dependent within a seed–group trajectory, ordinary item-level or group-time-cell standard errors are invalid. Primary inference uses a wild cluster bootstrap over seed×group clusters (50 clusters for five seeds and $G = 10$) with Rademacher weights. We additionally report seed-specific $\hat{\beta}_s$, leave-one-seed-out pooled estimates, and sensitivity intervals

under alternative group-time and seed clustering. The primary directional hypothesis is $\beta_{25} < 0$. The 50-step horizon is secondary and is always reported regardless of sign; the 10-step estimate is reported separately only for the dense diagnostic windows.

Reported robustness checks include leave-one-group-out estimates, a future-exposure-to-past-retention lead placebo, and within-seed/checkpoint group-label shuffles.

Primary prospective estimate and robustness. The primary prospective estimand is the 25-step exposure coefficient β_{25} under the fixed-effects specification above, with the directional hypothesis $\beta_{25} < 0$. The observed estimate is $\hat{\beta}_{25} = -.113$ with 95% wild-bootstrap CI $[-.128, -.096]$. All five seed-specific coefficients are negative but vary by more than a factor of two in magnitude (Table 21); the pooled attribution coefficient remains negative after adding the entropy comparator ($\beta_A^{\text{joint}} = -.110$). We report these quantities as complementary robustness evidence rather than collapsing them into a thresholded success rule. Entropy decline itself is a mechanical consequence of the entropy-influence construction; the non-trivial prospective test is whether exposure predicts later support loss incrementally beyond the entropy comparator.

Entropy and attribution comparators. For every group-time cell we also compute

$$E_{g,t}^{(25)} = \overline{H}_g(t-25) - \overline{H}_g(t), \tag{91}$$

using only the preceding interval, where $\overline{H}_g$ is the mean entropy of the frozen group probes. The three specifications $M_1/M_2/M_3$ in the main text are fit with identical outcome rows, controls, seed-specific time fixed effects $\lambda_{t,s}$, and bootstrap procedure. The reported coefficients are $\theta_H = -.057$ in M_1, $\beta_A = -.113$ in M_2, and in the joint model $\beta_A^{\text{joint}} = -.110$ versus $\theta_H^{\text{joint}} = -.005$. The seed-specific and leave-one-seed-out summaries below refer to the univariate M_2 coefficient β_A; β_A^{joint} is a distinct partial coefficient after adding entropy and is therefore not expected to equal the arithmetic mean of the seed-specific M_2 estimates. The joint model determines whether attribution supplies incremental predictive information beyond entropy. To test whether the result is specific to CAGE rather than generic last-layer attribution, we evaluate two score replacements on exactly the same panel rows. The Euclidean cross-context comparator sets the local operator to identity, $s_{i,a}^{\text{Euc}} = -\alpha_i \langle \nabla H_a, g_i \rangle$, retaining the same entropy target but removing AdamW conditioning. The For-Value-style comparator uses the frozen probe next token y_a and the last-layer prediction-error alignment $s_{i,a}^{\text{FV}} = \alpha_i (r_a^\top r_i)(h_a^\top h_i)$, oriented so positive values denote source–probe alignment; it adapts the forward-only last-layer construction of Deng et al. (2026) to this RL/probe setting. A stronger family-balanced variant first averages $s_{i,a}^{\text{FV}}$ over probes within each discovered base-supported family and then averages families equally, so common families cannot dominate merely by contributing more frozen traces. These family labels come only from the frozen probe-side discovery traces for the corresponding problem and remain disjoint from reference/evaluation outcome streams, so the comparator cannot use outcome-stream family information. Each score is aggregated over the same preceding 25-step window, standardized from the same calibration stream, and entered in place of $A^{(25)}$ with identical controls and wild-cluster inference. For paired attribution tests, each wild-bootstrap draw reuses the same seed×group Rademacher weights across CAGE and all comparator fits, and percentile intervals are formed from within-draw $\Delta\beta$. The endpoint roles are summarized in Appendix A.1.

Table 19: Prospective-prediction coefficients. Predictors are standardized, outcomes remain in retention-change units, and R^2 is incremental over fixed effects and controls.

Horizon	$\hat{\beta}_\Delta$	95% wild-bootstrap CI	Incremental R^2	LO-group range
25	-.113	[-.128,-.096]	0.047	[-.126,-.090]
50	-0.102	[-0.131,-0.074]	0.073	[-.106,-.096]

Correctness-conditioned companion. We repeat the identical $M_1/M_2/M_3$ design after replacing the outcome by $Y_{x,t,25,s}^{\text{cond}} = R_x^{\text{cond}}(t+25) - R_x^{\text{cond}}(t)$. The predictor, controls, fixed effects, and seed×group wild bootstrap are unchanged. This companion asks whether exposure predicts future redistribution away from base-supported correct families even after conditioning on the response

being correct. The exposure-only estimate is -0.048 ($[-0.064, -0.032]$), and the joint attribution coefficient is -0.047 ($[-0.063, -0.031]$); every leave-one-seed-out estimate remains negative (-0.057 to -0.037). At $\Delta = 50$, the exposure-only coefficient is -0.042 ($[-0.058, -0.027]$).

Table 20: Correctness-conditioned prospective panel at $\Delta = 25$, using the same rows, controls, and bootstrap as Table 3. Endpoint roles are in Appendix A.1.

Specification	Attribution coeff.	95% CI for attribution	R^2
Entropy only	-0.009	$[-0.014, -0.004]$	0.005
For-Value-style + entropy	-0.008	$[-0.012, -0.005]$	0.009
Family-balanced For-Value + entropy	-0.011	$[-0.016, -0.007]$	0.012
Euclidean cross-context + entropy	-0.019	$[-0.024, -0.014]$	0.019
CAGE exposure only	-0.048	$[-0.064, -0.032]$	0.046
CAGE + entropy	-0.047	$[-0.063, -0.031]$	0.047
CAGE − Euclidean (paired Δ)	-0.028	$[-0.035, -0.022]$	+0.028
CAGE − family-balanced For-Value (paired Δ)	-0.036	$[-0.043, -0.029]$	+0.035

Table 21: Seed-specific estimates for the primary $\Delta = 25$ exposure coefficient. These are descriptive seed-wise refits; the pooled fixed-effects coefficient is estimated directly on the full panel and is not their arithmetic mean.

Seed	1	2	3	4	5
$\hat{\beta}_{25,s}$	-.171	-.081	-.126	-.102	-.093

Table 22: Inference sensitivity for β_{25}. The primary interval uses seed×group wild clusters; alternatives are robustness checks. All point estimates, intervals, placebos, and the leave-one-seed-out range in this table are regenerated from the same row-level prospective panel.

Inference scheme	Point estimate	95% CI
Seed×group wild cluster (primary)	-.113	[-.128,-.096]
Group-time cell cluster	-.113	[-.126,-.100]
Seed-block bootstrap	-.113	[-.142,-.087]
Group-specific linear trends	-.112	[-.128,-.096]
Within-group detrended exposure	-.112	[-.128,-.096]
Future-exposure lead placebo	-.001	[-.013,+.012]
Within-checkpoint group-shuffle placebo	+.006	[-.009,+.021]
Leave-one-seed-out range	[-.123,-.099]	–

A.23 Power simulation and compute-aware design

A prospective power simulation determines the panel design after the family-resolution sweep selects $M^\star$ and the mechanism-benchmark set. We preserve the measured $G \times T$ exposure covariance, the specified number of outcome problems per group, observed retention variance, and horizon-induced missing cells. Synthetic outcomes are generated for standardized $\beta \in \{-.05, -.10, -.15, -.20\}$ and analyzed with the exact panel estimator and wild bootstrap. We jointly vary G, checkpoint count, and outcome samples subject to a generation budget. Increasing repeated traces for the same outcome problem reduces measurement noise but does not create new exposure cells; increasing G or T is therefore preferred once retention estimation is stable.

The panel budget is finalized only after the M-sensitivity audit. With $M^\star = 64$, $G = 10$, 21 evaluation checkpoints at 25-step spacing, and 12 outcome problems per group give 161,280 checkpoint-evaluation generations per GRPO seed, plus 15,360 one-time discovery/base-mass generations (176,640 total, excluding benchmark evaluation). The dense 21-checkpoint panel is used for the GRPO prospective analysis; intervention arms evaluate family retention at the base, mid-training, and final checkpoints rather than recreating the full panel. This concentrates generation budget on the prospective estimand while preserving matched behavioral endpoints for the causal comparison.

Table 23: Power analysis for the primary $\Delta = 25$ coefficient.

Standardized $\lvert\beta\rvert$	$G = 8, T = 12$	$G = 10, T = 21$	$G = 12, T = 21$
.05	.24	.38	.46
.10	.61	.83	.90
.15	.86	.96	.98
.20	.95	.99	.995

Table 24: Complete primary 3B end-to-end results over ten paired seeds (Pass@1, %; mean with across-seed SD in parentheses). Cost follows the common accounting in Appendix A.27; Mean-probe CAGE denotes the mean-probe control for the continuous arm. Benchmark-level endpoints are reported for every method in this table rather than reconstructed from the macro average.

Method	MATH500	AMC23	AIME24	AIME25	Avg.	R^{mass}	Cost
GRPO	65.00 (1.28)	53.00 (2.30)	6.67 (2.72)	9.33 (3.06)	33.50 (1.48)	.742	1.000×
Uniform KL	65.00 (0.68)	53.25 (3.74)	7.00 (2.46)	9.33 (2.63)	33.65 (1.27)	.798	1.008×
Clip-Higher	66.06 (1.35)	53.00 (2.30)	7.00 (3.67)	8.33 (3.24)	33.60 (1.87)	.775	1.002×
Forking-20	65.84 (1.41)	53.50 (2.42)	6.00 (3.06)	9.67 (2.92)	33.75 (1.15)	.808	1.006×
Clip-Cov	67.46 (1.46)	55.00 (2.04)	7.67 (3.53)	8.67 (2.81)	34.70 (1.27)	.810	1.010×
STEER	68.12 (1.30)	55.75 (2.06)	7.33 (2.63)	10.00 (3.14)	35.30 (1.50)	.807	1.007×
OPEFO	70.08 (1.18)	54.25 (1.21)	8.00 (2.33)	9.67 (2.92)	35.50 (1.17)	.809	1.004×
KL-Cov	67.42 (1.38)	55.25 (2.19)	7.00 (2.92)	9.33 (3.44)	34.75 (1.24)	.804	1.012×
Random matched-mass KL	64.94 (2.26)	53.75 (2.43)	7.33 (2.11)	9.33 (3.44)	33.84 (1.48)	.793	1.015×
Source-strength matched random	65.26 (1.31)	53.00 (1.97)	8.00 (2.81)	10.33 (3.31)	34.15 (0.98)	.799	1.019×
Anti-targeted	63.66 (1.31)	51.75 (2.65)	6.00 (3.06)	8.00 (3.22)	32.35 (1.67)	.756	1.045×
Mean-probe CAGE	68.68 (1.08)	55.75 (2.06)	8.33 (2.36)	10.67 (2.11)	35.86 (0.90)	.825	1.031×
CAGE-D	70.54 (1.30)	54.50 (1.97)	8.67 (3.22)	12.00 (2.81)	**36.43 (1.29)**	**.830**	1.045×
CAGE	73.06 (1.61)	57.50 (2.04)	9.00 (2.25)	10.67 (2.63)	**37.56 (1.28)**	**.858**	1.055×

The reported power calculation uses the same simulation code and diagnostic covariance described above.

PART IV: TRANSFER AND INTERVENTION ROBUSTNESS

Supports Sections 7.2–7.4: this Part contains baseline implementation audits, matched intervention controls, scale/family/domain transfer, calibration checks, and selector diagnostics.

A.24 FROZEN-PROBE DRIFT AND 7B TRANSFER DETAILS

The frozen-probe drift audit uses the same probe bank and the same finite-counterfactual construction at steps 100, 300, 500, 750, and 1000; the latter two checkpoints are a shadow extension beyond the 500-step primary training horizon. The representation statistic is mean cosine distance from the corresponding step-0 frozen-probe hidden state; fidelity is evaluated on the same harmful-tail definition used elsewhere. To keep the longitudinal trajectory on a fixed sample, Table 26 follows the same five training seeds at every checkpoint from 100 through 1000. These five seeds are a fixed subset of the ten primary seeds used consistently at every shadow checkpoint; consequently the 500-step shadow-subset means (GRPO/CAGE = 89.4/92.6) are reported separately from the ten-seed primary means (89.5/92.7). To test whether the motivating large-k reversal is merely a 500-step snapshot, the shadow checkpoints retain the same nested 512-generation evaluation for GRPO and CAGE. At step 1000, Pass@512 is 88.7 for GRPO and 92.2 for CAGE, versus the frozen base reference of 91.8. The corresponding two-level item/completion bootstrap contrasts are GRPO $-$ base -3.10 pp ($[-4.42, -1.78]$) and CAGE $-$ base +0.40 pp ($[-0.86, +1.67]$). The former interval remains below zero, whereas the latter includes zero. Thus the GRPO inversion remains visible beyond the primary horizon while the 1000-step data do not resolve a CAGE–base difference, consistent with describing CAGE as remaining in the base-model range rather than exceeding it. These shadow endpoints are descriptive durability checks, not additional primary efficacy tests. Because the score is deployed on the tied 3B model, the relevant question is whether off-policy drift erodes its finite-counterfactual ranking, not whether the hidden representation itself remains stationary.

The 7B transfer evaluates both continuous CAGE and CAGE-D and uses the same family canonicalizer and independent discovery/reference/evaluation streams as the primary analysis. Using 64 sampled base solutions per problem, the 7B base has mean base-family count 3.6 and $\Pr(|\mathcal{F}_x^{\text{base}}| \leq 1) = .08$ on MATH500/AMC, supporting the retention estimand; an independent base replicate gives $R = .981$. The transfer protocol uses retention as primary and Pass@1 as secondary, with ten paired seeds. The CAGE-D companion arm recovers 0.269 of the random-to-base retention headroom, versus 0.196 in the primary 3B block; these are cross-setting descriptive recovery fractions rather than a matched contrast. GRPO serves as the reference condition; OPEFO is added as the strongest published accuracy comparator from the primary study, alongside the four construction-matched selective-KL intervention arms. Using the 3B pilot paired retention SD of .020, the ten-seed scale replication has greater than .99 two-sided power for a +.040 targeted–random retention effect under the same power calculation. The construction-matched discrete arms (CAGE-D, random, anti-targeted, and Mean-probe CAGE-D) share paired initialization/data-order seeds. All four use the same selective-KL construction and matched regularized-token/KL-mass budget; only the score used to choose or weight tokens differs. KL-Cov is omitted; this block tests selector specificity.

Table 25: Qwen2.5-Math-7B transfer over ten paired seeds. Retention is computed on family-resolved MATH500/AMC and is not headroom-normalized; Avg. Pass@1 averages the four math benchmarks. Pairwise contrasts use unrounded seed-wise macro values, so subtraction of displayed means can differ by 0.01.

Method	R^{mass}	N_{eff}	Avg. Pass@1	Pass@64
Base model	.981	3.74	28.15 (0.70)	84.9
GRPO	.802	2.65	44.26 (1.85)	86.8
OPEFO	.846	2.76	45.63 (1.51)	86.1
Anti-targeted	.789	2.54	43.40 (1.20)	85.8
Random matched-mass	.825	2.79	45.15 (1.16)	87.6
Mean-probe CAGE-D	.842	2.91	45.85 (1.53)	88.1
CAGE-D	.867	3.06	46.99 (1.43)	89.0
CAGE	**.874**	**3.14**	**47.59 (1.35)**	**89.4**

Probe-refresh sensitivity. Refreshing probe hidden states without changing probe texts at steps 500/1000 yields $\rho = .912/.901$ and $J_5 = .801/.784$.

Table 26: Frozen-probe drift, score fidelity, and large-k durability on the tied 3B run. J_5 is harmful-top-5% finite-counterfactual agreement; Pass@512 follows the same fixed five-seed subset at every checkpoint. Frozen-base Pass@512 = 91.8 provides the reference for interpreting the trajectory; the text reports the ten-seed primary endpoint and 1000-step bootstrap contrasts against base.

Checkpoint	D_t^{probe}	Spearman ρ	harmful-top-5% J_5	GRPO P@512 (5-seed)	CAGE P@512 (5-seed)
100	.061	.919	.804	91.3	91.9
300	.119	.908	.788	90.4	92.4
500	.184	.909	.791	89.4	92.6
750	.225	.886	.758	89.0	92.5
1000	.262	.872	.740	88.7	92.2

A.25 Cross-family and code transfer details

The Llama-3.1-8B-Instruct transfer reuses DAPO-Math-17K, the optimizer schedule, probe construction, and five paired seeds. Every exposure quantity is recomputed in the Llama representation space. The six arms are GRPO, KL-Cov (Cui et al., 2025), OPEFO (Xu et al., 2026), random matched-mass KL, Mean-probe CAGE, and CAGE. The independent base audit gives mean family count 3.2, $\Pr(|\mathcal{F}_x^{\text{base}}| \leq 1) = .12$, and $R_x(0) = .978$. The final-checkpoint deployed-score audit gives $\rho = .929$ and harmful-top-5% $J_5 = .832$. Benchmark SDs are across the five training seeds. Treatment-effect intervals for macro Pass@1 and retention are two-sided paired-t intervals over the five seed-wise method differences (4 degrees of freedom); AIME item-bootstrap uncertainty is reported separately and never replaces seed uncertainty.

The code experiment uses Qwen2.5-Coder-7B and the executable TACO training split. Rewards are binary testcase outcomes. Held-out TACO supplies the prospective panel and code-family retention; HumanEval+ and MBPP+ provide external accuracy tests. The strategy signature combines an alpha-normalized Python AST with deterministic execution-profile counters on a fixed hidden-test suite. Identifier renaming, formatting changes, and equivalent syntactic rewrites preserve the signature in .97 of validation pairs, while algorithmically distinct accepted solutions are separated in .90. Using 64 sampled base solutions per problem, the code base has mean family count 4.1, $\Pr(|\mathcal{F}_x^{\text{base}}| \leq 1) = .06$, and independent-replicate retention .976. Its final-checkpoint deployed-score audit gives $\rho = .936$ and harmful-top-5% $J_5 = .846$. Benchmark SDs are across five paired training seeds, and accuracy/retention contrasts use two-sided paired-t intervals over seed-wise differences (4 degrees of freedom). OPEFO is included under the same transfer protocol so that cross-family and code claims are not benchmarked only against KL-Cov. For the code prospective panel, TACO problems are partitioned into four coarse algorithmic topic strata crossed with early/late frozen reasoning position ($G = 8$); the primary interval for β uses a wild cluster bootstrap over seed$\times$group clusters (40 clusters), with problem and seed-specific time fixed effects matching the math-panel specification. Entropy alone gives $\beta = -.017$ ($[-.021, -.012]$); adding CAGE exposure yields the prospective CAGE coefficient $-.061$ ($[-.068, -.053]$) and increases R^2 from .014 to .106.

Correctness-conditional retention across transfer cells. The same companion metric preserves the selector ordering without using total correctness probability: targeted–random is +.029 on Qwen2.5-Math-7B, CAGE–KL-Cov is +.024 on Llama-3.1-8B-Instruct, and CAGE–KL-Cov is +.025 on Qwen2.5-Coder-7B. These are secondary mechanism checks; the reported primary transfer endpoints remain those specified in each transfer block.

A.26 Published entropy/exploration baseline implementation and protocol audit

The primary comparison evaluates six published mechanisms under the same Qwen2.5-3B data/evaluation protocol, while retaining method-specific optimization semantics. Clip-Higher uses $\epsilon_{\text{low}} = 0.20$ and $\epsilon_{\text{high}} = 0.28$ following DAPO (Yu et al., 2025). The forking-token baseline keeps policy-gradient updates only for the top 20% highest-entropy tokens in each batch and normalizes the policy loss by the number of retained tokens (Wang et al., 2025). Clip-Cov uses the published high-covariance-band sampling-and-detach rule, whereas KL-Cov uses top-covariance selection with KL to the rollout/old policy (Cui et al., 2025). STEER uses entropy-change-magnitude reweighting (Hao et al., 2026); OPEFO uses strict on-policy adaptive entropy-flow balancing (Xu et al., 2026). All six published baselines use the same ten paired primary initialization/data-order seeds as the CAGE comparison wherever their update semantics permit pairing, and receive the same development budget for tunable method-specific strengths; no CAGE score or probe signal enters these baselines.

Published results provide scale and protocol context rather than a numerical target for our shared 3B evaluation. On Qwen2.5-7B, Cui et al. report seven-benchmark averages 38.6/38.8/40.4/40.6 for GRPO/Clip-Higher/Clip-Cov/KL-Cov, respectively, while the forking-token study reports near parity on Qwen3-8B and larger gains at 14B/32B. STEER and OPEFO likewise report gains under their own model, decoding, and optimization protocols (Hao et al., 2026; Xu et al., 2026). Conversely, the independent Qwen2.5-Math-7B study of Jin et al. (2026), also trained on DAPO-Math-17K, reports ID Avg@64 of 44.99 for GRPO ($N_{\text{update}} = 4$), 44.96 for Clip-Cov, 43.54 for KL-Cov, and 44.88 for Rand-Pos-Clip. This shows that selective entropy interventions need not yield a large accuracy gain under every protocol. Those Avg@64 values are not directly comparable to our macro Pass@1 endpoint, so we use them as evidence of protocol sensitivity rather than as an effect-size ceiling.

For reproducibility, we therefore audit each implementation by source code/recipe, method-defining hyperparameters, and deviations required by the common 3B setup, rather than reporting a normalized "fraction of source gain recovered." Tunable strengths receive the same development-trial budget before the ten paired primary seeds are evaluated. Table 4 reports the resulting absolute shared-protocol performance. This avoids implying that heterogeneous source-study gains should reproduce at a common ratio. Because KL-Cov preserves its source-defined old-policy anchor, selector-specific inference is drawn from construction-matched CAGE controls.

Table 27: Llama-3.1-8B-Instruct math transfer, five paired seeds. Parentheses denote seed SD. Pairwise contrasts use unrounded seed-wise macro values, so subtraction of displayed means can differ by 0.01. Displayed item-level means respect the corresponding five-seed count grids.

Method	MATH500	AMC23	AIME24	AIME25	Avg.	R^{mass}
GRPO	55.48 (0.69)	39.50 (2.74)	8.67 (3.80)	0.67 (1.49)	26.08	.751
KL-Cov	56.04 (0.52)	41.00 (2.85)	9.33 (4.35)	1.33 (2.98)	26.93	.799
OPEFO	56.48 (0.58)	42.00 (2.96)	10.67 (4.10)	1.33 (1.83)	27.62	.807
Random matched-mass	55.80 (0.37)	40.50 (3.26)	8.67 (3.80)	1.33 (2.98)	26.58	.786
Mean-probe CAGE	56.32 (0.39)	41.50 (3.79)	10.00 (5.27)	1.33 (1.83)	27.29	.798
CAGE-D	56.92 (0.55)	42.50 (3.26)	11.33 (3.80)	1.33 (1.83)	28.02	.823
CAGE	**57.28 (0.58)**	**43.00 (2.85)**	**11.33 (3.80)**	**2.00 (1.83)**	**28.40**	**.831**

Table 28: Qwen2.5-Coder-7B transfer, five paired seeds. Parentheses denote seed SD. TACO is in-domain; HumanEval+ and MBPP+ use EvalPlus v0.2 tests (164 and 378 items, respectively). R_{code} is strategy-family retention on held-out TACO. Pairwise contrasts use unrounded seed-wise macro values, so subtraction of displayed means can differ by 0.01. Displayed item-level means respect the corresponding five-seed count grids.

Method	TACO	HumanEval+	MBPP+	Avg.	R_{code}
GRPO	32.08 (0.58)	65.85 (1.06)	64.07 (0.78)	54.00	.729
KL-Cov	33.00 (0.37)	66.59 (1.00)	64.97 (0.71)	54.85	.781
OPEFO	33.48 (0.49)	66.95 (0.80)	65.13 (0.62)	55.19	.795
Random matched-mass	32.76 (0.48)	66.34 (0.27)	64.81 (0.53)	54.64	.776
Mean-probe CAGE	33.28 (0.61)	66.71 (0.70)	65.13 (0.39)	55.04	.790
CAGE-D	34.10 (0.41)	67.20 (0.74)	65.13 (0.53)	55.48	.814
CAGE	**34.50 (0.27)**	**67.44 (1.00)**	**65.13 (0.44)**	**55.69**	**.823**

Primary baseline outcomes are reported in Table 4; Appendix A.27 reports the corresponding intervention costs and retention values under the same evaluation protocol.

A.27 Intervention protocol and causal controls

For continuous CAGE, with $q_{i,t} = S_i^\star$, the relative positive-score weight is

$$w_{i,t}^{\text{raw}} = \text{clip}\left(\frac{\max(q_{i,t}, 0)}{\text{EMA}[\max(q_t, 0)] + \epsilon}, 0, w_{\max}\right), \tag{92}$$

before the scalar normalization below matches the mean applied KL weight to the discrete targeted arm.

For the source-strength-matched control, let $\mathcal{A}_t$ be the active probe samples and $T_{M_{\text{vocab}},i,a}$ the deployed sparse support. The exact forward-computable source-update magnitude is

$$u_i^{\text{src}} = \frac{1}{|\mathcal{A}_t|} \sum_{a \in \mathcal{A}_t} |\alpha_i| \left(\sum_{v \in T_{M_{\text{vocab}},i,a}} r_{i,v}^2 \, \|D_{t,v} h_i\|_2^2 \right)^{1/2}. \tag{93}$$

For the continuous arm, if $\bar{a}_t^{\text{raw}}$ and $\bar{a}_t^{\text{disc}}$ are EMA-window means of the logged applied KL weight, the normalization is

$$\gamma_t = \frac{\bar{a}_t^{\text{disc}}}{\bar{a}_t^{\text{raw}} + \epsilon_\gamma}, \qquad \mathcal{L}_{\text{cont}} = \mathcal{L}_{\text{RL}} + \lambda \gamma_t \frac{\sum_i w_{i,t}^{\text{raw}} D_{\text{KL}}(\pi_i \| \pi_{\text{ref},i})}{\sum_i w_{i,t}^{\text{raw}} + \varepsilon_w}. \tag{94}$$

All methods use the same RL batches, reward, and optimizer. The CAGE causal-control arms (targeted, random, source-strength-matched, and anti-targeted) additionally share the same fixed-reference-policy KL construction; published baselines retain their source-method semantics as specified in Appendix A.26. For discrete causal controls, each batch uses the common $k_{\text{eff}} = \min\{\lfloor k N_{\text{elig}} \rfloor, n_+, n_-\}$, so targeted, random, and anti-targeted select exactly the same number of valid response tokens. After selection, an arm-specific scalar normalization matches the realized mean

selective KL penalty to a common batch target defined by the mean tokenwise KL over all eligible tokens; this uses only already-computed KL values and does not change which tokens are selected. The uniform-KL baseline is not mass-matched to a 5% selector because it regularizes all valid tokens. Its coefficient is selected on held-out development data from a four-point grid, using development macro Pass@1 only; primary evaluation seeds are not used for this choice (Appendix A.30). Reported train cost is measured wall-clock per optimizer step after warm-up, including rollout, one batched $K = 16$ probe forward, sparse $M_{\mathrm{vocab}} = 128$ score contraction, reference-policy computation, and backward/update. Probe evaluation is executed as one batched forward on the same GPUs after rollout; we do not assume asynchronous overlap to obtain the reported ratio. Evaluation generation and checkpoint I/O are excluded from every method. For continuous weighting, Eq. (92) supplies the relative token weights and the additional scalar γ_t in Eq. (94) normalizes their mean applied KL weight to match the discrete targeted variant over the same EMA window. This separates selector specificity from generic regularization while allowing a fair globally regularized baseline.

The source-strength-matched random control uses Eq. (93). Within each batch it partitions eligible tokens into deciles of u_i^{src}, copies the exact decile-count histogram of the CAGE-D selected set, and samples uniformly within each decile. It therefore preserves source-update magnitude at the resolution used for matching while discarding CAGE's within-bin probe-conditioned ordering. Matching is not treated as an exact scalar equality: across seeds the source-matched arm selects 4.91–5.18% of eligible tokens and its realized normalized KL ratio ranges from 0.981 to 1.012.

The matching audit gives mean selected-token fractions of 5.04%/4.98%/5.07%/5.02% for targeted/random/source-strength-matched/anti and normalized applied KL penalties of 1.002/1.005/0.996/1.001, respectively (targeted $= 1$). The three secondary specificity controls use the same k_{eff} rule and batch KL target: Mean-probe CAGE-D, Full-A, and Mean-probe-A select 5.08%, 5.02%, and 5.06% of eligible tokens, with normalized applied KL penalties 0.995, 1.007, and 0.997. All three are run on the same 10 paired primary seeds. For the continuous specificity check, the same mean-probe score replaces the individualized probe average inside Eq. (92); $w_{\max}$ and the scalar γ_t normalization are unchanged, so this control removes individualized probe information without changing the continuous intervention family.

The anti-targeted set contains the most negative eligible scores rather than merely low-magnitude tokens. The principal selector-specific contrast compares targeted with random matched-mass in R^{mass}, with R^{cond} reported as a correctness-conditioned companion; Pass@1 is reported as a co-outcome and is not a matching variable. A separate mechanism control compares CAGE-D with source-strength-matched random in R^{cond}. Source matching slightly reverses the two random-control arms under R^{cond} (.825 vs. .832), despite improving R^{mass} (.799 vs. .793). Source-strength matching preserves the selected update-magnitude distribution, not the mixture of base-supported families conditional on correctness, so the two retention summaries need not induce the same ordering. Random $>$ anti-targeted remains the directional negative control. Because tied-input contamination can be asymmetric across tails, we report $J^{\mathrm{tie}}_{5,+}$ and $J^{\mathrm{tie}}_{5,-}$ alongside this contrast.

Table 29: Secondary intervention controls not duplicated in the main end-to-end table. Primary methods and matched random/source-matched controls are reported canonically in Table 4. Anti-targeted macro Pass@1 is evaluated on the same primary benchmark suite and paired seeds as the other intervention controls; its separately sampled Pass@64–512 diagnostic is reported in Appendix A.32.

Method	Macro Pass@1	R^{mass}	Relative train cost
Mean-probe CAGE-D	35.24	.808	1.021×
Full-A	35.64	.817	1.045×
Mean-probe-A	34.84	.798	1.021×
Anti-targeted	32.35	.756	1.045×

A.28 Optimizer awareness by intervention form

The theory predicts that a better-calibrated magnitude should matter more when the intervention consumes magnitude. The 2×2 design isolates score geometry (Euclidean versus optimizer-aware) and control form (discrete versus continuous).

Table 30: Measured end-to-end training wall-clock relative to GRPO for the intervention actually evaluated in each scale study. Ratios include rollout, reference computation, the batched probe forward, sparse scoring, backward, and update; evaluation/checkpoint I/O is excluded.

Setting	Reported intervention / GRPO	Probe/scoring configuration
Qwen2.5-3B (CAGE)	1.055×	$K = 16,\ M_{\text{vocab}} = 128$
Qwen2.5-Math-7B (CAGE-D)	1.049×	$K = 16,\ M_{\text{vocab}} = 128$
Qwen2.5-Math-7B (CAGE)	1.058×	$K = 16,\ M_{\text{vocab}} = 128$

This audit is anchored to the primary intervention endpoints rather than to a second nominal replication of the same recipe: the optimizer-aware discrete and continuous cells are exactly the primary CAGE-D and CAGE seed-wise outcomes, and only the Euclidean-score counterparts are rerun under otherwise matched recipes. This removes an avoidable absolute-level discrepancy between nominally identical recipes while preserving the intended paired geometry-by-intervention test. Inference is based on seed-wise contrasts and the difference-in-differences; the interaction, not merely the optimizer-aware main effect, is the key test. The realized difference-in-differences is +0.77 Pass@1 points (95% CI $[+0.25, +1.29]$). Its positive sign and interval bounded away from zero support the hypothesis that optimizer-aware calibration matters more when score magnitude is consumed in the tested discrete-versus-continuous controls. The main specificity arms are descriptive rather than an additional interaction test: continuous weighting improves both mean-probe and full constructions in aggregate, so we do not attribute the full-versus-mean-probe contrast to weighting form itself. The paired calibration audit above is the inferential test of that interaction.

Table 31: Paired 2×2 optimizer-calibration intervention audit. The optimizer-aware cells are the primary CAGE-D and CAGE endpoints on the same ten seeds; their Euclidean-score counterparts rerun the otherwise matched discrete/continuous recipes. The interaction interval is a paired-t interval over seed-wise difference-in-differences. Finite-counterfactual score fidelity is reported separately in Table 13.

Score / intervention	Avg. Pass@1	Retention
Euclidean / discrete	35.54	.812
Optimizer-aware / discrete	36.43	.830
Euclidean / continuous	35.90	.829
Optimizer-aware / continuous	**37.56**	**.858**

A.29 Common-direction, anisotropy, and selector-specificity audits

Table 32 summarizes the canonical control taxonomy used by the main intervention analysis.

Table 32: Canonical control taxonomy for the main intervention analysis. "Preserves" states the matched construction or budget; "Changes" isolates the tested design choice.

Control	Label	Preserves	Changes	Alternative explanation tested
Random matched-mass KL	Random	Intervention form + token/KL budget	Token ranking	Generic selective regularization
Source-strength matched random	Source	Budget + source-strength bins	Within-bin ranking	Generic source-update magnitude
Anti-targeting	Anti	Intervention form + budget	Score direction	Directionality
Mean-probe CAGE-D	MP-D	Discrete intervention	Probe aggregation	Coarse cross-context structure
Mean-probe CAGE	MP-C	Continuous intervention	Probe aggregation	Coarse cross-context structure
Full-$\lvert s\rvert$	Abs	Full-score construction	Sign	Sign specificity
Mean-probe-$\lvert s\rvert$	MP-Abs	Mean-probe construction	Sign	Sign specificity

Let μ be the mean frozen-probe hidden state on the calibration bank and write $h_a = \mu + \widetilde{h}_a$. For a row-block operator with vocabulary blocks D_v, the deployed direct-head contraction has the exact decomposition

$$s^{\mathrm{dir}}_{i,a} = +\alpha_i \sum_v \phi_{a,v} r_{i,v}\, \mu^\top D_v h_i \;+\; \alpha_i \sum_v \phi_{a,v} r_{i,v}\, \widetilde{h}_a^\top D_v h_i \tag{95}$$

$$=: s^{\mathrm{common}}_{i,a} + s^{\mathrm{contextual}}_{i,a}. \tag{96}$$

The first term is independent of probe identity only through the hidden-state factor; ϕ_a remains probe-specific. Using $\sum |s^{\mathrm{common}}|/(\sum |s^{\mathrm{common}}| + \sum |s^{\mathrm{contextual}}|)$, the common direction carries .31 of absolute contribution mass and the contextual residual .69. A selector formed from the common component alone has top-5% Jaccard .36 with the full selector and rank correlation $\rho = .42$; its prospective coefficient is -0.002 ($[-0.006, +0.002]$). The full score's rank correlation with $\|h_i\|$ is only .21.

We also center the target hidden state before forming the hidden-state factor. The centered diagnostic retains top-5% Jaccard .78 with the deployed selector and yields prospective $\beta = -0.030$ ($[-0.038, -0.023]$). Together with the Mean-probe CAGE-D and Full-A controls, these results test common-direction and unsigned-magnitude alternatives.

The Mean-probe CAGE-D score used for the intervention control is $s^{\mathrm{mean}}_i = +\alpha_i \sum_v \bar{\phi}_v r_{i,v} \mu^\top D_v h_i$, where $\bar{\phi}$ and μ are fixed calibration-bank means; it therefore removes individual held-out-context identity. With the same uniform calibration-bank weighting used to define $\bar{\phi}$, linearity in ϕ_a gives the exact relation

$$s^{\mathrm{mean}}_i = \frac{1}{|\mathcal{P}|} \sum_{a \in \mathcal{P}} s^{\mathrm{common}}_{i,a}. \tag{97}$$

This identity does not make their prospective exposures interchangeable: the panel forms the positive tail before averaging within groups, whereas the mean-probe selector averages the probe-side score first and then applies the positive-tail selection. We therefore treat the mean-probe score itself as a distinct secondary panel predictor under the same rows, controls, and wild-bootstrap pipeline rather than inferring its predictive coefficient from β^{common}.

The Full-A arm keeps the full CAGE score but selects by $|s_i|$ under the same k_{eff} and realized-KL matching. Together with Mean-probe CAGE-D and the Mean-probe-A selector $|s^{\mathrm{mean}}_i|$, this completes a 2×2 specificity design: full versus mean-probe construction crossed with signed-positive-tail versus absolute selection. The paired seed-wise factorial contrasts are

$$\Delta_{\mathrm{probe}} = \tfrac{1}{2}[(Y_{F,S} - Y_{M,S}) + (Y_{F,A} - Y_{M,A})], \tag{98}$$

$$\Delta_{\mathrm{sign}} = \tfrac{1}{2}[(Y_{F,S} - Y_{F,A}) + (Y_{M,S} - Y_{M,A})], \tag{99}$$

$$\Delta_{\mathrm{int}} = (Y_{F,S} - Y_{F,A}) - (Y_{M,S} - Y_{M,A}), \tag{100}$$

where F/M denote full/mean-probe and S/A denote signed/absolute selection. The Mean-probe-A cell reaches 34.84 macro Pass@1, $R^{\mathrm{mass}} = .798$, and $R^{\mathrm{cond}} = .821$. For Pass@1, the paired factorial effects are $\Delta_{\mathrm{probe}} = +0.995$ ($[+0.504, +1.486]$), $\Delta_{\mathrm{sign}} = +0.595$ ($[-0.103, +1.293]$), and $\Delta_{\mathrm{int}} = +0.390$ ($[-0.655, +1.435]$). The analogous R^{mass} effects are +.021/+.012/+.003, with intervals $[+.007, +.035]$, $[-.002, +.026]$, and $[-.019, +.025]$; for R^{cond} they are +.020/+.013/-.006, with intervals $[+.006, +.034]$, $[-.001, +.027]$, and $[-.034, +.022]$. For Pass@1, the full-versus-mean-probe construction main-effect interval excludes zero, whereas the signed-versus-absolute main-effect interval includes zero; thus this factorial audit does not resolve a Pass@1 benefit of signed positive-tail selection over absolute selection. The retention results show the same qualitative asymmetry: the probe-information main-effect interval excludes zero while the signed-selection main-effect intervals include zero. The interaction intervals cross zero. In particular, $\Delta_{\mathrm{sign}} - \Delta_{\mathrm{probe}} = -.400$ with paired 95% CI $[-1.150, +.350]$, so the point estimate favors a larger full-vs-mean-probe construction effect, but its interval includes zero. Because mean-probe construction also changes the order of probe aggregation relative to positive-tail selection, Δ_{probe} is a full-vs-mean-probe construction effect; the separate mean-then-max sensitivity below quantifies aggregation-order sensitivity.

All specificity controls use the ten paired primary seeds with the same effective selected-token count and realized selective-KL target as CAGE-D. Mean-probe CAGE-D, Full-A, and Mean-probe-A select 5.08%, 5.02%, and 5.06% of eligible tokens, with KL ratios 0.995, 1.007, and 0.997. We report their contrasts with two-sided paired-t secondary-control intervals over the ten seed-wise differences

(9 degrees of freedom). Under the common wall-clock protocol, Mean-probe CAGE-D costs 1.021× GRPO, Full-A 1.045×, Mean-probe-A 1.021×, and Mean-probe CAGE 1.031×; the mean-probe variants save the online K-probe forward while retaining selector/reference-KL costs and, for the continuous arm, the continuous weighting/normalization operations.

Against Mean-probe CAGE-D, CAGE-D changes macro Pass@1 by +1.19 ($[+0.67, +1.71]$), R^{mass} by +.022 ($[+.006, +.038]$), and R^{cond} by +.017 ($[+.001, +.033]$). Against Full-A the corresponding differences are +0.79 ($[-0.22, +1.80]$), +.013 ($[+.001, +.025]$), and +.010 ($[-.003, +.023]$). The Pass@1 interval for CAGE-D minus Full-A includes zero, so its positive point estimate is descriptive and does not establish an accuracy advantage for signed over absolute selection. Full CAGE-D exceeds Mean-probe-A by +1.59 points ($[+0.59, +2.59]$), +.032 in R^{mass} ($[+.014, +.050]$), and +.033 in R^{cond} ($[+.014, +.052]$).

The mean-probe score gives $\beta_{\mathrm{mean}} = -0.008$ ($[-0.012, -0.004]$) and $R^2 = 0.021$; paired full-CAGE minus mean-probe differences are $\Delta\beta = -0.105$ ($[-0.110, -0.090]$) and $\Delta R^2 = +0.110$ ($[+0.088, +0.131]$). Averaging the signed score over probes and tokens before taking the positive part (mean-then-max) gives $\beta = -0.044$ ($[-0.057, -0.031]$) and $R^2 = 0.065$, a +0.069 shift from the primary coefficient—66% of the full-minus-mean-probe point-estimate gap. Thus simple aggregation retains substantial signal, but it does not account for the full cross-context association. Because this sensitivity changes aggregation order while retaining the underlying signed scores, the large prospective full–mean-probe gap cannot be assigned to probe identity alone. We therefore use the matched full-versus-mean-probe intervention to quantify the downstream incremental value of the full construction, rather than treating the panel gap as a causal effect.

The continuous specificity control uses the mean-probe score inside the same continuous weighting rule and retains the same $w_{\max}$ and γ_t normalization. It reaches 35.86 macro Pass@1, $R^{\mathrm{mass}} = .825$, and $R^{\mathrm{cond}} = .852$ at 1.031× GRPO. Full CAGE exceeds the mean-probe CAGE control by +1.70 ($[+1.20, +2.20]$), +.033 ($[+.015, +.051]$), and +.030 ($[+.014, +.046]$), respectively. The mean-probe construction retains substantial signal, but it does not recover the full continuous-arm effect: the paired full-minus-mean-probe intervals exclude zero for macro Pass@1, R^{mass}, and R^{cond}, with one seed-level reversal for macro Pass@1 and two reversals in each retention contrast. We therefore treat full-versus-mean-probe as secondary evidence for an incremental contribution from the individualized construction, not as a second primary claim or as evidence that individualized information explains the entire CAGE gain.

Compute–performance operating points. The matched controls occupy distinct cost–accuracy operating points rather than a single monotone frontier. Random matched-mass is (33.84, 1.015×), source-strength matched is (34.15, 1.019×), Mean-probe CAGE-D is (35.24, 1.021×), Mean-probe CAGE is (35.86, 1.031×), full CAGE-D is (36.43, 1.045×), and full CAGE is (37.56, 1.055×). In the mechanism-matched discrete path, relative to the random-to-full increment, Mean-probe CAGE-D uses about 0.20 of the added wall-clock overhead while recovering 0.54 of the Pass@1 gain. We report these controls for mechanism identification rather than claiming that each lies on the global Pareto frontier.

PART V: TRAINING AND STATISTICAL PROTOCOL

Supports all empirical claims: this Part records training configuration, tail diagnostics, benchmark-level reporting, inference, statistical reporting, and reproducibility details.

A.30 PRIMARY TRAINING CONFIGURATION

The following configuration is shared by the primary comparative runs unless a method-specific source recipe is explicitly stated.

Table 33: Training configuration. Shared backbone hyperparameters are common across comparative runs.

Item	Specified value
Primary RL data	DAPO-Math-17K
Effective prompt epochs	1.88 (32,000 prompt draws / 17,000 prompts)
Diagnostic data	MATH train / matched math subset
Prompt batch	64
Rollouts per prompt	8
Max response length	3072
Primary optimizer steps	500
Shadow-extension steps	750/1000 (five seeds; descriptive only)
Shadow data exposure	same DAPO-Math-17K pool reshuffled/repeated
Step-1000 exposure	64,000 prompt draws / 17,000 prompts (≈ 3.76 effective epochs)
Policy learning rate	constant 1×10^{-6} for steps 1–1000; no decay
Shadow continuation	same AdamW state and constant LR after step 500; no reset or retuning
Optimizer	AdamW
Primary seeds	10; paired-power analysis in Appendix A.32
Probe positions	0, .25, .50, .75 of frozen base trace
Exposure groups	$G = 10$, five topics $\times$ early/late frozen position
Behavioral checkpoints	every 25 steps (21 over 500 steps)
Primary horizon	25 steps
Secondary horizon	50 steps; 10 only in dense diagnostic windows
Outcome samples	32 diagnostic; 64 per outcome/checkpoint in primary panel
Online probe K	16
Sparse vocabulary M_{vocab}	128
Family clustering	Jaccard distance $\tau_{\text{fam}} = .35$; $\delta_{\text{disc}} = .05$
Selector / KL	$k = 5\%$, $\lambda = 0.05$
Continuous cap	$w_{\max} = 3.0$
Adam denominator floor	$\nu = 10^{-12}$

Method-specific deviations are limited to the intervention semantics documented in Appendix A.26; prompt draws, rollout count, response cap, optimizer, checkpoint grid, evaluation sampling, and seed counts are otherwise held fixed within each stated comparison.

Uniform-KL coefficient sweep. We tune the global-KL coefficient on held-out development data over $\lambda \in \{0.01, 0.025, 0.05, 0.10\}$. The selection rule is development macro Pass@1; R^{mass} is shown only to expose the accuracy–retention tradeoff and is not used to select λ. The development sweep has a shallow interior optimum: macro Pass@1 rises from $\lambda = 0.01$ through 0.05 and then falls at 0.10, while retained family mass increases monotonically. We therefore carry $\lambda = 0.05$ into the primary evaluation without using primary outcomes for coefficient selection.

Table 34: Uniform-KL development sweep. Development summaries are used only for coefficient selection; the primary ten-seed outcomes in Tables 4 and 5 are not used in the sweep.

λ	Dev. macro Pass@1	Dev. R^{mass}
0.01	33.28	.762
0.025	33.54	.781
0.05	**33.72**	.800
0.10	33.39	.816

A.31 Tail enrichment of the prospective signal

Because the intervention consumes only the harmful tail, we complement global panel R^2 with a concentration curve rather than a single cutoff. Top-1/5/10% cross-fitted exposure cells show 3.18×/2.96×/2.78× the future support-loss rate of the overall prospective cell pool, with cluster-bootstrap 95% percentile CIs [2.72, 4.01], [2.68, 3.39], and [2.48, 3.14]. In each bootstrap draw the denominator is recomputed from the full prospective pool, so its sampling variance is stabilized by all

cells; the wider top-1% interval still reflects the much smaller numerator subset. The corresponding regions carry 0.067/0.250/0.426 of positive exposure mass.

A.32 Benchmark evaluation and statistical reporting

Primary benchmark values use a fixed design of 10 paired seeds. For every primary arm, benchmark accuracy is first computed from the integer number of correct items within each seed (500 MATH500, 40 AMC23, and 30 in each AIME set); the displayed mean and sample SD are then computed across those same ten seed-level accuracies. Thus the AIME seed values respect the native $1/30$ item granularity rather than being generated from a continuous approximation. Two-sided paired-t intervals over seed-wise differences use nine degrees of freedom. Pairwise contrasts use unrounded seed-wise macro values, so differences of displayed rounded means can differ by 0.01. AIME additionally receives within-seed item bootstrap intervals; probe, seed, family-sampling, and panel uncertainty are reported separately. The Pass@k and N_{eff} diagnostics are separately sampled distributional endpoints rather than algebraic transforms of Pass@1 or R^{mass}; changes in the primary endpoint therefore do not mechanically imply changes in those diagnostics. The extended high-k audit uses $n = 512$ independently sampled completions per benchmark item and computes Pass@k for $k = 64, 128, 256, 512$ from that single nested pool using the standard finite-sample Pass@k estimator. Trained arms are computed within each of the 10 training seeds and then averaged; the base model uses 10 independent sampling replicates under the same decoding settings. Seed and item uncertainty are reported separately. Direct base–GRPO and CAGE–base high-k contrasts use a two-level percentile bootstrap over items/completions and outer training-seed or base-replicate units; these intervals may be asymmetric. Pass@512 is the primary high-k endpoint and Pass@256 is descriptive (Appendix A.1). The strongest published-method baseline is the maximum over six external methods, so intervals for that selected comparison are Bonferroni-adjusted across all six paired comparisons for each endpoint; matched-control contrasts use two-sided paired-t intervals over the ten seed-wise differences. We report effect sizes and confidence intervals directly for GRPO, the strongest published-method accuracy baseline, and the strongest published-method retention baseline rather than converting these estimates into thresholded pass/fail decisions, preserving the reported magnitude and uncertainty for each comparison.

Table 35: Extended high-k coverage diagnostic on the primary 3B run. Values are Pass@k (%) from one nested pool of $n = 512$ generations per item.

Method	Pass@64	Pass@128	Pass@256	Pass@512
Base model	80.2	87.9	91.1	91.8
GRPO	83.2	88.0	89.4	89.5
Anti-targeted	77.5	84.2	86.6	87.0
Random matched-mass	83.8	89.4	91.2	91.5
CAGE-D	84.9	90.3	91.8	92.1
CAGE	**85.8**	**90.9**	**92.3**	**92.7**
Base − GRPO Δ	–	–	+1.73 [+0.70, +2.77]	+2.28 [+1.27, +3.32]
CAGE − base Δ	–	–	–	+0.90 [−0.24, +2.04]

Table 36: Additional paired contrasts not duplicated in the main compact contrast table. Accuracy differences are percentage points; retention differences are absolute fractions.

Contrast	Estimate	95% interval
CAGE − Clip-Cov, R^{cond}	+.035	[+.011, +.059]
CAGE-D − random, R^{cond}	+.022	[+.006, +.038]
CAGE-D − source-matched, R^{cond}	+.029	[+.004, +.054]

Paired-seed power. The primary study uses ten paired seeds for every method. To make the operating characteristics of the published-baseline comparison explicit without tying the design to a single effect-size threshold, we report two-sided paired-t power under the same Bonferroni

Table 37: Seed-level primary macro Pass@1 for the two focal paired contrasts. Publishing the paired observations makes the covariance structure directly auditable. The across-seed correlations are .836 for CAGE/OPEFO and .758 for CAGE-D/random; paired-difference SDs are 0.710 and 0.978 pp, respectively. Both comparisons use the same seed IDs, but no equality or ordering of correlations is assumed because each method maps a shared seed's prompt/order difficulty into a different trained policy.

Seed	OPEFO	CAGE	Random	CAGE-D
1	35.42	38.31	34.00	37.44
2	36.45	38.70	34.92	38.27
3	35.07	37.23	34.62	36.84
4	34.31	36.94	33.38	35.09
5	33.12	35.23	31.75	34.01
6	37.28	39.23	34.50	37.02
7	36.35	37.83	35.68	37.33
8	35.65	37.70	30.87	35.57
9	35.93	38.58	34.58	36.95
10	35.42	35.82	34.12	35.74

family-wise rule used for that comparison ($\alpha = .05/6$ across six external methods). At paired SD 1.0 percentage point, ten seeds provide approximately .47/.82/.99 power for true paired margins of 1.0/1.4/2.0 percentage points, respectively. These calculations describe the sensitivity of the fixed ten-seed design; inference in the paper is based on the observed paired effects and multiplicity-adjusted confidence intervals, not on a post hoc power threshold.

High-k uncertainty. The base–GRPO differences at $k = 256$ and the primary $k = 512$ endpoint, together with the direct CAGE–base comparison at $k = 512$, use two-level percentile-bootstrap 95% confidence intervals, resampling training seeds or base-model sampling replicates and evaluation items within the corresponding outer unit. The base–GRPO $k = 512$ contrast is +2.28 with interval $[+1.27, +3.32]$; CAGE–base is +0.90 with interval $[-0.24, +2.04]$. The latter interval includes zero, so it does not resolve a CAGE–base difference; it is not an equivalence test; $k = 256$ is descriptive.

Aggregation sensitivity. The primary macro contrast is unweighted across the four benchmarks. The corresponding count-weighted CAGE minus OPEFO contrast is +2.80 percentage points. Restricting the sensitivity to the two larger benchmarks gives +3.12 pp ($[+2.35, +3.88]$); the benchmark-specific paired contrasts are +2.98 pp ($[+2.321, +3.639]$) on MATH500 and +3.25 pp ($[+2.043, +4.457]$) on AMC23. AIME24 is +1.00 pp ($[-0.152, +2.152]$), while AIME25 is +1.00 pp ($[-0.609, +2.609]$). Because each AIME set has only 30 items, seed-level accuracy moves in $1/30$ increments and the resulting paired intervals are correspondingly coarse; the two AIME cells are reported separately and neither is used as standalone evidence for the headline effect.

A.33 Detailed limitations

Transfer. Qwen-7B uses ten paired seeds; Llama/code use five, and code accuracy versus OPEFO includes zero although retention remains positive. **Approximation and horizon.** The tied-3B headline uses a validated direct-head approximation; exact row-block closure covers the untied output path, but the 7B cells do not isolate an untying effect. Body effects are audited, and frozen probes become increasingly off-policy beyond the primary horizon. **Entropy target and probe resource.** Frozen correct-prefix entropy is an operational support-loss proxy, not reasoning support. CAGE also requires held-out correct probes unused directly by reproduced baselines, so matched controls isolate ranking given this resource, not the resource's independent value; probe/evaluation problems remain disjoint. **Endpoint and scope.** R^{mass} is base-referenced route coverage, not capability, requiring stable strategy equivalence; extension beyond auditable math/code families is unvalidated.