PDSA: Separate Policy and Memory State for Resource-Bounded Adaptation

Anonymous authors
Anonymous manuscript

Abstract

An adaptive agent can improve by retaining facts, changing its policy, or trying more answers. Predictive Dual-State Adaptation (PDSA) separates bounded memory from an independently restorable numerical policy and proposes action-conditioned future utility for admission. We study controlled arithmetic and list-program repair with a frozen shared semantic base, public execution feedback, and hidden audit cases excluded from updates and selection. Five studies examine shared semantic qualification, candidate/update budget, calibrated update strength, policy-carrier design, and feedback resolution. Larger candidate budgets with evolving public feedback improve selected answers equally in updating and frozen regimes; neither calibrated strength nor an internal carrier establishes its prespecified feedback-specific gain. In the final 32-problem comparison, dense feedback distinguishes every active candidate batch, drives 83 optimizer steps across problems (at most three per problem), and changes attained-prefix distributions and most candidate streams. Yet dense-learning audit loss is 0.703125, versus 0.7015625 for frozen selection, 0.653125 for exact-reward updates, and 0.73125 for dense cyclic mismatch; none of the registered dense-learning contrasts meets the materiality and uncertainty rule. Configured search reaches 0.028125. Subsequent post-hoc CPU scoring finds no missed audit-perfect solutions in the saved pools: all four methods retain only two fully solved problems under pool-best selection. Pool-best mean loss remains 0.609375–0.684375 on the same audit cases, locating substantial residual error in observed candidate coverage. Separately, admission improves over no-write but not a strong rule. These results distinguish feedback variability, policy reuse, candidate coverage and selection from incremental task benefit; learned predictive admission and full PDSA advantage remain unestablished.

1 Introduction

An agent can make progress by retaining an observation, learning from execution feedback, or trying more answers without changing its policy. These mechanisms can yield the same final score. To attribute improvement to within-problem learning, the comparison must separate additional attempts, informative feedback, changes in numerical state, actual later use of that state, and improvement beyond a non-updating method with the same information.

Predictive Dual-State Adaptation (PDSA) poses a broader question: can action-conditioned future utility guide selective memory admission while numerical policy updates improve later attempts on the same unfinished problem? It separates slow foundation parameters θ, a bounded numerical policy carrier Φ, and independently bounded authoritative records M. The Memory-Mediated Learning Architecture (MMLA) report supplies these state definitions and conditional contracts (Zou & Donz, 2026b). Its earlier θ/M studies (Zou & Donz, 2026a) concern distinct experimental objects; the two citations identify versions of the same work. This paper evaluates policy adaptation and single-event memory admission separately, and contributes controlled measurements of updates, functional reuse, candidate coverage and public selection in PDSA instances.

Our central empirical question is whether a working adaptation mechanism provides incremental task benefit. We study arithmetic and list-program repair with a frozen, shared semantic base, actual memory reads, public execution feedback, and audit cases withheld from selection and updates.

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.

Five studies vary shared semantic qualification, candidate/update budget, update strength, carrier design, and feedback resolution. Each has its own inputs and matched controls; their means are not successive measurements of one learning curve. We center the feedback intervention and complete saved-pool analysis, using the preceding studies to locate the capabilities they actually test.

The latest comparison makes feedback more discriminative without changing the carrier or candidate ceiling across its learning arms. Dense behavioral scores distinguish every active batch, including batches whose exact rewards are all zero. Updates affect attained later-prefix distributions and most candidate streams diverge. Nevertheless, dense-feedback learning does not meet its registered audit-benefit criterion against frozen selection, exact-reward learning, or a dense cyclic-mismatch control. Configured symbolic search remains substantially stronger.

The contribution is a set of diagnostic separations that can be reused when testing within-problem adaptation. Equal-candidate-budget frozen sampling separates extra attempts from updates. The qualified common base controls acquired grounding, while attained-prefix and reset interventions distinguish numerical change from functional use. The reward-resolution comparison establishes that nonconstant feedback, actual reuse and candidate divergence can coexist with no established incremental selected-answer benefit. Hidden audit evaluates that benefit independently of public selection. The cyclic control tests a specific reward mismatch, whose gradient interpretation we make explicit. Post-hoc scoring of the saved pools finds no missed audit-perfect solution: even pool-best selection leaves substantial error on the fixed audit cases. The model experiments through 21 September 2026, closed out on 22 September, and the subsequent 23 September CPU diagnosis support these separations.

2 Related work

Policy adaptation and external memory already coexist in metacognitive test-time reasoning (Li et al., 2025). Policy of Thoughts combines tree exploration, execution feedback and transient LoRA updates with GRPO within an instance (Jiao et al., 2026). TTT-Discover targets a single high-reward discovery using an entropic objective and PUCT-based reuse of prior solutions (Yuksekgonul et al., 2026). Our flat, bounded candidate pools use completion-level REINFORCE with a leave-one-out baseline, following the RLOO formulation studied for LLM alignment by Ahmadian et al. (2024). Public executable rewards replace the learned preference reward and KL-shaped alignment objective of that work. TTRL instead obtains test-time rewards from majority agreement (Zuo et al., 2025). We test selected-output increments over same-information frozen sampling, with separately restorable state and held-out exact audit, rather than reproduce these systems or compare their benchmark performance.

Multiple samples and functional tests underpin code-generation evaluation (Chen et al., 2021). Our pools instead adapt to feedback and may stop early, so their coverage is not iid pass@32. Proposal quality and verification also determine how test-time computation helps (Snell et al., 2024). Our decomposition measures the best saved candidate and the additional loss of the public selector, rather than a compute-optimal allocation rule. Self-Refine iteratively revises outputs using model-generated feedback without parameter updates (Madaan et al., 2023); our frozen comparator receives aggregate execution feedback, not a program–critique–revision history, and does not represent all nonparametric repair methods.

Downstream utility is established in memory learning. UMA scores branches of a query-agnostic memory state with mean QA reward (Zhang et al., 2026b); Memory-R1 learns operations including NOOP (Yan et al., 2026). Mem-α learns memory construction from downstream performance (Wang et al., 2025), and Mem-T uses tree-guided hindsight credit assignment (Yue et al., 2026). A-MAC combines future utility with observable admission features (Zhang et al., 2026a); AgeMem combines long- and short-term management (Yu et al., 2026); MemRL learns episodic utilities without weight updates (Zhang et al., 2026c).

Dual state, low-rank adaptation, reward learning, and factorial controls are existing tools. The PDSA target combines own-successor admission pricing, independent label/audit futures, and separately restorable policy and memory. Our contribution is a controlled diagnosis separating numerical updates, actual policy reuse, observed candidate coverage and public selection. The memory studies supply separately scoped component evidence for PDSA.

3 STATE, MODEL AND FEEDBACK INTERVENTIONS

3.1 SEPARATE LIFETIMES AND SHARED INFORMATION

A completed observation supplies candidate row c_t. A trusted assembler checks schema, namespace and version, then commits one legal row or exact NULL:

$$M_{t+1} = T(M_t, c_t, a_t), \qquad T(M_t, c_t, \mathrm{NULL}) = M_t.$$

A tombstone is a stored deletion record, not NULL. Eight 512-byte rows plus 40 bytes of metadata occupy 4,136 serialized bytes; renderings, snapshots, policy parameters and optimizer state are additional costs. Authority means the system recognizes the record, not that its contents are semantically true. Public query keys permit deterministic lookup, not learned language routing. Named facts are rendered losslessly from actual resident rows; missing facts are not supplied by the evaluator.

The common model is $\bar{\theta} = (\theta_0, \omega^*_{\mathrm{sem}})$: the pinned MiniCPM5-1B-SFT base (OpenBMB, 2026) and a shared, frozen offline LoRA adapter (Hu et al., 2021). Every neural arm keeps this adapter enabled. The initial online carrier is the independent rank-four output residual $z_{\bar{\theta},\Phi}(s) = z_{\bar{\theta}}(s) + BAh_{\bar{\theta}}(s)$. Here s is the attained token prefix and $h_{\bar{\theta}}(s)$ its last hidden vector. The internal alternative adds $2B_\Phi A_\Phi x$ in parallel to each complete q/v projection, including its frozen semantic adapter, at zero-based layers 16–23, with independent rank-16 factors. Only one online carrier is active. Comparing these designs changes position, rank and parameter distribution together, not one isolated architectural variable.

Online factors are FP32, with random nonzero input factors and zero output factors, empty AdamW state (Loshchilov & Hutter, 2019), zero weight decay, gradient-norm cap 1 and projection to $[-16, 16]$. The output studies use 10^{-4} or a publicly calibrated rate; the internal comparisons use 10^{-4}. At zero B, the first task gradient of A is zero while B can learn; later nonzero gradients are not guaranteed. Policy, optimizer, RNG, matcher and prompt state persist within an unresolved request and reset between independent branches and requests. Memory and the common base occupy separate restoration domains. Restoring a state does not roll back consumed computation.

3.2 THE SAME CONSTRAINED DISTRIBUTION FOR SAMPLING AND LEARNING

A public grammar restricts JSON structure, operation names by public input type, arity, integer ranges, operation count and EOS. It supplies neither correct order nor correct operands. With legal support $G(s)$ at prefix s,

$$p_G(v \mid s) = \frac{\mathbf{1}[v \in G(s)] \exp z_{\bar{\theta},\Phi,v}(s)}{\sum_{u \in G(s)} \exp z_{\bar{\theta},\Phi,u}(s)}.$$

Sampling uses temperature one, no top-k/top-p and no KV cache. Current-weight replay reconstructs the same support and EOS. Sampling and replay use the same native eager BF16 8×512 execution path; padding and dummy positions are charged (Table 38).

For K completions, reward r_k gives leave-one-out advantage $\hat{A}_k = K(r_k - \bar{r})/(K-1) = r_k - \sum_{i \neq k} r_i/(K-1)$ and objective

$$\mathcal{L}_\Phi = -\frac{1}{K} \sum_{k=1}^{K} \hat{A}_k \sum_{j=1}^{|y_k|} \log p_G(y_{k,j} \mid x, y_{k,<j}).$$

The objective averages completions and sums answer/EOS log probabilities; it does not normalize each completion by length. Internal updates accumulate eight completion-wise backwards before one clip/step. Constant rewards give an exact no-op, including optimizer state. Cyclic mismatch (historically labeled sham) left-rotates rewards to $(r_2, \ldots, r_K, r_1)$ only for the update. It is not a neutral zero-mean placebo: for iid completions at a fixed policy and prompt with parameter-independent rewards, its raw ascent-gradient expectation is $-g/(K-1)$ rather than zero, where $g = \mathbb{E}[r \nabla_\Phi \log p_G(y \mid x)]$. At $K = 8$ this is $-g/7$. Appendix E derives the relation and its limits; it does not predict clipped AdamW steps or task scores. Every batch is used once at its generating weights; later batches are freshly on-policy. Reasoning-Time Training (RTT) means numerical learning and subsequent use within the same unfinished problem. Offline semantic fitting and post-closure sensitivity do not meet that definition.

Table 1: Compositional task families. Target templates are evaluator-only. Levels are fixed template strata, not calibrated general difficulty.

Family	Level	Target templates
Arithmetic	1	mul→add; add→mul
Arithmetic	2	clip→mul→add; mul→clip→add
List	1	filter_min→take; take→filter_min
List	2	filter_min→sort→take→append; append→filter_min→sort→take

3.3 Feedback resolution at a fixed internal carrier

The latest study varies candidate-level update feedback on four public cases. For actual output a_i and public expected output y_i, define exact reward $R_e = \frac{1}{4}\sum_i \mathbf{1}[a_i = y_i]$. Behavioral similarity and dense reward are

$$s_i = \begin{cases} (1 + |a_i - y_i|)^{-1}, & \text{arithmetic}, \\ 1 - \dfrac{d_{\text{edit}}(a_i, y_i)}{\max(1, |a_i|, |y_i|)}, & \text{lists}, \end{cases} \qquad R_d = \tfrac{1}{2}R_e + \tfrac{1}{8}\sum_i s_i.$$

Here d_{edit} uses unit insertion/deletion/substitution on integer sequences, not JSON characters. Invalid programs, execution failures or wrong output types receive both rewards zero with the full case denominator. Only exact public perfection closes a branch. The mixture weight is fixed, not selected on audit, and no policy-ordering or generalization guarantee is claimed for this proxy.

Four arms share the same internal initialization, rate, information and candidate ceiling: F freezes Φ; E updates with R_e; D updates with R_d; S updates with cyclic-mismatched R_d. Public scoring and selection in S remain truthful. All arms receive the same aggregate-feedback schema: exact loss, passed, total, validity, error, and a dense score rounded to six decimals. Full stored precision is used for selection and updates. Selection minimizes the tuple $(1 - R_e, -R_d, \text{generation ordinal})$: dense score only breaks an exact-loss tie. Thus D versus F measures update benefit beyond selection using the same dense information, while D versus E tests which reward drives the updater. This common packaging/tie rule differs from earlier studies; historical means are not pure reward effects. E and D do not use different final ranking rules. Their reward objectives change advantages and later on-policy trajectories; update counts are post-intervention outcomes, not a variable on which we match or select cases. The comparison estimates the effect of these online regimes, not a pure information-resolution effect at equal optimizer counts or FLOPs.

The RLOO objective targets a single completion's expected training reward at the current policy and prompt. Our primary endpoint is the hidden exact loss of a public-selected answer from an adaptive finite pool. Improving mean reward, sampling a good tail candidate and selecting it are distinct; there is no monotonic implication between these objectives. This mismatch is an interpretive possibility, not an identified cause of the measured gap.

4 Tasks, timing and estimation

Current-fact qualification supplies explicit workflow order and tests parameter binding from arrived records. Compositional repair instead gives a behavioral goal and four public I/O examples; the model chooses operation order and operands. Facts and M remain fixed throughout a request. Twenty hidden cases score behavioral equivalence, not equality to a target AST.

Integer facts use scale $\{-3, -2, -1, 1, 2, 3\}$, offset $[-7, 7]$, fixed clip bounds $(-16, 16)$, minimum $[-3, 3]$, limit $[1, 4]$, and suffix $[-8, 8]$. Arithmetic inputs are distinct integers in $[-32, 32]$; list inputs are distinct six-integer lists with entries in $[-8, 8]$. Public and audit streams are separate and inputs disjoint within a problem. Gold programs, templates and audit cases are excluded from model inputs, grammar and writers. The DSL permits 1–8 operations, integer literals $[-128, 128]$ and take counts $[0, 8]$; valid syntax can still fail execution, for example with reversed clips.

Each branch generates all eight candidates before public scoring. Its next prompt includes aggregate feedback from the current batch's public-best candidate, not that candidate's program or a per-case

Table 2: Study-specific units and changes. All neural comparisons condition on one fitted common base. These rows cannot be pooled into independent tasks or interpreted as a continuous learning curve.

Study	Physical units	Intervention / comparison
Shared semantics	32 states; 16 pairs	Offline adaptation; two input presentations per state.
Candidate budget	32 requests, one stream	Rank-four real/sham/frozen; up to 32 candidates.
Update strength	16 problems, two repeats	Publicly calibrated output LR versus original rate, frozen and sham.
Carrier design	16 problems, two repeats	Internal versus calibrated output carrier, frozen and internal sham.
Feedback resolution	32 problems, one stream	Fixed internal carrier; exact/dense/sham update reward, common dense selection.

correction trace. The primary endpoint selects from the whole pool so far. Initial batches are shared only after current-weight parity checks; subsequent branches generate under their own state, feedback and RNG. Exact public success ends further sampling and learning, with the result carried to fixed endpoints. It does not imply hidden-audit perfection. Only rounds 1–3 update in the three later studies; round 4 generates and selects. The original candidate-budget study also charged a fourth-round update for subsequent auxiliaries.

For round r, $C_r \sim p_{\Phi_{r-1}}(\cdot \mid x_r)$ precedes scoring and any update to Φ_r. Endpoint E_r selects a candidate already generated in $\bigcup_{j \le r} C_j$. Round-four candidates can therefore reflect at most three earlier updates, and the pool can select a still earlier candidate. All outputs and public choices are locked before audit. Full loop details and the historical fourth-round distinction appear in Appendices Q.2 and D.

Configured search enumerates length-one through length-four products of the public primitive set, allowing repeats, with operands bound by a fixed mapping from actual resident facts. It uses the same public examples and the study's public selection rule. This binding is an inductive advantage, not an exhaustive full-DSL oracle or an equal-FLOPs method. There are 120 arithmetic and 340 list configurations per problem. Search is evaluated once per physical problem, including in studies with two sampling repeats.

The latest 32 problems have four instances of each template; eight separate engineering problems supply checks, not inherited weights. All 40 tuples are distinct and avoid available historical tuple/semantic inventories. There are 400 historical input-value overlap exposures; common-adapter training overlap is unexcluded and pretraining overlap unknown. This is a new-group development evaluation within familiar templates, not independent-domain or training-subject confirmation. Earlier studies retain their own disclosed overlap and construction rules (Appendix Q).

Here *prespecified* (called *registered* in historical tables) means an internal protocol frozen before execution, not external preregistration. The companion gives a redacted protocol extract and identity hashes; these are not independent timestamps. Primary loss is the failed fraction of 20 hidden cases, equally averaged over complete problems. For feedback resolution the registered differences are F−D, E−D and S−D. Saved 10,000-draw whole-problem paired bootstrap intervals use seed 20260920019 and three Bonferroni comparisons at familywise $\alpha = 0.05$. Gain requires a decrease of at least 0.05 and a strictly positive adjusted lower bound. No request, invalid candidate or early success is removed. In the two 16-problem studies, repeats are first averaged within problem; they are not 32 independent problems. The original budget study has two, not three, registered comparisons. Study-specific endpoints and saved intervals are retained without new statistical resampling. When all observed paired differences are zero, the empirical bootstrap interval is $[0, 0]$; this conditional finite-sample result does not establish population equivalence.

Table 3: Feedback resolution on 32 physical problems, one paired sampling stream. Loss is the hidden exact failure fraction; all final selected programs execute. Optimizer steps are summed across problems, with at most three per problem.

Method	Audit loss	Cases /640	Perfect /32	Optimizer steps
Frozen selection (F)	0.7015625	191	2	0
Exact-reward update (E)	0.6531250	222	2	31
Dense-reward update (D)	0.7031250	190	2	83
Dense cyclic mismatch (S)	0.7312500	172	2	85
Configured search	0.0281250	622	28	—

Table 4: Registered control-minus-D improvements. Positive favors D. Saved whole-problem paired intervals are Bonferroni-adjusted for three comparisons; none meets reduction 0.05 with strictly positive lower bound. W/T/L counts D wins/ties/losses.

Contrast	Estimate	Adjusted interval	W/T/L	Gain
F−D	-0.0015625	$[-0.0625000, 0.0656250]$	1/29/2	No
E−D	-0.0500000	$[-0.1578125, 0.0468750]$	1/25/6	No
S−D	+0.0281250	$[-0.0875000, 0.1500000]$	5/20/7	No

5 RESULTS

5.1 DISCRIMINATIVE FEEDBACK AND DIVERGENT CANDIDATES WITHOUT REGISTERED GAIN

All 32 formal problems complete in the feedback-resolution comparison (Table 3). All final selected neural programs execute; not every candidate does. None of the three registered dense-learning comparisons meets the materiality and interval rule (Table 4). E has the smallest neural point estimate. Its descriptive 0.0484375 reduction from F is not a registered primary contrast, and there is no exact-sham arm: S shuffles dense, not exact, rewards.

The manipulation changes more than parameters. All 462 active arm-round candidate groups have nonconstant dense scores, including 275 with all-zero exact rewards. These are exposures containing shared first batches and dependent rounds, not 462 independent problems. D alone has 115 active groups, 66 converted from exact-all-zero to dense-variable, and 83 optimizer steps in total across the 32 problems, with at most three steps per problem. Of 29 applicable attained round-two prefix probes, 28 have positive TV and one has zero. D's saved candidate-token streams differ from F/E/S on 27/25/28 of 32 problems. Thus no update and identical candidate streams cannot fully explain this result. Discrimination and functional use still do not establish alignment with hidden exact correctness.

E/D/S take 31/83/85 updates: D versus E is not equal-update or equal-FLOPs. All arms share the dense public field and exact-first selection. Their first endpoint agrees by construction. Figure 1 shows the saved later endpoints; overlapping curves are retained. Family results, public-success/hidden-failure counts, keep/reset auxiliaries, winner timing and every problem remain in Appendix D. Snapshot effects are not substituted for full online-regime effects.

The formal call uses 199 optimizer steps and 3,088 generations, comprising 2,928 actual main candidates and 160 auxiliaries; 3,696 logical candidate exposures include sharing. Appendix D retains complete physical-position, backward, wall-time, engineering and restoration accounting. Equal candidate ceilings do not imply equal computational cost.

5.2 CANDIDATE COVERAGE VERSUS PUBLIC SELECTION

A subsequent post-hoc CPU evaluation scores the previously unscored members of these locked pools, without generating new candidates. For a saved pool $\mathcal{C}$ and its original public winner c^*, define

$$L_{\text{pool}} = \min_{c \in \mathcal{C}} L_{\text{audit}}(c), \qquad R_{\text{select}} = L_{\text{audit}}(c^*) - L_{\text{pool}}, \qquad L_{\text{selected}} = L_{\text{pool}} + R_{\text{select}}. \quad (1)$$

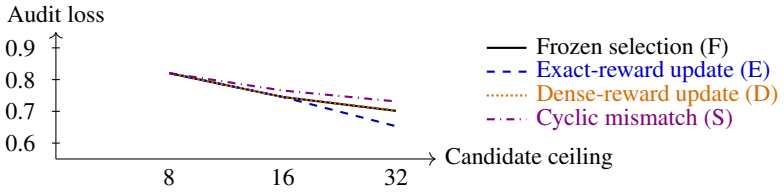

Figure 1: Saved fixed endpoints of the feedback-resolution study. Curves overlap exactly at eight candidates; F/E/D also overlap at 16. Changes within a curve combine sampling, feedback packaging, selection and, for learning arms, updates.

Table 5: Final saved-pool decomposition over all 32 problems. Lower loss is better. Pool and selected perfect counts refer to the same 20 audit cases per problem. Post-hoc audit-best selection is unavailable at deployment; original public choices are unchanged.

Method	L_{selected}	L_{pool}	R_{select}	Pool perfect	Selected perfect
Frozen (F)	0.7015625	0.6421875	0.0593750	2/32	2/32
Exact update (E)	0.6531250	0.6093750	0.0437500	2/32	2/32
Dense update (D)	0.7031250	0.6390625	0.0640625	2/32	2/32
Cyclic mismatch (S)	0.7312500	0.6843750	0.0468750	2/32	2/32

Invalid programs retain loss one. All 32 problems remain in each denominator; the original selected outputs and primary comparisons are unchanged. Table 5 separates observed candidate coverage from the additional loss of public selection.

For each method, both pools containing an audit-perfect candidate already have an audit-perfect public winner. The other 30/32 saved pools contain no such candidate on the fixed 20 cases. Even selecting by audit leaves mean loss 0.609375–0.684375 and the same complete-problem counts. Thus the final failures cannot be attributed entirely to overlooking correct answers already present in these pools. Selection still leaves partial-case headroom: F/E/D/S have positive regret on 16/15/17/13 problems. D improves pool-best loss over F by only 0.003125, while its regret is 0.0046875 larger, leaving the original F$-$D difference of -0.0015625.

The three nonzero F$-$D selected-loss differences show distinct finite-pool outcomes (Table 17). Arithmetic problem 7 has worse pool-best loss under D (0.95 versus 0.50); list problem 7 has better pool-best loss (0.70 versus 0.80) but larger selection regret (0.30 versus 0). List problem 15 improves both pool-best loss (0.35 versus 0.75) and regret (0 versus 0.25). Their selected-loss differences are -0.50, -0.20 and $+0.65$; the other 29 are zero. This post-hoc description exhausts changes in selected loss, not changes in the pools or the previously measured 27/32 candidate-stream divergences. D's winners originate in rounds 1, 2 and 4 after 0, 1 and 3 preceding updates, respectively; these counts do not give the total updates on those problems.

These are descriptive identities, not separately randomized generation and selection effects. The audit-best minimum is optimistic on the same 20 cases and unavailable at deployment; it bounds selection within the observed pool, not intrinsic model capacity or future sampling. Appendix D.1 gives all endpoint accounting, scoring costs and dependent ranking counts.

5.3 Supporting mechanisms and separate memory evidence

The evidence map in Table 2 distinguishes capabilities from incremental task gains. Shared adaptation establishes controlled current-fact grounding. In the earlier budget study all three arms improve from 0.91875 to 0.78125 as attempts and feedback evolve; the update increment remains zero. Post-hoc, all 32 paired candidate streams match, and each online arm's final winner predates its first actual update on 29/32 requests. Public calibration and an internal carrier subsequently produce small, uncertain task differences, while the carrier's finite supervised diagnostic establishes a usable gradient path. These are distinct studies, with full measurements in Appendix A and study-specific appendices. In particular, the carrier study's 111 internal-real batches and 18 optimizer steps do not describe the later dense-feedback arm's 83 steps.

Eight single-event prefixes compare every legal slot and NULL in each action's own successor memory, holding policy fixed with no later writes. Four label futures choose the action before four independent audit futures. Mean loss is 0.43125 for label choice and the relevance/source-trust rule, versus 0.5140625 for NULL: gains of zero and 0.0828125 against the respective controls. The rule's nonzero loss is not a solved-admission ceiling. A separate explicit-workflow policy check is a ceiling: all rewards are one, all control audit losses zero, and updates zero. The delayed-demand study instead gives selected/NULL/public-demand losses of 0.75 each (Appendices J–K). The full-PDSA efficacy factorial on the qualified common base, including learned admission and a strong simple joint comparator, remains unmeasured. The earlier engineering switch study is retained in Appendix G; it is not that efficacy factorial.

6 Discussion and conclusion

PDSA requires distinguishing what changed from what helped. Shared semantic training establishes current-fact grounding; larger candidate budgets with evolving public feedback improve selected answers equally with and without updates; calibrated strength and an internal carrier produce small, uncertain differences. Dense feedback removes exact-reward degeneracy in the observed active batches and produces frequent updates, functional reuse and divergent candidates, yet no registered incremental audit benefit. The strong search baseline demonstrates substantial remaining neural headroom without implying equal computational or inductive assumptions. The saved-pool diagnosis finds no missed audit-perfect solutions and substantial residual pool-best loss. It locates errors in observed candidate coverage as well as imperfect public selection, without identifying an intrinsic capacity limit.

These results narrow explanations but do not uniquely identify a failure cause. The dense proxy may be misaligned with hidden exact correctness; step size, credit assignment, limited updates and task coverage remain possible contributors. Distribution change alone supplies no answer-quality guarantee, and a small learning rate alone does not force an unchanged answer (the conditional bound in Appendix Q.6 requires feature scales and margins). The three later comparisons change different factors on different groups. Their means cannot identify a cross-study causal improvement. The cyclic control's interpretation is given in Section 3 and Appendix E.

Inference remains conditional on one fitted common base, familiar primitives, finite programs, small development sets and disclosed overlaps. Complete per-method FLOPs, energy, independent-domain transfer, learned future-blind admission, full-system joint benefit and negative loss-scale interaction have not been established. Table 25 consolidates execution and material-availability limits, including the internal-carrier parent's missing final full-state receipt. The resulting diagnostic distinguishes feedback variability, functional policy reuse and selected-answer improvement instead of treating the first two as proxies for the third. Finite-pool scoring further separates residual candidate errors from selection regret; the cyclic control has the explicit gradient interpretation given above. These findings concern the tested PDSA components.

AI usage

AI assistance supported the state formalization, hypotheses and experimental design; conditional mathematical arguments and proof drafting; implementation of methods, synthetic-data generators and analysis; interpretation of retained results; literature organization; and manuscript drafting, revision, translation and figure preparation. Reported measurements come from retained execution artifacts, not generated estimates. AI-assisted author pre-review informed the revision. Human authors remain responsible for the final claims, disclosure and submission.

Reproducibility statement

The main text specifies the state domains, probabilities, update objective, controls and endpoints. The appendices retain qualification criteria, per-request results, costs, earlier failures and data limitations. The LaTeX source package rebuilds the document. The anonymous companion contains the fitted common adapter, initial policy states, original bound inputs and allocation records for twelve studies, pinned source, structured traces and results. The core pool includes all 1,429 program

texts and 3,696 logical memberships, public winners, cases and saved audit scores. File, tensor and saved-data checks verify identities, pairings, denominators and the 384 decomposition rows without executing programs. Re-execution of raw-pool scoring and fixed-skeleton parameter rebinding remain unrun; no model rerun from the package is claimed. Individual hidden-case outcome flags and some historical environment/receipt fields were not recorded. A per-study index identifies coverage and gaps, including inherited readout results outside this raw-asset export. These checks do not constitute an independent reproduction of model generation or scoring. Table 25 identifies the remaining evidence and material gaps.

REFERENCES

Arash Ahmadian, Chris Cremer, Matthias Gallé, Marzieh Fadaee, Julia Kreutzer, Olivier Pietquin, Ahmet Üstün, and Sara Hooker. Back to basics: Revisiting REINFORCE-style optimization for learning from human feedback in LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12248–12267. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.662. URL https://aclanthology.org/2024.acl-long.662/.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde, Jared Kaplan, Harri Edwards, Yura Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, Will Guss, Alex Nichol, Igor Babuschkin, Suchir Balaji, Shantanu Jain, Andrew Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. Evaluating large language models trained on code, 2021. URL https://arxiv.org/abs/2107.03374v1. arXiv:2107.03374v1.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685v2*, 2021. URL https://arxiv.org/abs/2106.09685v2.

Zhengbo Jiao, Hongyu Xian, Qinglong Wang, Yunpu Ma, Zhebo Wang, Zifan Zhang, Dezhang Kong, and Meng Han. Policy of Thoughts: Scaling LLM Reasoning via Test-time Policy Evolution, 2026. URL https://arxiv.org/abs/2601.20379v1. arXiv:2601.20379v1.

Yang Li, Zhiyuan He, Yuxuan Huang, Zhuhanling Xiao, Chao Yu, Meng Fang, Kun Shao, and Jun Wang. Adapting Like Humans: A Metacognitive Agent with Test-time Reasoning, 2025. URL https://arxiv.org/abs/2511.23262v1. Version 1; version-specific text consulted 13 September 2026.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101v3*, 2019. URL https://arxiv.org/abs/1711.05101v3.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Sean Welleck, Bodhisattwa Prasad Majumder, Shashank Gupta, Amir Yazdanbakhsh, and Peter Clark. Self-refine: Iterative refinement with self-feedback, 2023. URL https://arxiv.org/abs/2303.17651v1. arXiv:2303.17651v1.

OpenBMB. MiniCPM5-1B-SFT. Model repository, fixed revision, 2026. URL https://huggingface.co/openbmb/MiniCPM5-1B-SFT/tree/a60b37f1fc409c54e1e337b0723aaac6f92dfec0. Revision a60b37f1fc409c54e1e337b0723aaac6f92dfec0.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling llm test-time compute optimally can be more effective than scaling model parameters, 2024. URL https://arxiv.org/abs/2408.03314v1. arXiv:2408.03314v1.

Yu Wang, Ryuichi Takanobu, Zhiqi Liang, Yuzhen Mao, Yuanzhe Hu, Julian McAuley, and Xiaojian Wu. Mem-α: Learning Memory Construction via Reinforcement Learning, 2025. URL https://arxiv.org/abs/2509.25911v1. Version 1; version-specific text consulted 13 September 2026.

Sikuan Yan, Xiufeng Yang, Zuchao Huang, Ercong Nie, Zifeng Ding, Zonggen Li, Xiaowen Ma, Jinhe Bi, Kristian Kersting, Jeff Z. Pan, Hinrich Schütze, Volker Tresp, and Yunpu Ma. Memory-R1: Enhancing Large Language Model Agents to Manage and Utilize Memories via Reinforcement Learning, 2026. URL https://arxiv.org/abs/2508.19828v5. Version 5; version-specific text consulted 13 September 2026.

Yi Yu, Liuyi Yao, Yuexiang Xie, Qingquan Tan, Jiaqi Feng, Yaliang Li, and Libing Wu. Agentic Memory: Learning Unified Long-Term and Short-Term Memory Management for Large Language Model Agents, 2026. URL https://arxiv.org/abs/2601.01885v3. Version 3; version-specific text consulted 13 September 2026.

Yanwei Yue, Boci Peng, Xuanbo Fan, Jiaxin Guo, Qiankun Li, and Yan Zhang. Mem-T: Densifying Rewards for Long-Horizon Memory Agents, 2026. URL https://arxiv.org/abs/2601.23014v2. Version 2; version-specific text consulted 13 September 2026.

Mert Yuksekgonul, Daniel Koceja, Xinhao Li, Federico Bianchi, Jed McCaleb, Xiaolong Wang, Jan Kautz, Yejin Choi, James Zou, Carlos Guestrin, and Yu Sun. Learning to Discover at Test Time, 2026. URL https://arxiv.org/abs/2601.16175v2. Version 2; version-specific text consulted 13 September 2026.

Guilin Zhang, Wei Jiang, Xiejiashan Wang, Aisha Behr, Kai Zhao, Jeffrey Friedman, Xu Chu, and Amine Anoun. Adaptive Memory Admission Control for LLM Agents, 2026a. URL https://arxiv.org/abs/2603.04549v1. Version 1; version-specific text consulted 13 September 2026.

Kehao Zhang, Shangtong Gui, Sheng Yang, Wei Chen, and Yang Feng. Learning to Remember: End-to-End Training of Memory Agents for Long-Context Reasoning, 2026b. URL https://arxiv.org/abs/2602.18493v2. Version 2; version-specific text consulted 13 September 2026.

Shengtao Zhang, Jiaqian Wang, Ruiwen Zhou, Junwei Liao, Yuchen Feng, Zhuo Li, Yujie Zheng, Weinan Zhang, Ying Wen, Zhiyu Li, Feiyu Xiong, Yutao Qi, Bo Tang, and Muning Wen. MemRL: Self-Evolving Agents via Runtime Reinforcement Learning on Episodic Memory, 2026c. URL https://arxiv.org/abs/2601.03192v2. Version 2; version-specific text consulted 13 September 2026.

Junyi Zou and Avrova Donz. MMLA: How Memory Lets the Past Shape the Future, 2026a. URL https://arxiv.org/abs/2606.28876v3. Public version 3; historical title retained.

Junyi Zou and Avrova Donz. MMLA: Memory-Mediated Learning Architecture for Predictive Dual-State Adaptation, 2026b. URL https://arxiv.org/abs/2606.28876v4. Version 4, 14 September 2026; scientific evidence cutoff 13 September 2026.

Yuxin Zuo, Kaiyan Zhang, Li Sheng, Shang Qu, Ganqu Cui, Xuekai Zhu, Haozhan Li, Yuchen Zhang, Xinwei Long, Ermo Hua, Biqing Qi, Youbang Sun, Zhiyuan Ma, Lifan Yuan, Ning Ding, and Bowen Zhou. TTRL: Test-Time Reinforcement Learning, 2025. URL https://arxiv.org/abs/2504.16084v3. Version 3; version-specific text consulted 13 September 2026.

A SUPPORTING INTERVENTION OVERVIEW

A.1 COMMON SEMANTICS, CANDIDATE COVERAGE AND STRONGER UPDATES

Shared offline adaptation gives 16/16 audit-perfect greedy programs and 8/8 successful rebinding pairs in each of four family/input strata, paired over 32 physical states and 16 rebinding groups. The original criteria are 15/16 and 7/8 with valid untruncated programs. Its fixed 288-update fit follows 16 discarded engineering updates; qualification does not select a checkpoint. This establishes controlled grounding, not RTT (Appendix H).

The subsequent one-step compositional study has 13 all-zero and three nonconstant reward groups. Three real and three sham updates produce actual later-prefix distribution differences, but real/reset/sham greedy losses remain 0.965625; search has loss 0.03125 and solves 14/16 requests. All 158/160 executable scored outputs and two failures remain (Appendix M). Interface and earlier diagnostic failures are retained separately, not combined into an adaptation learning curve.

In the 32-request candidate-budget study, real, sham and frozen methods each have loss 0.78125, pass 140/640 cases and solve 2/32 requests. Both registered differences are zero request-wise, with saved empirical intervals $[0, 0]$ under the interpretation in Section 4. All three improve from 0.91875 at eight candidates through 0.81250 at 16 to 0.78125 at up to 32. This combines sampling, public selection and evolving feedback packaging, not an isolated sample-count or update-count effect. Round-one equality is by design; later equality is an observation. Search passes 621/640 cases and solves 28/32.

Post-hoc descriptive comparison under the paired random-number setting finds identical saved candidate streams and final selected program texts on all 32 requests across the three main arms. In each online arm, 29/32 selected candidates predate that request's first actual update; this is not a count of requests without updates and does not identify a numerical cause. Each online arm takes 39 main steps; 15/30 applicable later-prefix probes have positive TV. Auxiliary keep/reset effects include both improvements and degradations. Appendix B preserves all endpoints, the before/after update counts, costs and failures.

A.2 CALIBRATED UPDATES AND AN INTERNAL CARRIER RETAIN A TASK-BENEFIT GAP

Public-only calibration selects output-carrier learning rate 0.01, within its empirical bounds but below target median TV 0.01. On 16 new problems with two paired repeats, calibrated real reduces loss by 0.0296875 over frozen and 0.015625 over sham. Neither meets materiality or a positive adjusted interval.

The internal carrier lowers supervised diagnostic NLL from 0.97620 to 0.08928 in 32 fixed updates; those weights are discarded. This is limited trainability, not RTT. Separate public calibration selects 10^{-4}, below its target but within both real/sham bounds. In comparison, frozen-minus-internal real loss is 0.0125 with interval $[0, 0.05]$; sham-minus-real is -0.0125 with $[-0.05, 0]$; output-minus-internal real is zero with $[0, 0]$. No registered upgrade is established. Equal output/internal means include one problem's opposite repeat-level differences of $+0.4$ and -0.4. A better cyclic-mismatch mean has the conditional control interpretation in Appendix E; the zero interval follows the statistical convention in Section 4.

In the earlier carrier study (Table 6), internal real encounters 111 batches: 75 all-zero and 36 nonconstant, with 18 formal updates and retained later-prefix functional use. The completed table reuses 31 immutable repeats and a separately authorized last-repeat completion. Its fresh-model check cannot replace the missing parent-wide final state receipt (Appendix C).

Table 6: Separate paired-group studies, 19–20 September. Each neural row has 16 physical problems, two repeats and 640 case exposures to 320 distinct audit cases. Search runs once per problem; its denominator is 16, not 32. Columns from different studies are not a paired comparison.

Method	Update-strength loss	Carrier-study loss
Frozen sampling	0.7234375	0.703125
Original output real (10^{-4})	0.7234375	—
Calibrated output real (0.01)	0.6937500	0.690625
Calibrated output sham (0.01)	0.7093750	—
Internal real (10^{-4})	—	0.690625
Internal sham (10^{-4})	—	0.678125
Configured search	0 (16/16)	0 (16/16)

B MULTI-ROUND FEEDBACK: COMPLETE DEVELOPMENT RESULTS

The evidence cutoff is 18 September 2026, 08:17 UTC. All 32 requests, including zero-reward and early-success requests, completed. Each row below is a paired request with 20 audit cases. The three main selected programs are evaluated separately; their request-wise losses happen to agree. The main column therefore displays each of real feedback, cyclic sham and frozen public-selected sampling, not a pooled average. Display indices are within-family labels, not independent training seeds. The accompanying aggregate CSV reproduces the table without private run paths.

B.1 REGISTERED STATISTICS AND DEPENDENCE

The primary endpoint is the audit loss of the earliest minimum-public-loss candidate at the full available budget. Public feedback, not audit, selects the output. The saved paired differences are exactly zero for real versus frozen sampling and real versus sham on all requests. The registered 10,000-draw paired bootstrap, with Bonferroni adjustment for two comparisons, therefore returns $[0, 0]$ for both. This is a finite empirical statement conditional on the fixed subject and observed groups, not a powered equivalence test or evidence of zero effect on a new population.

The primary analysis retains complete requests and their paired methods. The 32 requests contain 30 parameter tuples and 31 semantic–parameter groups. Repeated parameter exposure can induce dependence beyond the request unit; the bootstrap does not model all such population dependence. One known historical parameter-tuple overlap remains. The available historical composition manifests, not every prior training corpus, define the overlap check. Shared-adapter training overlap is not excluded and unknown pretraining data cannot be certified clean. No independent-domain result or training-seed population inference is claimed.

Post-hoc description of the saved candidate streams. After the registered analysis, we compared all saved answer-token sequences and program texts at corresponding generation positions, without running a model, executing a program or recomputing scores. On all 32 requests the real, sham and frozen streams match, including their invalid candidates; their publicly selected final program texts also match. This is a direct trace comparison, not an inference from equal audit losses. It does not imply identical parameters or token distributions: the separate prefix probes detect differences. The selected programs first appeared in rounds 1, 2 and 3 on respectively 24, 5 and 3 requests, and on no request in round 4. In each online branch, 29 selected programs were generated before any actual update, two after one update and one after two. These descriptive counts retain closed and zero-signal requests; they do not identify a numerical root cause or replace the registered primary endpoint.

B.2 LEARNING SIGNALS AND AUXILIARY CONTRASTS

For each online arm, rounds 1–4 complete respectively 12, 8, 11 and 8 optimizer steps across all requests. At the round-1, round-2 and round-4 candidate-generation boundaries, the cumulative counts are 0, 12 and 31; after those rounds they are 12, 20 and 39. These are charged operations, not the number of updates experienced by every candidate. Selected-output endpoints use candidates generated before the update of their respective round. A fourth-round candidate can therefore reflect

Table 7: Multi-round feedback comparison, 18 September. Loss is lower-better over 20 audit cases per request. All 32 selected programs in each row execute. Candidates and cases are dependent within the 32 paired requests.

Method	Mean audit loss	Cases /640	Perfect /32
Real-feedback online policy	0.78125	140	2
Cyclic-sham online policy	0.78125	140	2
Frozen public-best-of-32	0.78125	140	2
Initial greedy	0.98906	7	0
Aggregate-feedback context	0.97969	13	0
Configured symbolic search	**0.02969**	**621**	**28**

Table 8: Predetermined budget endpoints on the same 32 requests. Each main method has the displayed loss. Update counts are cumulative across requests for *each* online arm, not per-request counts. *Before* counts steps completed before that round's new candidates; *after* includes that round's charged steps. Early closures are carried forward without further work. Round 1 is shared by design.

Endpoint	Candidates	Each loss	Updates before	Updates after
Initial greedy	—	0.98906	0	0
Round 1	8	0.91875	0	12
Round 2	16	0.81250	12	20
Round 4	32	0.78125	31	39

at most three preceding round updates. The final eight updates affect subsequent auxiliary probes, not the already generated primary candidate pool.

Each main method encounters 124 eight-candidate groups. Real and sham each record 83 all-zero groups and 41 nonconstant groups. Of the latter, two close the problem on public success before updating; the remaining 39 produce actual updates. There are no all-one or other-constant groups. Sixteen requests update in each online branch. Two list requests finish at round 2; they remain in the primary denominator and are carried forward to later budget endpoints without additional sampling.

The round-3 functional check uses the first actually sampled completion at the earliest position with multiple legal semantic choices. Its prefix is reached within a still-open problem and is not supplied by a label. Fifteen of 30 applicable requests have positive TV in each online branch; the other 15 have zero. The two closed requests are inapplicable rather than failed probes. These measurements do not count as improved task scores.

Real final keep/reset differs on three requests, with two favorable and one unfavorable change; sham differs on two. No one-step anchor yields an audit-perfect request. A fixed-snapshot reset effect is different from the total effect of an online regime whose future samples and feedback may diverge. Neither replaces the registered selected-output comparison.

B.3 Complete resource boundary

Discarded engineering work accounts for 97 generations, 10 updates and 18,649,088 physical tokens. The formal package accounts for 2,988 generations, 94 updates and 830,550,016 physical tokens. Formal updates comprise 39 real, 39 sham and eight in each four-sample anchor. The 992 logical candidates per main method total 2,976 across methods; sharing the first 256 candidates reduces actual main-candidate generation to 2,464. The remaining generations serve registered endpoints and other checks. Each main method has 983/992 executable candidates. Invalid candidates retain their failure and reward.

All 2,208 saved main/anchor strict probability checks agree on support, EOS and log probabilities at recorded precision, with maximum error zero. This subset is not the full count of 2,821 replay calls, which also includes other probes and engineering work. Source and input identities, 32 pre-audit output locks, frozen common weights and state restores passed their recorded integrity checks. No

Table 9: Arithmetic requests in the multi-round comparison. Updates show real/sham actual main steps. Candidates are per main method. Tokens include all per-request branches and auxiliary work.

Request	Each main loss	Search loss	Context loss	Updates	Candidates	Tokens (10^6)
1	1.00	0.00	1.00	0/0	32	30.040
2	1.00	0.00	1.00	0/0	32	27.619
3	1.00	0.00	1.00	0/0	32	30.257
4	1.00	0.00	1.00	2/2	32	28.324
5	1.00	0.00	1.00	0/0	32	27.898
6	1.00	0.00	1.00	0/0	32	28.033
7	0.75	0.00	1.00	2/2	32	29.479
8	1.00	0.00	0.95	0/0	32	28.967
9	1.00	0.00	1.00	0/0	32	30.110
10	0.50	0.00	1.00	1/1	32	29.839
11	0.75	0.00	1.00	2/2	32	27.365
12	0.70	0.00	1.00	3/3	32	27.730
13	1.00	0.45	1.00	0/0	32	29.696
14	1.00	0.00	0.95	0/0	32	29.270
15	1.00	0.00	1.00	0/0	32	29.815
16	0.75	0.00	1.00	3/3	32	25.915

Table 10: List requests in the multi-round comparison. Updates show real/sham actual main steps. Candidates are per main method. Tokens include all per-request branches and auxiliary work.

Request	Each main loss	Search loss	Context loss	Updates	Candidates	Tokens (10^6)
1	1.00	0.00	1.00	0/0	32	25.944
2	0.95	0.25	1.00	3/3	32	26.120
3	1.00	0.10	1.00	0/0	32	26.202
4	0.35	0.00	1.00	2/2	32	22.122
5	1.00	0.00	0.95	2/2	32	23.876
6	0.75	0.00	1.00	3/3	32	25.993
7	0.80	0.00	0.95	4/4	32	25.113
8	1.00	0.00	1.00	0/0	32	24.338
9	0.00	0.15	0.55	0/0	16	9.622
10	1.00	0.00	1.00	0/0	32	25.010
11	0.45	0.00	1.00	2/2	32	25.616
12	0.00	0.00	1.00	0/0	16	9.699
13	0.65	0.00	1.00	1/1	32	24.523
14	0.60	0.00	1.00	4/4	32	23.495
15	0.20	0.00	1.00	3/3	32	26.821
16	0.80	0.00	1.00	2/2	32	25.698

inference was rerun for this paper. The fixed native eager BF16 frame remains 8×512; physical cost includes every padding and dummy position.

Two earlier bounded attempts did not reach any formal request. One stopped at its budget after 97 engineering generations and 10 updates. Another was stopped by the owner after at least 77 recorded generations and five updates; its complete cost is unavailable. These are incomplete attempts, not negative formal results, and their costs are not silently added to the completed invocation using an invented total.

B.4 Frozen base and restoration

The same public revision and qualified shared adapter are used in every neural condition. The common semantic adapter has 688,128 parameters in 32 tensors (2,752,512 float32 bytes). The rank-four online policy occupies 2,113,536 float32 parameter bytes, excluding gradients and optimizer moments. Before/after tensor identities and read/restore checks were retained. Resetting policy, optimizer and RNG preserves the common adapter and memory; restoration does not erase incurred

Table 11: Auxiliary greedy contrasts, distinct from the primary public-selected endpoint. The final-state denominator is 30 because two requests had already closed. One-step anchors retain all 32 requests and use the predetermined first four initial samples, not a selected subset.

Policy / endpoint	Keep loss	Reset loss	Requests
Real, final multi-round state	0.988333	0.986667	30
Sham, final multi-round state	0.981667	0.986667	30
Real, one-step four-sample anchor	0.965625	0.989063	32
Sham, one-step four-sample anchor	0.982813	0.989063	32

Table 12: Whole-invocation cost, including discarded engineering work, main branches, anchors, replay, probes and restoration. Counts are actual operations, not successful learning events. Per-method GPU time/FLOPs and energy were not separately measured.

Quantity	Measured value
Model loads	1
Generation requests / generated tokens	3,085 / 108,866
Physical forward calls	207,324
Physical token positions	849,199,104
Dummy/padding token positions	796,730,970
Backward calls / optimizer steps	104 / 104
Replay calls / replayed tokens	2,821 / 98,268
Executor calls / tool attempts	57,838 / 10,927
Configured symbolic programs	7,480
Symbolic-search timer (μs)	173,102
Supervisor wall time (s)	19,662.150
Peak CUDA allocated / reserved (bytes)	4,741,867,008 / 5,356,126,208
Final output bytes	253,180,276

cost. Private provenance retains the complete model/checkpoint and result hashes. These checks establish the recorded state boundaries, not an independent-domain performance guarantee.

C UPDATE-STRENGTH AND INTERNAL-CARRIER STUDIES

These development studies use the qualified, always-enabled semantic base, public task primitives, named resident facts and typed grammar. Each separately allocates 16 physical problems (two per template) and two paired sampling repeats. Repeats share four public and 20 audit cases: 640 neural exposures correspond to 320 distinct audit cases. Search is selected and audited once per problem, then shared between repeats. Neither study supplies training-seed or independent-domain inference. Allocations exclude available historical parameter/semantic inventories, not unknown pretraining or all common-adapter training content. These latter overlaps remain unexcluded.

Both use eight candidates per batch, at most four rounds, and updates only in rounds 1–3. Round 4 generates and selects without optimization. Exact public success closes a branch before further updating; constant rewards remain no-ops. Initial sharing requires the registered current-weight path check; later batches use each branch's own policy, RNG and aggregate feedback. Primary selection minimizes public loss, with earliest ordinal breaking ties. Audit cannot select a rate, checkpoint, program or endpoint. This differs from the earlier candidate-budget study's charged fourth-round auxiliary update.

C.1 PUBLIC DISTRIBUTION-CHANGE CALIBRATION

The output study has eight separate calibration problems, six with eligible nonconstant rewards. Each trial restores the same initial policy and empty optimizer, then tests one rate in $\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$ on public feedback. Attained-prefix changes select a scalar, not a checkpoint; trial weights are discarded. Target median group-mean TV is 0.01, with median ceiling 0.05, maximum prefix TV 0.20 and maximum KL 0.50. The bounded fallback selects 0.01: median TV is 0.00574795. Rate 0.1 violates the bounds (maximum TV 0.54684, KL 0.65105). These are empirical calibration criteria, not guarantees on all prefixes or task outcomes.

Calibrated real passes 196/640 audit exposures with 2/32 perfect repeats; frozen and original-rate real each have 177/640 and 1/32; calibrated sham has 186/640 and 2/32. Frozen-minus-calibrated-real and original-minus-calibrated-real reductions are both 0.0296875, with adjusted interval $[0, 0.0796875]$. Sham-minus-real is 0.015625 with $[-0.01875, 0.059375]$. The whole-group bootstrap uses 10,000 draws, seed 20260918017, three Bonferroni comparisons at familywise 0.05, and requires reduction at least 0.05 and a strictly positive lower bound. No comparison passes. This does not uniquely identify parameterization as cause.

C.2 INTERNAL CARRIER AND FINITE TRAINABILITY

Sixteen projection-local rank-16 increments act on q/v at zero-based layers 16–23, alpha 32, dropout zero, FP32 factors, random nonzero input factors and zero output factors. The complete original module and frozen semantic adapter stay enabled. Internal policy has 688,128 parameters (2,752,512 bytes), compared with 2,113,536 parameter bytes for the rank-four output carrier; gradients and moments are additional. This jointly changes position, rank and parameter distribution. No single architectural cause is isolated.

An eight-example supervised diagnostic takes 32 fixed updates at 2×10^{-4} with answer/EOS-token-normalized masked CE. Aggregate NLL falls from 0.9761997642 to 0.0892833724. Diagnostic weights and optimizer are discarded. This is training-support learnability, not feedback-driven RTT. A separate eight-problem calibration tests $10^{-5}, 10^{-4}, 10^{-3}$ with real and sham trial steps; four problems are eligible, yielding 24 updates. Both feedback types must satisfy the empirical bounds. The chosen 10^{-4} is bounded but below target; 10^{-3} is excluded. Only the scalar rate enters comparison; no diagnostic labels or fitted trial weights are inherited.

F/O/I/S denote frozen, calibrated output-real (0.01), internal-real (10^{-4}) and internal-sham (10^{-4}). Passing exposures are 190, 198, 198 and 206 of 640; each has four perfect repeats. Search passes all 320 distinct cases on 16 problems. Inference averages repeats, resamples all methods together over 16 groups, and uses 10,000 draws, seed 20260919019, three Bonferroni-adjusted comparisons and the same gain criterion. All 16 output/internal group means agree, but one arithmetic problem has output/internal losses $0.7/0.3$ in one repeat and $0.3/0.7$ in the other. A win and a loss cancel; equality is not imposed at the repeat level.

Formal updates are output-real 17, internal-real 18 and sham 19. Internal real has 75 all-zero and 36 nonconstant batches among 111 actual batches. Each online arm has ten applicable positive-TV later-prefix records. Whole-package updates total 113: 32 supervised, three engineering, 24 calibration and 54 formal. These are operations, not 113 useful-learning events. The paired group means and all repeat-level outcomes accompany the source as descriptive aggregate data.

C.3 Completion provenance, costs and limits

The original carrier invocation stopped at a policy-load budget boundary after preserving 31 complete repeat outputs. A prospectively authorized completion retained them unchanged and ran only the last registered repeat at its original seed. Its shared initial batch publicly closed all four branches, correctly adding zero updates. All 32 outputs then entered the original audit/analysis. The original incomplete invocation and failed partial-repeat costs remain.

The parent performed its 39th policy load before its counter refused the operation at cap 38. Combined actual loads are 41 versus 40 recorded increments. This is a retained accounting fault, not inferred compliance. The original 31 per-repeat recovery checks remain; the fresh completion's full theta/state check cannot replace the missing parent-wide final receipt. These limits qualify the combined evidence without deleting outcomes or changing scores.

Table 13: Recorded whole-package costs. Carrier totals include the original attempt, failed partial work and last-repeat completion. No per-method energy or matched-FLOPs efficiency is inferred.

Quantity	Update strength	Carrier plus completion
Launcher wall seconds	25,512.056	26,771.466
Generation requests	3,464	3,038
Physical forward calls	268,915	276,462
Physical token positions	1,101,475,840	1,132,388,352
Optimizer operations	119	113
Configured programs	3,800	3,680
Output bytes	556,522,686	386,839,045

Both keep native eager BF16 8×512, temperature one and no KV cache. Padding/dummy positions are physical work, not distinct data tokens. The output study charges 13 engineering, 24 calibration and 82 formal updates; its final common and policy-restoration checks pass. The internal study's more limited parent-wide proof is not supplied by another process. Earlier unsuccessful attempts remain separate where cumulative costs are unavailable. Energy and some per-method times were not measured.

The subsequent fixed-carrier feedback-resolution comparison is complete and reported separately in Section 5.1 and Appendix D. Its new information/selection interface and new physical problems prevent a pure reward-effect interpretation of cross-study differences. It does not repair the parent-state provenance limitation described above.

D Fixed-carrier feedback resolution: full evidence

The formal invocation on 21 September completed all 32 physical problems, one paired sampling stream per problem, after reusing checks from eight engineering problems. No engineering policy or optimizer was inherited. Four instances of each of the eight templates appear; 20 hidden audit cases are shared between methods within a problem. A case exposure, arm, round or candidate is not an additional independent problem. The fixed method compares the frozen internal carrier, exact-reward learning, dense-reward learning and dense cyclic sham, with configured symbolic search retained. Primary scores remain the original saved outcomes. The pool diagnosis below uses additional post-hoc CPU scores obtained on 23 September.

Table 14: Paired descriptive decomposition: control loss minus treatment loss. Each selected difference equals its pool and regret differences. F−E is additional arithmetic, not a fourth registered primary test. No new intervals or tests are computed.

Contrast	Selected difference	Pool difference	Regret difference
F−D	-0.0015625	+0.0031250	-0.0046875
E−D	-0.0500000	-0.0296875	-0.0203125
S−D	+0.0281250	+0.0453125	-0.0171875
F−E	+0.0484375	+0.0328125	+0.0156250

Table 15: Residual selection differences at the final endpoint. Pair counts compare distinct programs within each problem's best-public-exact-loss bucket; undetermined includes a tie on either axis. These dependent pairs are not an independent statistical sample.

Arm	Positive regret	Audit spread in tie	Concordant	Discordant	Undetermined
F	16/32	14/32	306	112	2377
E	15/32	12/32	309	110	2374
D	17/32	13/32	374	163	2520
S	13/32	9/32	270	137	3338

D.1 Complete saved-pool diagnosis

For a fixed saved pool $\mathcal{C}$ and the originally public-selected candidate c^*, the post-hoc decomposition is

$$L_{\text{pool}} = \min_{c \in \mathcal{C}} L_{\text{audit}}(c), \qquad R_{\text{select}} = L_{\text{audit}}(c^*) - L_{\text{pool}}, \qquad L_{\text{selected}} = L_{\text{pool}} + R_{\text{select}}.$$

Invalid programs retain loss one. The saved pools contain all 3,696 candidate exposures, or 1,429 distinct problem–program items. Deduplication binds the problem identity, exact program text, audit-case identity and executed scorer version; programs from different problems do not share a score. Original endpoint, auxiliary and search scores cover 104 items. The subsequent bounded CPU evaluation adds the missing 1,325 scores on the original 20 cases per problem, without repairing programs or changing public choices. It makes 26,063 actual case-executor calls rather than the 26,500 upper bound because the original interpreter short-circuits invalid programs. Invalid loss remains one; infrastructure failure is distinct from a program error.

All 32 problems, four arms and endpoints 1, 2 and 4 are present in 384 unique rows. Every row satisfies the decomposition to absolute tolerance 10^{-12}, and all original selected-endpoint losses are preserved. Duplicate removal saves scoring calls only: candidate exposure counts, early closure and the 32-problem denominator are unchanged. The anonymous companion provides these derived rows and their winner source round and number of preceding updates, along with a deterministic checker against the original selected scores. Without raw programs and cases, this checker does not independently recompute pool minima or the original deduplication.

The final endpoint is in Table 5; Tables 14 and 16 give paired differences and corresponding case counts. For every method, the set of two audit-perfect pools equals the set of two audit-perfect public winners. No already sampled audit-perfect answer was missed. The other 30 pools per method lack a perfect member on these cases. The audit-best member is a finite-sample minimum on the same 20 cases, hence optimistic and unavailable to a deployable selector. It neither replaces a primary result nor bounds future sampling or intrinsic model expressivity.

Table 17 is an exhaustive description of nonzero selected-loss differences among all 32 problems. The three differences sum to -0.05, giving the original mean -0.0015625 with 29 zeros. Pool losses can change even when selected losses do not. These rows describe worse candidate coverage, better coverage offset by worse selection, and joint coverage/selection improvement, respectively. They neither identify unique causes nor support extrapolation beyond the saved pools; no program structure or operand error is inferred from the loss values.

Positive regret remains on 16/15/17/13 problems for F/E/D/S. Within each problem's best-public-exact-loss bucket, compare unique program pairs by dense reward and audit correctness. Concor-

Table 16: Case counts derived from the complete saved-score means, not new evaluations. Every audit-perfect pool has an audit-perfect original public selection; pool-best selection increases partial-case counts only.

Arm	Selected cases	Pool-best cases	Selected problems	Pool-perfect problems
F	191/640	229/640	2/32	2/32
E	222/640	250/640	2/32	2/32
D	190/640	231/640	2/32	2/32
S	172/640	202/640	2/32	2/32

Table 17: All three problems with nonzero final F−D selected-loss differences. Each loss/regret cell is F / D, on the same 20 audit cases. The final two columns refer to D's selected candidate, not to the problem's total update count. These existing-output descriptions are post-hoc, not prespecified subgroup tests.

Problem	L_{selected}	L_{pool}	R_{select}	Round	Prior updates
Arithmetic 7	0.50 / 1.00	0.50 / 0.95	0.00 / 0.05	1	0
List 7	0.80 / 1.00	0.80 / 0.70	0.00 / 0.30	2	1
List 15	1.00 / 0.35	0.75 / 0.35	0.25 / 0.00	4	3

dance means the two orderings agree; discordance means they disagree; undetermined includes a tie on either axis. Table 15 retains all three categories. D's 374 concordant, 163 discordant and 2,520 undetermined pairs do not establish overall anticorrelation. Pairs within a problem share programs and cases; these are dependent descriptive counts, not calibrated correlations or independent statistical observations. The additive differences likewise do not identify separate causal generation and selection effects. The additional F−E decomposition is arithmetic only, with no new confidence interval or confirmation claim for E. Original registered contrasts are unchanged.

The scoring invocation completes in 0.669186 seconds with 1,325 program scores and 26,063 executor calls; its saved output occupies 496,012 bytes. It uses a 2 GiB address-space limit; actual peak RSS and CPU hardware were not recorded. There is no new model load, forward, sampling or training. A retained summary snapshot contains a preterminal `INVALID` status written before the final transition; its numerical counts agree with the final complete receipt. The stale marker is not a scoring failure and the original file is unchanged.

Completed support check; prospective validation not run. A separate metadata-only projection reads seven bound historical allocation files: 184 records, 240 tuple exposures and 158 unique tuples. The fixed supports contain 90 arithmetic and 476 list tuples, leaving 12 and 396 after the historical exclusions. The previously fixed eight-template allocation order admits 8, 16 or 24 balanced physical groups, not 128 or 200. This check takes 0.051576 seconds, outputs 2,049 bytes and creates no new cases or scores. Capacity is not a power or throughput guarantee. No prospective model validation is run, and no expanded support or reused-tuple confirmation is substituted. This later inventory does not alter the original study's allocation and overlap accounting below.

D.2 Public information and the executed loop

The renderer, typed grammar, base, semantic adapter and rate are common. The internal policy comprises 688,128 FP32 parameters in q/v projections of zero-based layers 16–23, rank 16, alpha 32 and no dropout. The input-factor initialization is uniform on $[-1/\sqrt{d_{\text{in}}}, 1/\sqrt{d_{\text{in}}}]$, from a separate CPU generator with seed 20260919018; output factors start at zero. AdamW has rate 10^{-4}, betas $(0.9, 0.999)$, epsilon 10^{-8} and zero weight decay. The learned common adapter always remains enabled. The output-residual carrier is inactive, not jointly optimized.

For each problem and branch the actual order is:

1. Restore the registered initial policy, empty optimizer and paired sampling stream. Generate eight initial candidates once. Check their current-weight prefix log probabilities under

each separately reset branch before public scoring; otherwise use the registered separate-sampling fallback. All 32 initial batches passed sharing in this invocation.

2. In an open round, obtain eight completions under the branch's own current policy and prompt. Score the complete batch on the four public examples. Append its candidates to the pool and select endpoints at rounds 1, 2 and 4 using exact loss, descending dense reward, then earliest ordinal.
3. The next prompt's aggregate feedback comes from the current batch's best candidate under that same ordering. It contains the inherited exact fields plus the six-decimal dense field, not the generated program, a correction trace, or hidden feedback.
4. Exact public success closes the request. Otherwise E, D and S update only in rounds 1–3, using exact reward, dense reward, or left-rotated dense reward respectively. Constant update rewards are exact no-ops. Round 4 never updates. Closed endpoints carry without further work.
5. For an updated, still-unresolved online branch, the first actual round-two sample supplies the earliest semantic position with multiple legal choices. At fixed prefix, support and other inputs, compare attained and initial policy. The rule never selects a maximum-TV position or gold prefix. Final keep/reset greedy auxiliaries apply only to still-open branches.
6. Persist all neural outputs, choices, search selections and restoration records before opening any formal audit. Audit does not affect stopping, reward, rate, checkpoint, candidate choice or the set of problems.

TV is half the summed absolute probability difference on the same ordered legal support; KL is $\mathrm{KL}(p_{\mathrm{initial}} \| p_{\mathrm{attained}})$. For this metric only, saved legal log weights are normalized in CPU double precision before exponentiation. Sampling and same-weight strict replay retain their executed probabilities. This normalization is not a new policy, support change or epsilon smoothing. Functional use at an actual later prefix and final snapshot sensitivity are separate measurements.

The public score of an invalid program is zero under both rewards with the full four-case denominator, even if a prior case passed. Integer distance uses arithmetic values; list distance uses sequences of integers. The fixed half-exact/half-similarity mixture is not guaranteed to order policies by hidden exact loss. S uses real public feedback and selection; only the updater receives the left-rotated dense reward vector.

D.3 Allocation, inference and original denominators

Data seed 20260920019 derives independent Python random streams from SHA-256 of canonical sorted-key JSON arrays. Allocation order is engineering then comparison, arithmetic then list, template order from Table 1, then instance order. The descriptive stream key combines split, family, zero-based template and instance. A shuffled finite parameter support supplies the first tuple whose parameter and semantic-group identities are absent from the checked inventory and earlier allocations. Public and audit inputs use separate stream keys and exclude overlap within a problem. An additional sampling stream provides the paired generation seed. Construction metadata and reference programs are not model features.

Eight engineering groups have public inputs only. The 32 formal groups have four public and 20 audit cases each. All 40 parameter tuples are unique; known tuple and semantic-group overlap with checked historical inventories is zero. The inventory covers 144 rows and 118 unique tuples, not every historical training distribution. There are 400 historical input-value overlap exposures. Common-semantic training overlap is not excluded; pretraining coverage is unknown. These qualifications remain even though new problem identifiers and held-out audit streams are used.

The registered bootstrap resamples complete paired problems, 10,000 times with seed 20260920019, with all four methods retained together. Its three control-minus-D comparisons use Bonferroni intervals at familywise 0.05 and require a point reduction of at least 0.05 and a strictly positive lower bound. The saved intervals in Table 4 were not recomputed for this manuscript. There is one fitted common subject and one sampling stream per problem, not a training-seed population. E's descriptive advantage over F is not a newly registered fourth comparison; S is not an exact-sham comparator. No denominator or threshold is changed.

Table 18: Every arithmetic problem in the feedback-resolution comparison. Each loss uses 20 cases; no problem is excluded.

Problem	F	E	D	S	Search
1	0.50	0.50	0.50	0.50	0.00
2	0.45	0.45	0.45	0.45	0.00
3	1.00	1.00	1.00	1.00	0.00
4	0.85	0.85	0.85	1.00	0.00
5	1.00	1.00	1.00	1.00	0.00
6	0.30	0.30	0.30	1.00	0.00
7	0.50	1.00	1.00	1.00	0.00
8	1.00	1.00	1.00	1.00	0.00
9	0.60	0.60	0.60	0.55	0.00
10	0.45	0.45	0.45	0.45	0.00
11	1.00	0.30	1.00	1.00	0.00
12	0.95	0.95	0.95	0.90	0.00
13	1.00	1.00	1.00	1.00	0.45
14	0.10	0.10	0.10	0.10	0.00
15	0.70	0.70	0.70	0.50	0.00
16	1.00	1.00	1.00	0.10	0.00

Table 19: Every list problem in the feedback-resolution comparison. Each loss uses 20 cases; no problem is excluded.

Problem	F	E	D	S	Search
1	1.00	0.90	1.00	1.00	0.10
2	0.80	0.50	0.80	0.80	0.10
3	0.80	0.80	0.80	1.00	0.00
4	0.35	0.35	0.35	0.35	0.00
5	0.85	0.85	0.85	0.80	0.00
6	1.00	1.00	1.00	1.00	0.00
7	0.80	0.80	1.00	0.80	0.00
8	0.95	0.90	0.95	0.85	0.00
9	0.00	0.00	0.00	0.00	0.00
10	1.00	1.00	1.00	1.00	0.00
11	0.00	0.00	0.00	0.00	0.00
12	0.25	0.25	0.25	1.00	0.00
13	1.00	0.10	1.00	1.00	0.00
14	1.00	1.00	1.00	1.00	0.00
15	1.00	1.00	0.35	1.00	0.00
16	0.25	0.25	0.25	0.25	0.25

D.4 Secondary outcomes and endpoint provenance

All 462 active branch-round batches have nonconstant dense scores; 275 have exact-all-zero rewards. The 50 remaining positions among 512 registered arm-round positions are closed carries, not missing batches. D converts 66 of its own 115 active groups from exact-all-zero to dense-variable. It takes 83 updates; 28/29 applicable later-prefix probes have positive TV. The zero-TV probe is retained. E and S have 12/12 and 29/29 positive probes. These exposure counts include shared initial samples and dependent rounds.

D's candidate-token streams differ from F, E and S on 27, 25 and 28 problems respectively, as recorded in the original closeout. Candidate divergence does not imply better selected programs. The saved final primary selections have public exact success but imperfect audit on respectively 3, 4, 3 and 3 problems in F/E/D/S. All 32 problems remain in each primary denominator. Tables 22 and 23 retain the applicable auxiliary subsets and winner provenance rather than selecting a favorable family or update subset.

Table 20: Within-family means of the feedback-resolution comparison; 16 physical problems per family.

Family	F	E	D	S	Search
Arithmetic	0.7125000	0.7000000	0.7437500	0.7218750	0.0281250
List	0.6906250	0.6062500	0.6625000	0.7406250	0.0281250

Table 21: Feedback manipulation and functional-use evidence. Batches are dependent arm-round exposures; positive-TV counts use only applicable attained prefixes.

Arm	Active	Exact zero	Dense variable	Optimizer steps	Applicable	TV > 0
F	115	69	115	0	0	0
E	115	67	115	31	12	12
D	115	66	115	83	29	28
S	117	73	117	85	29	29

D.5 Integrity, resource boundaries and incomplete history

Every final selected neural program executes; only 3,641 of 3,696 public-scored candidate exposures do. All 3,088 generated outputs terminate without truncation. There are 2,928 distinct main generations plus 160 auxiliaries; the 3,696 logical main exposures include 768 shared duplicates. Search runs once on each physical problem, enumerating 7,360 configurations. Its operands are bound from resident facts and its selection uses the same exact-first, dense-tie rule. This study allows 60 seconds per search and requires complete enumeration; no completed problem truncated its support. It is neither a full-DSL optimum nor an equal-FLOPs neural comparator.

All 2,264 completely saved strict replay operand records agree on mask, EOS and log probability at recorded precision, maximum error zero. The charged total is 4,880 replayed completions: successful checks without complete operand serialization are not silently added to the saved-evidence denominator. There are 6,368 finite connected gradient-tensor records and zero projected parameter elements. Zero first-step input-factor gradients at zero output factors are permitted. The saved closeout verifies 60 source identities, 32 locked problem outputs, 32 independent recovery records, and final foundation/common-adapter/initial policy identities. Memory restoration preserves payload and rendered prompt; its provenance epoch can change. Restores do not erase costs.

The formal call uses 278,445 forwards, each with 4,096 physical positions. Of its 1,140,510,720 positions, 1,068,223,190 are dummy/padding positions; 108,687 generated answer tokens measure a different quantity. Peak CUDA allocation/reservation is 7,356,746,752/8,717,860,864 bytes, peak RSS is 3,193,044,992 bytes, and saved output is 447,578,011 bytes. The executed package binds the same Python 3.11.15, PyTorch 2.11.0+cu128, Transformers 5.12.1 and CUDA-build 12.8 inventory. Hardware records remain limited as described in Table 38; no new machine query supplies missing historical measurements.

The preceding invocation completed eight public engineering groups but did not reach formal evaluation before stopping. Its 62 updates, 862 generations and 7,309.422849 seconds are charged separately and not relabeled a complete formal study. The fresh formal invocation reused its checks, not its weights. The sum in Table 24 is two calls, not one faster or longer single run. Per-method Transformer FLOPs and energy were not measured. The new successful final-state records do not fill the earlier carrier parent's missing full-state receipt. The subsequent CPU pool scoring adds hidden evaluation of old candidates, not new model experiments. This document integration performs saved-result arithmetic and source checks only; it does not repeat scoring or confer a new independent-review status.

Table 22: Prespecified final keep/reset auxiliaries. Only still-open requests are applicable; these snapshot effects do not replace online-regime primary effects.

Arm	Applicable	Keep loss	Reset loss
E	26	0.9615385	0.9807692
D	27	0.9629630	0.9814815
S	27	0.9814815	0.9648148

Table 23: Final winners' source round and preceding actual updates. Counts are problems, not update operations. Public-perfect but audit-imperfect selections remain failures on their missed cases.

Arm	Source rounds 1/2/3/4	Prior updates 0/1/2/3	Public/audit gap
F	16/8/6/2	32/0/0/0	3
E	13/8/8/3	26/6/0/0	4
D	16/9/4/3	16/9/4/3	3
S	16/8/2/6	16/8/2/6	3

Table 24: Recorded feedback-resolution cost: earlier engineering invocation, complete formal invocation, and their sum. A two-call sum is not one invocation; energy and per-method FLOPs were not measured.

Quantity	Engineering	Formal	Two-call sum
Model loads	1	1	2
Optimizer steps	62	199	261
Microbatch backwards	496	1,592	2,088
Generations	862	3,088	3,950
Generated tokens	29,976	108,687	138,663
Forward calls	74,994	278,445	353,439
Physical positions	307,175,424	1,140,510,720	1,447,686,144
Executor cases	11,314	46,999	58,313
Tool attempts	2,848	11,203	14,051
Search configurations	1,840	7,360	9,200
Launcher seconds	7,309.422849	26,996.634683	34,306.057532

E Interpreting the cyclic reward-mismatch control

The executed update left-rotates the reward vector and then applies the same leave-one-out centering as the real update. This section clarifies that specific control; it changes no historical reward, score or comparison. The argument uses standard score-function identities underlying REINFORCE and RLOO (Ahmadian et al., 2024), not a new optimization theorem.

Fix the policy, prompt and parameter-independent legal support. Let $y_1, \dots, y_K$ be iid complete answers, including EOS, and let their rewards have no direct parameter dependence. Write $s_i = \nabla_\Phi \log p_G(y_i \mid x)$, $g = \mathbb{E}[r_i s_i]$ and $\bar{r} = K^{-1} \sum_j r_j$, assuming the expectations and differentiation under the sum are valid. The score identity gives $\mathbb{E}[s_i] = 0$. Independence therefore implies

$$\mathbb{E}[r_j s_i] = 0 \quad (j \neq i), \qquad \mathbb{E}[\bar{r} s_i] = g/K.$$

For the real raw ascent estimator,

$$\widehat{g}_{\mathrm{real}} = \frac{1}{K} \sum_i \frac{K}{K-1} (r_i - \bar{r}) s_i, \qquad \mathbb{E}[\widehat{g}_{\mathrm{real}}] = g.$$

A fixed cyclic permutation π has no fixed points. Although $\mathbb{E}[r_{\pi(i)} s_i] = 0$, the batch mean still contains r_i; hence

$$\widehat{g}_{\mathrm{cyc}} = \frac{1}{K} \sum_i \frac{K}{K-1} (r_{\pi(i)} - \bar{r}) s_i, \qquad \mathbb{E}[\widehat{g}_{\mathrm{cyc}}] = -\frac{g}{K-1}.$$

For eight candidates the expectation is $-g/7$. A zero batch sum of advantages does not make their score-weighted gradient neutral. Thus the reported cyclic control probes a particular mismatch, not generic random feedback or an unbiased zero-gradient placebo.

These are expectations of raw, unclipped and unpreconditioned gradients. Gradient clipping, AdamW moments and normalization, projection, finite precision, conditioning on public-success stopping, and subsequent policy and prompt evolution prevent a direct inference about expected parameter steps or audit loss. A better empirical cyclic-control score does not establish that reversing the task gradient is beneficial. The original cyclic-control measurements remain intact.

For comparison, hold a realized batch $(r_i, s_i)_{i=1}^K$ fixed and average over the uniform distribution on all $K!$ permutations, allowing fixed points. Then $\mathbb{E}_\pi[r_{\pi(i)} \mid (r_j, s_j)_j] = \bar{r}$, so

$$\mathbb{E}_\pi[\widehat{g}_\pi \mid (r_j, s_j)_j] = 0.$$

Sampling uniformly from all permutations realizes this property in expectation; restricting to permutations without fixed points does not. It still gives no zero-mean guarantee for a nonlinear AdamW parameter step. This alternative control has not been evaluated in the reported studies.

F Protocol identity and evidence availability

Prespecified decisions are the internally frozen, pre-execution method, allocation, stopping and statistical rules. The anonymous companion's `PRESPECIFIED_PROTOCOL.md` extracts these from the retained protocol; its hashes identify files but do not supply an independent timestamp or external preregistration. Scientific conclusions use executed artifacts, not source-review decisions.

The recommended companion omits the unexecuted raw-pool and optional model example entries. It supplies the original core programs, memberships, public winners and resident facts, four public and twenty audit cases per problem, and pinned scorer/interpreter source, but no validated standalone scientific replay entry. Audit cases were withheld from generation, updating and selection, not from the offline evaluator. Original post-hoc pool scoring is complete; the later raw-pool recheck and fixed-skeleton public-parameter rebinding remain unrun. Neither supplies additional evidence in this paper. Only file/static and saved-data checks are recommended as executable checks in the companion. The six-class material sheet distinguishes actual files, unrecorded fields, out-of-scope inherited assets and unverified execution modes. It preserves the distinction between original and anonymized file hashes. No new model, program evaluation or statistical resampling was used to prepare this export.

Table 25: Execution scope and material limits. A later successful check does not supply evidence missing from an earlier invocation.

Study / material	Retained evidence	Remaining boundary
Earlier structured/interface studies	Original failures and partial runs, followed by separately identified successors.	Engineering completion is not task benefit; failed outputs are not repaired.
Shared semantic base	One fixed fitted subject; paired qualification and rebinding results.	Not online RTT or cross-subject confirmation.
Internal carrier	31 saved repeats plus one completed repeat; supervised trainability and later-prefix checks.	Parent final full-theta/state receipt is missing; the fresh completion's identity cannot replace it.
Feedback comparison	All 32 formal groups complete; original outcomes, selected outputs, restoration and cost records.	Earlier engineering invocation's incomplete terminal is retained; two-call costs remain separate.
Pool scoring	Complete later CPU scores and 384-row decomposition.	The summary's embedded preterminal status is stale; the separate terminal is complete. No new model generation occurred.
Anonymous companion	Fitted adapter, saved initial-policy states, twelve-study inputs/source/traces, full core pool and saved scores; file/tensor/data checks.	No model rerun or raw-pool recheck. Intermediate checkpoint/cache payloads are omitted; unrecorded fields and inherited-readout raw assets remain missing.

G MEASURED OUTPUT AND STATE–PROGRAM DIAGNOSTICS

The September 15 structured-output engineering study used four physical worlds (two per family), each under the four policy/memory switches. These 16 trajectories are paired executions, not 16 independent tasks. All 684 generations were JSON-valid, schema-valid, EOS-terminated and untruncated; 619 were execution-valid. All 576 sampled-completion replays agreed on masks, EOS and selected-token probability, over 11,662 replayed tokens. Final programs were executable in 11/16 trajectories, task-perfect in 0/16, and passed 10/320 audit cases. An inspection associated all 65 runtime rejections with reversed clip bounds, without repairing their outputs.

Of 128 workflow sampling groups, 106 had all-zero rewards and 22 were nonconstant. There were 11 actual online workflow updates and eight updates in fixed-prefix isolation controls. The latter include sham-feedback conditions and are not eight useful learning events. At two list prefixes, real/keep total variation was 2.9439×10^{-5} and 1.0661×10^{-4}; reset gave zero. Sham/keep also changed the distributions, and all four cells within each prefix generated identical greedy programs and scores. This is a finite functional-reuse witness, not useful feedback-specific improvement. The two arithmetic probes had no reward signal and remained no-ops. This study took 254.979 seconds including the supervisor and charged 7,936,832 processed tokens; energy was not measured.

G.1 PAIRED PUBLIC-FACT INPUTS

A subsequent zero-update diagnostic used 32 sealed attained snapshots, 16 per family, only after all necessary facts had arrived. It generated four samples plus one greedy program per snapshot and neural input condition: 320 requests total. The minimal-input condition lists the sufficient current public facts; the normal-input condition supplies the same facts through memory and recent-observation/feedback packaging. The public-rule reference compiles from arrived release records cross-checked against actual memory reads. It neither reads hidden answers to fill parameters nor provides a program prefix to the model. Table 26 records whole-program audit perfection; all 64 neural greedy programs were executable.

Normal-minus-minimal greedy audit loss was 0.43125 for arithmetic and 0.50000 for lists; registered Bonferroni-adjusted intervals were $[0.209375, 0.65625]$ and $[0.25000, 0.75000]$ respectively. The original 10,000-draw paired bootstrap retained the snapshot and both conditions, conditional on this frozen model and these engineering worlds. It is not independent-domain or trained-seed inference. Ten arithmetic and twelve list normal-input greedy programs exactly matched the un-

Table 26: State–program diagnostic, 15 September 2026: complete programs passing all 20 audit cases, each out of 16 snapshots per family. Best-of-four uses four public feedback cases, earliest sample on ties, then independent audit cases; it never selects on audit outcomes.

Input / selection	Arithmetic	List
Public-rule reference	16/16	16/16
Minimal facts, greedy	6/16	8/16
Normal deployment, greedy	0/16	0/16
Minimal facts, public-selected best-of-four	10/16	12/16
Normal deployment, public-selected best-of-four	1/16	2/16

changed syntax example. These tags can overlap old-value, sensor and sign coincidences; they do not uniquely identify a cognitive mechanism. Different ASTs are not automatically semantically different. The measured input intervention bundles framing, examples, ordering, recent observations and length; their individual effects are not identified.

All 320 completions met the grammar, while 306 were executable; the 14 runtime failures were arithmetic samples. The public reference passed 640/640 audit cases. No model update, output repair or retrospective score replacement occurred. This diagnostic took 83.703 supervisor seconds, one load, 6,489 forwards and 1,559,980 processed tokens. These measurements motivated the shared semantic adaptation evaluated next.

H SHARED SEMANTIC ADAPTATION: PROTOCOL AND QUALIFICATION

The completed bounded intervention preserved the model revision, public grammar, prompt builders, tasks and rewards. It added offline parameters ω_{sem} on query/value projections in transformer layers 16–23 (rank 16, scale 32), with θ_0 and initial online Φ frozen. Its sole objective is answer/EOS-token-normalized teacher-forced cross entropy under p_G; prompt and padding receive no loss. Gradients pass through the transformer to ω_{sem} rather than a detached frozen output. AdamW uses learning rate 2×10^{-4}, physical microbatch 8, effective batch 16, eight fixed epochs and no checkpoint selection on qualification results.

The sealed training set contains 576 supervised sequences, not 576 independent worlds: 48 parameter tuples in 24 rebinding groups, two input forms and six derived views per tuple. Engineering, development and qualification groups are disjoint at the complete parameter/rebinding-group level; changing a world ID does not create independence. The rule labels use only arrived, input-visible facts. Qualifications retain 16 states and eight rebinding pairs per family, with both minimal and normal input forms. Each stratum requires at least 15/16 greedy audit-perfect, 16/16 executable, no truncation, and 7/8 rebinding pairs following changed public facts. This is controlled new-group qualification, not an independent task-family claim.

The execution discarded 16 tiny-fit updates, then performed 288 formal updates, froze the fixed final weights and evaluated the registered strata. Two effective-batch objective measurements each use two physical microbatches and total NLL divided by all answer/EOS tokens, never an unweighted mean of microbatch means. Native eager BF16 forward shapes are fixed at 8×512 for sampling, prefix replay and batched loss; every dummy and padding position is costed. Hard limits are one model load, 890 generations, 304 optimizer steps, 608 backward calls, 121 million physical tokens, 3,600 seconds and 2 GiB output. The physical-token limit can bind before the nominal generation or forward limits.

The bounded execution completed and met the original shared-base criteria (Table 27). All four adapted strata had 16/16 audit-perfect greedy programs, 320/320 audit cases, 16/16 execution-valid programs, no truncation and 8/8 rebinding pairs following both members on the registered distinguishing input. Training-support and fixed-endpoint development each passed 8/8 greedy states in every stratum. One development list/normal sampling group included a partial reward; it did not select or extend training. All 64 adapted qualification sampling groups had reward vector $(1, 1, 1, 1)$.

Normal-input audit loss decreased by 0.8625 for arithmetic and 0.7500 for lists. The registered adapted-minus-unadapted estimates were -0.8625 with adjusted interval $[-0.96875, -0.75000]$

Table 27: Shared semantic adaptation, 16 September 2026. Greedy counts are out of 16 states per family/input, with both input forms paired on the same states. Rebinding counts are out of eight pairs and use the registered behavior-distinguishing input, not whole-audit perfection. Before counts are a descriptive reaggregation of the saved qualification pair records. A task-perfect program passes all 20 audit cases.

Family / input	Before	After	Rebinding before	After
Arithmetic / minimal	13/16	16/16	6/8	8/8
Arithmetic / normal	0/16	16/16	0/8	8/8
List / minimal	15/16	16/16	7/8	8/8
List / normal	4/16	16/16	0/8	8/8

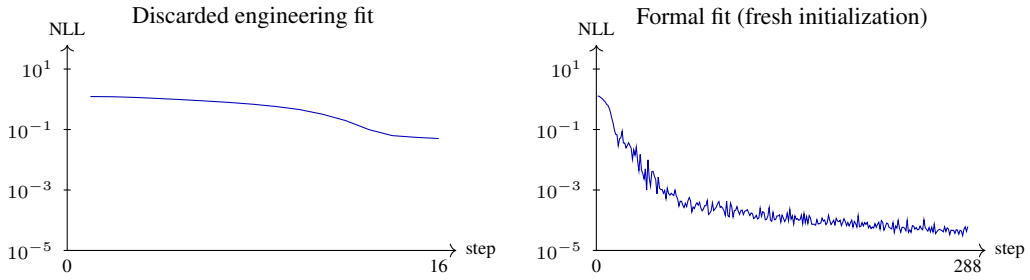

Figure 2: Recorded answer/EOS-token-weighted teacher-forced training objective under the frozen grammar. Vertical axes are logarithmic. All 16 engineering and 288 formal updates are shown separately: engineering weights and optimizer were discarded before formal fitting. These are batch objectives, not qualification accuracy or within-problem learning benefit.

and -0.7500 with interval $[-0.9375, -0.5625]$. These are the original 10,000-draw paired-bootstrap intervals, Bonferroni-adjusted for two comparisons (familywise $\alpha = 0.05$), over eight rebinding groups per family, conditional on the fixed fitted subject. They are not uncertainty over training seeds or unseen rule families. No example-copy, argument-mismatch or order-error tags were observed in adapted qualification. This observation does not uniquely identify why the unadapted model failed.

The run used 304 optimizer steps (16 discarded engineering, 288 formal), 608 backward calls, 890 generations, 18,409 forwards and 75,403,264 physical token positions, including 69,946,489 dummy/padding positions. Supervisor wall time was 1,847.048 seconds; peak allocated CUDA memory was 7,650,932,224 bytes. The final output tree was 27,373,314 bytes; cumulative worker write accounting was 28,421,575 bytes. Energy was not measured. All costs stayed within the frozen caps. Twelve before/after/frozen probability probes agreed exactly at the recorded precision, with zero maximum batch log-probability error; they also preserved the shared-adapter identity. This is actual BF16 evidence for these probes, not a theorem of universal numerical identity. Checkpoint reload reproduced eight selected qualification requests; independent memory restoration reproduced read/generation, and online-policy reset preserved the common adapter with an empty optimizer.

Earlier failures remain separate: a preceding numerical/identity stop occurred after 17 generations and zero updates, while a later static candidate failed integration review without loading a model. Neither was a completed negative adaptation experiment. Qualification here refers to the subsequent measured neural outcomes, not the static integration checks. The qualified $(\theta_0, \omega^*_{\mathrm{sem}})$ is frozen as the common base for every later neural cell and neural sampling baseline. Symbolic search uses no neural adapter and receives the same public information. Reinitializing/restoring Φ retains ω^*_{sem}; its input factor is random and its output factor is zero. Offline improvement is not within-problem RTT, and old functional witnesses do not transfer automatically to the new common base.

I Intervention and evidence map

Table 28: Evidence map for the common base and remaining system claims. The fixed-state, one-step, candidate-budget, calibrated-strength, carrier and feedback-resolution interventions are complete, without the registered primary policy benefit. Predictive-student and joint benefits remain untested; earlier base-specific witnesses are not inherited automatically.

Study	Primary comparison	What a failure means
Policy isolation	Real/sham feedback crossed with keep/reset Φ, fixed M	Initial task: all-one/no-op ceiling. One-step composition: unchanged greedy outcome. Multi-round: no primary gain over frozen sampling or sham despite updates and functional reuse. Dense feedback: more differentiated rewards, updates and divergent candidates without registered audit benefit.
Admission teacher	All slots plus NULL; label choice, independent audit	Completed: margin over NULL, zero over strong rule. No current justification for predictive student training in this environment.
Future-blind admission	Learned future value versus current risk, strong features, downstream reward	Teacher success but student failure limits estimation or transfer, not input integrity by itself.
Online system	Four cells and a strong simple combination	Joint benefit without negative interaction does not establish complementarity.
Harm and recovery	Wrong/stale observations, rejection, deletion, restores	Persistent errors or uncharged recovery limit safe-use and efficiency claims.
Independent confirmation	New registered rule/task family and one stress profile	Controlled-domain success alone does not establish open-domain transfer.

J COMPLETED FROZEN-BASE OPPORTUNITY RESULTS

Both tables report the 16 September 2026 completed development invocation, not new-environment results. Group numbers below are within-family display indices, not independent training seeds. Every registered group is included.

Table 29: Fixed-state policy check. Each loss is over 20 audit cases. RK/RR are real keep/reset; SK/SR are sham keep/reset. Each pair has loss 0/0; the compact table prints their common value. Feedback and best4 are the unchanged-wrapper feedback and public-selected useful sampling controls. All 32 sampled candidates have reward 1; total optimizer and backward counts 0.

Group	Cursor	Rewards	RK/RR	SK/SR	Feedback	Best4
Arithmetic 1	8	$(1, 1, 1, 1)$	0	0	0	0
Arithmetic 2	12	$(1, 1, 1, 1)$	0	0	0	0
Arithmetic 3	20	$(1, 1, 1, 1)$	0	0	0	0
Arithmetic 4	28	$(1, 1, 1, 1)$	0	0	0	0
List 1	8	$(1, 1, 1, 1)$	0	0	0	0
List 2	12	$(1, 1, 1, 1)$	0	0	0	0
List 3	20	$(1, 1, 1, 1)$	0	0	0	0
List 4	28	$(1, 1, 1, 1)$	0	0	0	0

The label-choice advantage over NULL is 0.0828125; over the strong rule it is 0. Only the policy component has zero-loss ceiling evidence. All admission groups retain nonzero strong-rule loss. Four futures/cases/cells within a prefix stay dependent; these data do not support population-wide zero effect or a powered equivalence claim.

Table 30: Single-event admission audit loss. Each action has 4 label and 4 independent audit futures, each with 20 cases. Selection precedes audit. FIFO and current-risk are separate implemented controls, even where their values agree. The final column is the strong relevance/source-trust rule. Groupwise selected-minus-strong loss is exactly 0 in every group.

Group	Label choice	NULL	FIFO	Current-risk	Strong rule
Arithmetic 1	0.8375	0.8375	0.9000	0.9000	0.8375
Arithmetic 2	0.2250	0.3625	0.2250	0.2250	0.2250
Arithmetic 3	0.3125	0.3125	0.3125	0.3125	0.3125
Arithmetic 4	0.2500	0.5625	0.2500	0.2500	0.2500
List 1	0.6875	0.8750	0.6875	0.6875	0.6875
List 2	0.7500	0.7750	0.7500	0.7500	0.7500
List 3	0.1625	0.1625	0.1625	0.1625	0.1625
List 4	0.2250	0.2250	0.2250	0.2250	0.2250
Mean	0.43125	0.51406	0.43906	0.43906	0.43125

K Delayed-demand compositional repair and interface localization

The 17 September frozen-base development invocation completed all eight structure requests and eight single-event admission groups. It retained the qualified common adapter and the original DSL interpreter, but changed the decision problem: public I/O examples specify a composition without giving its operation order. Each structure request has four sampled candidates and one greedy output; the 40 outputs are not 40 independent tasks. Configured symbolic search enumerates length-one through length-four programs using the same public examples and resident configuration, selects by public loss, and breaks ties in fixed enumeration order. It is not exhaustive full-DSL search and is not compute-matched to neural generation.

Table 31: Completed composition requests. Each audit loss uses 20 cases. Valid counts are out of five dependent neural programs. Search is selected on public feedback only. The last column describes the four sampled rewards.

Request	Valid	Greedy	Public best4	Search	Reward group
Arithmetic 1	1/5	1	0.8	0	Nonconstant
Arithmetic 2	0/5	1	1	0	All zero
Arithmetic 3	0/5	1	1	0	All zero
Arithmetic 4	0/5	1	1	0	All zero
List 1	0/5	1	1	0	All zero
List 2	0/5	1	1	0	All zero
List 3	0/5	1	1	0	All zero
List 4	0/5	1	1	0.5	All zero
Mean loss		1.000	0.975	0.0625	

All 40 programs were schema-valid, EOS-terminated and untruncated. Saved programs, token traces and request IDs were cross-checked against the original raw generation stream and hashed result/terminal records before static inspection. The 19 rejected integer-task programs eventually apply a list-only primitive; all 20 list-task programs start with integer-only multiplication. Three programs also contain reversed clipping bounds, but only after an earlier type failure. No program was repaired, re-executed or rescored. This identifies an interface support mismatch, not a guarantee that restricting types will solve operation selection, numerical binding or useful learning.

These are zero relative increments, not evidence that all future tasks are solved or all actions are equivalent. A reliable shared read path and an independently audited margin over an appropriate strong rule are prerequisites for a future predictive learner.

The invocation used one model load, 624 generations, 21,385 physical forwards, 87,592,960 processed token positions (83,491,715 dummy/padding), and zero backward or optimizer calls. The supervisor measured 2,044.532 seconds; peak allocated CUDA memory was 4,705,604,608 bytes.

Table 32: Completed delayed-demand admission audit losses. Each group has four label and four independent audit futures, each of 20 cases; all nine own-successor actions are evaluated with no later writes. LRU, frequency, public-demand and current-risk losses equal the selected and NULL columns in every row. They remain separate implemented rules.

Group	Label-selected	NULL	FIFO
Arithmetic 1	.75	.75	.75
Arithmetic 2	0	0	.2375
Arithmetic 3	.75	.75	.75
Arithmetic 4	.50	.50	.50
List 1	1	1	1
List 2	1	1	1
List 3	1	1	1
List 4	1	1	1
Mean	.7500	.7500	.7796875

Configured search evaluated 1,840 programs; total executor calls were 10,961. Forty strict probability replays matched across 1,355 tokens, and eight save/reload plus independent memory restorations passed. Energy was not measured. The memory groups reuse two object worlds across demand/capacity variants; known historical parameter overlap and unexcluded shared-adapter training overlap preclude an independent-domain claim.

L A SHARED TYPED INTERFACE AND ONE-STEP FEEDBACK

On 17 September, a new development package completed 16 registered compositional-repair requests, eight per family, without selecting requests by outcome. All neural conditions share the frozen semantic base and a combined interface: actual resident facts are rendered as named numbers, and the grammar restricts operations by the publicly given input type only. Neither rendering nor support uses correct operation order, gold parameters or audit results. Memory and facts remain fixed during each open request.

Four candidates receive public-feedback rewards. Real and cyclic-sham branches each start at the same random nonzero input factor A, zero output factor B, and empty AdamW state (learning rate 10^{-4}, zero weight decay), and take at most one update. At this initialization, the task gradient of A is zero; this is one-step B adaptation on initial random features, not multi-step joint factor learning. Keep/reset evaluations share each branch computation. Outputs and public-best-of-four selections are persisted before independent audit. Configured symbolic search uses the same public examples and resident facts, not audit-based selection.

Validity does not imply useful candidates. All 160 originally scored neural outputs are schema-valid, EOS-terminated and untruncated; 158 execute. The two invalid arithmetic samples retain their original error and loss of one. A read-only posthoc inspection found reversed clip bounds; no output was repaired or rescored. There are 13 all-zero and three nonconstant four-candidate reward groups, with no all-one or other-constant groups. The 32 branch update calls yield three real and three sham optimizer/backward steps plus 26 exact no-ops. These six updates are not six useful learning events. None of the 64 initial sampled programs is audit-perfect; six requests contain a partially correct sampled program. That coverage description never selects a candidate using audit results.

Functional reuse without a changed greedy outcome. All 192 current-weight probability replays agree, including support and EOS; the maximum recorded log-probability discrepancy is zero. The functional probe uses the earliest multi-option semantic position reached in the actual fifth attempt of the same unresolved request, with non-policy inputs fixed and no gold prefix. All 64 cell probes are applicable. Exactly the three updating groups have positive real-keep and sham-keep total variation; all 32 reset probes are zero. Nevertheless, all 16 real-keep token sequences equal their real-reset and sham-keep counterparts, and their audit losses are identical (Table 33). This establishes a finite update/reuse path on the shared base, not useful RTT.

Table 33: Completed one-step feedback comparison, 17 September 2026. Loss is the failed fraction of 20 audit cases per request. The 16 paired requests, not the 320 case evaluations or four cells, are the analysis groups. RK/RR mean real keep/reset; SK/SR mean sham keep/reset. All four have identical observed values and are displayed together, not pooled.

Condition	Arithmetic	List	All	Perfect /16
RK, RR, SK, SR (each)	.962500	.968750	.965625	0
Frozen greedy	.962500	.968750	.965625	0
Feedback in context	1.000000	.881250	.940625	0
Public-best-of-four	.893750	.937500	.915625	0
Configured search	.000000	.062500	.031250	14

The registered gain, control loss minus real-keep loss, is zero against reset and sham, with empirical intervals $[0, 0]$, and -0.934375 against search, with interval $[-0.993750, -0.850000]$. These are 10,000-draw paired-bootstrap intervals with Bonferroni adjustment for three registered comparisons at familywise $\alpha = 0.05$. They describe the paired groups conditional on one fitted subject; they do not prove population equivalence. No admission, predictive learner or joint training was included. Appendix M retains every request, the functional probes and complete cost boundary.

M Completed typed-interface intervention: every request

The following tables use within-family display indices, not training seeds. All 16 sealed semantic/parameter tuples are retained. Three tuples overlap earlier engineering data; overlap with shared semantic-adaptation training has not been excluded, and pretraining coverage is unknown. This is same-generator development, not independent-domain confirmation. The older 40-program result used different inputs and remains unchanged; comparing its executability with the present result does not identify a causal effect of type support separately from named-fact rendering.

Table 34: Per-request audit losses. RK/RR/SK/SR and frozen greedy agree in every row; the shared-value column represents each condition separately. Context uses the unchanged aggregate public feedback wrapper. Best4 is selected on public feedback with the fixed tie rule, never audit.

Request	Each policy / frozen	Context	Public best4	Search
Arithmetic 1	.95	1.00	.40	.00
Arithmetic 2	1.00	1.00	1.00	.00
Arithmetic 3	.95	1.00	.95	.00
Arithmetic 4	1.00	1.00	1.00	.00
Arithmetic 5	1.00	1.00	1.00	.00
Arithmetic 6	.80	1.00	.80	.00
Arithmetic 7	1.00	1.00	1.00	.00
Arithmetic 8	1.00	1.00	1.00	.00
List 1	1.00	1.00	1.00	.00
List 2	1.00	1.00	.95	.05
List 3	1.00	1.00	1.00	.45
List 4	.75	.75	1.00	.00
List 5	1.00	1.00	1.00	.00
List 6	1.00	1.00	1.00	.00
List 7	1.00	.30	1.00	.00
List 8	1.00	1.00	.55	.00
Mean	.965625	.940625	.915625	.031250

At $B_0 = 0$, for residual $Z = HA^\top B^\top$ and upstream derivative G, $\nabla_A L = (GB)^\top H = 0$ whereas $\nabla_B L = G^\top HA^\top$ can be nonzero. An initial zero A task gradient is therefore expected under zero weight decay and empty optimizer state. It does not imply detachment, nor does a nonconstant reward guarantee a nonzero B gradient in every possible instance. Four keep/reset cells reuse the two actual branch computations; 64 logical uses are not 64 updates or independently trained trajectories.

Table 35: All four-candidate reward groups and update counts. Each nonconstant group has one real and one sham update, not two successful learning events. All other TV entries are exactly zero; all real-reset and sham-reset TVs are zero. TV compares the reached semantic-prefix distribution under kept versus initial policy.

Request	Public rewards	Updates	Real-keep TV	Sham-keep TV
Arithmetic 1	$(0, 0, .5, 0)$	2	2.530×10^{-6}	1.840×10^{-5}
Arithmetic 2	$(0, 0, 0, 0)$	0	0	0
Arithmetic 3	$(0, 0, 0, 0)$	0	0	0
Arithmetic 4	$(0, 0, 0, 0)$	0	0	0
Arithmetic 5	$(0, 0, 0, 0)$	0	0	0
Arithmetic 6	$(0, 0, 0, .25)$	2	3.566×10^{-7}	4.061×10^{-5}
Arithmetic 7	$(0, 0, 0, 0)$	0	0	0
Arithmetic 8	$(0, 0, 0, 0)$	0	0	0
List 1	$(0, 0, 0, 0)$	0	0	0
List 2	$(0, 0, 0, 0)$	0	0	0
List 3	$(0, 0, 0, 0)$	0	0	0
List 4	$(0, 0, 0, 0)$	0	0	0
List 5	$(0, 0, 0, 0)$	0	0	0
List 6	$(0, 0, 0, 0)$	0	0	0
List 7	$(0, 0, 0, 0)$	0	0	0
List 8	$(0, 0, .5, 0)$	2	6.395×10^{-5}	1.002×10^{-4}

The 160 scored neural generations comprise four initial samples, four policy cells, frozen greedy and context-only per request. Public-best4 selects one of the initial four. Another 32 generations verify restoration rather than contribute new task successes. All 16 frozen-greedy programs execute; 62/64 initial samples execute. Arithmetic request 4 sample 4 and arithmetic request 7 sample 3 have reversed clips in the static posthoc inspection. The interpreter's original error and task loss remain intact.

All sampled candidates together pass 36/1,280 audit case evaluations; none is audit-perfect. These are dependent, descriptive coverage counts, not a best-on-audit method. The final-condition audit counts are 11/320 for each policy cell and frozen greedy, 19/320 for context, 27/320 for public-best4 and 310/320 for search. Search retains nonzero list errors and is not a full-DSL optimality oracle.

Table 36: Actual one-step package costs. All physical 8×512 positions are charged, including dummy/padding. Search and model costs are not equal-FLOPs comparisons. Offline common-base training is a separate shared cost reported in Appendix H.

Quantity	Measured value
Model loads / generation requests	1 / 192
Generated tokens / replayed completions	6,591 / 192
Forward calls / physical token positions	13,247 / 54,259,712
Dummy and padding positions	51,203,864
Real / sham optimizer steps	3 / 3
Backward calls	6
Executor case calls / tool attempts	19,156 / 4,032
Configured symbolic candidates	3,680
Search timer proxy (microseconds)	83,124
Complete invocation wall time (seconds)	1,271.127
Worker wall time (seconds)	1,261.477
Peak CUDA allocated / reserved (bytes)	4,733,218,816 / 5,347,737,600
Final output bytes	52,397,862

The shared base and its adapter remain frozen and enabled. Policy, optimizer, RNG, typed matcher and memory restoration checks pass; model outputs after recovery match their originals. The energy cost was not measured. Limits bound the invocation rather than grant an efficiency claim. The paired intervals in Appendix L use all 16 requests; no zero-signal group, failed sample or weak family is discarded.

N INHERITED READOUT EVIDENCE AND ITS LIMITS

The coverage-controlled latent-readout study completed nine formal trajectories: a matched-presentation latent baseline, an added full-width linear bridge, and the same bridge with frozen text-teacher output alignment, each with three paired subjects. All used a fixed SFT lineage, oracle row selection, rule-based writes, four slots, 32 events, and eight prespecified greedy full-vocabulary first-word queries. Each of two task families contained 64 episode groups per subject. The criterion required all eight answers in an episode to be correct.

None of the latent conditions met its registered overall qualification. The text and restoration checks passed. The already-recorded paired bridge-minus-baseline difference was -0.02604 with a Bonferroni-adjusted 97.5% interval $[-0.09635, 0.04167]$; teacher-alignment-minus-bridge was -0.05729 with interval $[-0.11719, 0.00000]$. These intervals resampled whole episodes within family with all eight queries and fixed paired subjects retained; they are not uncertainty over an arbitrary seed or word population. The repeated text reference is not three independent text-model training runs.

The recorded cost was 9,792 formal and 192 engineering optimizer updates, 29,952 backward calls, and 62,064 completed generation requests. Additional teacher preparation makes these interventions unequal in FLOPs. The strong text teacher had a different decoder-adaptation history from the frozen latent student. Its success does not isolate representation space as the unique failure cause, and does not qualify a newly loaded public checkpoint on release programs.

Earlier fitted-support failures and later restricted generation/update/ restore successes coexist with these negative results. Different weights, prompts, words, and budgets do not form a single causal learning curve. These results motivate separating pure latent readout from the subsequent policy-adaptation and memory-admission studies.

Table 37: Actual completed coverage-controlled readout results. Entries are successful whole episodes for each paired subject, each out of 64. This is not a result of the new workflow or a $\Phi \times M$ experiment.

Intervention	Binding family	Preference family
Matched latent baseline	42, 43, 41	38, 45, 45
Added linear bridge	31, 41, 47	42, 41, 42
Bridge plus teacher alignment	24, 41, 36	38, 43, 40
Strong text reference	64, 64, 64	64, 64, 64

O PROSPECTIVE ADMISSION AND JOINT-EFFECT PROTOCOL

The following design remains separate from the completed one-step and multi-round policy comparisons. No prospective outcome is counted as a result.

O.1 AN ACTION-CONDITIONED ORACLE BEFORE A FUTURE-BLIND LEARNER

Fix Φ, reader, candidate, and memory snapshot at one event. For every legal slot and NULL, form its own successor. During the registered scoring horizon, later memory writes are disabled. For exogenous workflow observations, branches share external randomness but retain their own memory and generated actions. An interactive extension must use each action's own transitions; another branch's subsequent observations cannot be forced onto it.

Let F_{label} and F_{audit} be independent future sets under the same prefix. The teacher chooses

$$a^* = \arg\min_{a \in \mathcal{A}_t} \widehat{L}(a; F_{\text{label}}), \quad \Delta_{\text{audit}} = \widehat{L}(\text{NULL}; F_{\text{audit}}) - \widehat{L}(a^*; F_{\text{audit}}).$$

Ties prefer NULL, then fixed slot order. Each loss is computed after a real model read of that action's successor, not from future logits that were unaffected by the action. Report all actions, teacher calls, label performance, and independent audit performance. A proposed material threshold is 0.05 normalized task loss, fixed before the new model run. The selection and uncertainty rules must be finalized before confirmation.

Only after independent audit establishes useful room should a future-blind student Q be trained. Deployment features contain observable relevance, source information, occupancy, recency, and the proposed action; they exclude future identifiers, gold actions, future tokens, and hidden checks. Compare against current-risk scoring, no-write, write-all/recency, a strong relevance/trust heuristic, and a matched downstream-reward learner. The linear feature prototype is neither UMA nor A-MAC; it is a mechanism-matched adaptation whose limitations and teacher budget must be reported. Changing Φ also changes the value-estimation regime and requires checking Q's applicability. The engineering pilot has no learned Q. A later deployed Q is proposed as a frozen admission sub-block of the bounded numerical carrier Φ, alongside the online action sub-block. Both must be charged and restored; it is not an unaccounted fourth persistent adaptive state. This parameter layout remains to be locked before student deployment.

O.2 JOINT EFFECTS, HARM, AND CONFIRMATION

This unmeasured efficacy comparison includes learned admission and a strong simple joint comparator on the qualified common base. It is distinct from the earlier four-world engineering policy/memory switch study in Appendix G, and from feedback real/sham crossed with keep/reset policy at a fixed memory snapshot. Let L_{pm} be the complete-task loss with online policy adaptation $p \in \{0, 1\}$ and memory $m \in \{0, 1\}$. Report all four means, policy and memory main effects, and

$$\tau_{\text{interaction}} = \mathbb{E}[L_{11} - L_{10} - L_{01} + L_{00}].$$

On this loss scale, a joint improvement over either single arm is not sufficient for complementarity; the interaction must also be negative with relevant uncertainty. Compare the full method to a strong simple combination of reliable memory and ordinary policy learning. Shared weights, tokenizer, non-target state, and useful compute are essential to interpretation.

The same protocol must expose wrong and stale observations, occupancy pressure, rejected writes, tombstones, old-value leakage, unchanged facts, and the independent cost of restoration. Report final-task loss and success, per-request accuracy, first error, write rate, and recovery diagnostics. An all-correct long-trajectory score is informative but must not be the only metric if it drives every method to zero. Freeze one independent rule/task-family confirmation and one doubled-length or over-capacity profile. Neither short-run qualification nor a renamed template warrants an unrestricted generalization claim.

O.3 Confirmation planning

The analysis code provides grouped paired bootstrap intervals and the interaction contrast. A normal-approximation power calculation is a planning aid, not a guarantee. Required group counts must come from independent development variance, measured throughput, the 0.05 proposed material loss difference, and remaining time. The number and family of primary comparisons, timeout policy, integrity failures, and multiplicity adjustment must be frozen before confirmatory results. Registered algorithmic failures count as outcomes; infrastructure-invalid or incomplete groups are reported explicitly, not silently discarded.

P State and information accounting

The 512-byte row consists of a 73-byte typed header, a 128-byte bounded canonical value region, 279 bytes of zero padding, and a 32-byte checksum. The header includes a 16-byte namespace, two-byte slot, two eight-byte counters, four-byte event index, one-byte tombstone flag, 32-byte key, and two-byte payload length. Eight occupancy flags and a 32-byte bank header add 40 bytes to eight full row allocations. The physical serialization test checks 4,136 bytes; it does not establish authenticity.

Fixed-size active state does not make its Python object overhead, transient rendering, model activations, optimizer moments, or snapshots free. The output carrier has $r(H + V)$ float32 parameters for hidden width H, vocabulary V, and rank r; its gradient and optimizer moments must be counted in peak training memory. Internal factors instead sum $r(d_{\text{in}} + d_{\text{out}})$ over selected projections, also charging gradients and moments. This accounting is a new explicit engineering profile and does not resolve unrelated accounting or proof questions in a separately sealed long report.

The deployment model receives a public goal, explicit query keys, rendered active records, a bounded recent-observation view, and already available execution feedback. The evaluator alone owns expected semantics and hidden final cases. The assembler only sees completed observations. An oracle teacher may score private future cases in its training/development split, but may not send them or their identifiers to deployment features. Each branch stays inside its same-prefix group across splitting.

Q Reproducibility and status of materials

Q.1 Task construction and information available to the model

For the candidate-budget comparison, family order is arithmetic then list, template order is that of Table 1, and each template has four instances. The data seed is 20260917016. For construction index $i = 0, \ldots, 31$, the internal stream key is the string `b` followed by the two-digit zero-padded index. Each random stream is Python `random.Random` seeded by the integer SHA-256 digest of compact, sorted-key, UTF-8 JSON encoding of [seed, stream name, index string]. The stream names are `parameters`, `public` and `audit`; key and namespace streams are separate. These construction indices and target-derived grouping hashes are not model features.

Arithmetic parameters are drawn in the order scale, offset, with lower and upper then fixed to -16 and 16. List parameters are drawn in the order minimum, limit, suffix. Their task domains appear in the main text and are narrower than the legal DSL literal domain. Public arithmetic cases sample four integers without replacement; audit samples 20 from the remaining integers. List cases draw six independent integers per list, rejecting duplicate lists; audit additionally excludes all public lists. Separate streams and exclusion are both used. The original interpreter executes the private reference template with the drawn facts to form expected outputs during data construction. Evaluation checks

output behavior, not AST equality. The gold template, reference program, audit inputs/outputs and structural grouping never enter the prompt, grammar support, assembler or admission rule.

The model sees the behavioral goal, public DSL documentation, the four public input/output pairs, input type, and named values decoded losslessly from the requested row of actual memory. There is no fallback for a missing row. The current policy experiment has an empty recent-observation channel. The native checkpoint chat template serializes this user message with a generation prompt and thinking disabled; no answer prefix is supplied. The interpreter accepts 1–8 operations, integers in $[-128, 128]$, and take counts in $[0, 8]$. It rejects mistyped/unknown operations, invalid arity, reversed clip bounds, more than 16 intermediate list elements, or intermediate integers with magnitude above $2^{31} - 1$. Type support restricts primitive families but supplies neither a target ordering nor the correct literals.

Q.2 Public feedback, loop order and selected-output estimand

This subsection specifies the historical candidate-budget study. The latest feedback-resolution loop and added score field are specified in Appendix D; its fourth round has no update. For a program evaluated on n cases, feedback has exactly the fields `loss`, `passed`, `total`, `valid_program` and `error`. A valid program has loss $1 - \text{passed}/n$; any interpreter exception returns loss 1, passed 0, the full denominator n, validity false and an error string, even if earlier cases passed. Public reward is $1 -$ loss on four cases, so its possible values are $\{0, \frac{1}{4}, \frac{1}{2}, \frac{3}{4}, 1\}$. Audit uses the same loss definition on 20 held-out cases and cannot supply an update, stop decision or chosen output. A whole request is audit-perfect only if all 20 cases pass.

The next prompt contains the JSON feedback object of the earliest minimum-public-loss candidate in the *latest batch*. It contains neither that candidate's program text nor per-case errors/answers or a correction trajectory. The initial four examples remain visible. Selection for a reported endpoint instead uses the whole candidate pool through that round, breaking public-loss ties by earliest global generation index. Thus feedback packaging and endpoint selection use different, explicitly defined scopes.

Candidate-budget study: executed per-request loop.

1. Fix common weights, memory and request; initialize random A_0, zero B_0, empty optimizer and the same sampling RNG in all branches. Generate one initial greedy auxiliary output and eight shared samples.
2. For each branch and round $r = 1, \ldots, 4$, if already publicly closed, carry the selected endpoint forward. Otherwise use the shared batch at $r = 1$, or sample eight fresh completions $C_r \sim p_{\Phi_{r-1}}(\cdot \mid x_r)$ with that branch's weights and RNG.
3. Score the complete batch on the public examples and append it to the branch pool. A zero public loss closes the branch before updating. If still open, real/sham performs one LOO update or an exact constant-reward no-op; frozen performs neither. Sham uses $(r_2, \ldots, r_8, r_1)$. Save the branch's weights, optimizer, RNG and feedback.
4. At rounds 1, 2 and 4, record the earliest public-best candidate in the pool. For still-open branches, generate the registered post-update greedy auxiliary output with the newly available aggregate feedback.
5. Retain the separate first-four-sample anchors, context reference, round-3 attained-prefix probe, final keep/reset outputs, search result and restoration checks. Lock all outputs and choices before opening audit.

Initial samples are shared only while policy, prompt, grammar and RNG agree. Later samples are generated under each branch's own current policy and replayed at that same weight version, never reused for extra optimization epochs after a weight change. A constant reward vector skips both backward and optimizer; its optimizer state is unchanged. Within-request weights, moments and RNG persist; restoring an independent branch does not restore the cost ledger. First-round selected-output equality is guaranteed by the shared batch. Later equality is not imposed by sharing.

The primary contrasts compare complete online regimes: real versus frozen public-selected sampling and real versus cyclic sham at the final candidate budget. The keep/reset probe holds the final

non-policy input fixed and instead changes only the policy snapshot. It is a different estimand. The three-way primary loss equality does not assert program identity or equality of all auxiliary behavior.

Q.3 Configured search and two distinct admission rules

In the candidate-budget study, configured search enumerates Cartesian products in increasing length 1–4, allowing repeated primitives, in the fixed orders $(\texttt{mul}, \texttt{add}, \texttt{clip})$ and $(\texttt{filter_min}, \texttt{sort}, \texttt{take}, \texttt{append})$. Operands map respectively to scale, offset, (lower, upper), minimum, no operand, limit and suffix, all obtained from the actual resident row. The neural methods have the same public facts and examples but must generate their own operands; this hard-coded binding is a useful inductive bias of search, not exhaustive integer search. Missing facts produce no search program. Enumeration retains the first strictly lowest public loss; it does not stop at a zero public loss. Caps are 4,096 programs or 10 seconds per request. Hitting either returns the current best and marks search incomplete; the executed package requires completion rather than silently dropping it. With no truncation, $\sum_{\ell=1}^{4} 3^{\ell} = 120$ arithmetic and $\sum_{\ell=1}^{4} 4^{\ell} = 340$ list programs are scored per request: 7,360 formal programs plus 120 engineering programs, 7,480 total. Executor calls and search time are charged separately from neural work.

In the earlier frozen-base admission check, the relevance/source-trust rule admits a candidate only if its key is currently required and its source is a release. It reuses the existing key's slot, otherwise takes the lowest-index empty legal slot, otherwise replaces the oldest publication (slot index breaks ties). Write-all/FIFO uses the same placement without the relevance/trust filter. Current-risk uses that placement only for a required key whose resident value is missing, different or tombstoned, irrespective of source trust; no-write always selects NULL. All act within the assembler's legal set; NULL leaves the bank unchanged, whereas a tombstone is a stored deletion record.

The delayed-demand study uses a different rule and environment. Seven or nine public objects share eight slots, with the last object's release as the admission candidate. Its publicly known focus object has demand weight 8 and every other object weight 1 in the phase-predictable condition; unpredictable demand is uniform. Four label accesses and four audit accesses are drawn independently, with replacement, from these weights. Facts do not change and no later writes occur in that horizon. Future prompts contain explicit program semantics but no public examples, past request answers or recent-observation shortcut for a rejected row.

If an existing-key or empty slot is available, every writing baseline uses the existing or lowest-index empty slot. Otherwise FIFO removes the oldest publication, LRU the least recently requested resident key (unseen keys first), and slot index breaks ties. Frequency maximizes retained counts in the last eight public requests; public-demand maximizes retained demand weight; current-risk maximizes coverage of the current requested object. These latter three evaluate every legal successor including NULL, preferring NULL then lowest slot on ties. They receive no realized future access. For both admission studies, label pricing averages four own-successor future losses to select an action, with NULL-first/slot-order ties, before opening four independent audit futures. The first study allows exogenous subsequent observations and records reappearances; the delayed-demand study keeps facts fixed. Neither is a learned future-blind admission policy.

Q.4 Historical groups, offline recipe and rebinding

The state–program diagnostic used 16 independently seeded engineering worlds per family (seed 2026916), with attained event counts $4(i \bmod 8 + 1)$ for ordinal $i = 0, \ldots, 15$. The later qualification instead allocated parameter/rebinding pairs from finite public support, excluded the inventoried earlier tuples, and grouped all derived views with their source pair. Its arithmetic support included scale 4 from later releases; list minimum ranged from -5 to 5. Pairs changed offset or minimum at fixed other parameters; seeded ordering allocated engineering/train/development/qualification pairs before outputs. The qualification uses the first controlled view at six arrived events, not the earlier diagnostic's variable prefix lengths. Both retain the original prompt forms and syntax examples. The different before-adaptation counts (6/16 versus 13/16 arithmetic minimal; 8/16 versus 15/16 list minimal) are therefore measurements on different constructions, not a difficulty ranking or evidence of outcome-based selection. The adaptation comparison itself is paired within the same qualification groups.

The recorded offline recipe is a single fixed LoRA configuration: query/value projections in layers 16–23, rank 16, alpha 32, no dropout, learning rate 2×10^{-4}, AdamW betas $(0.9, 0.999)$, epsilon 10^{-8}, zero weight decay, gradient norm cap 1, seed 2026917 and eight epochs. Online initialization uses seed 2026915. No model/rank/learning-rate sweep or qualification-based checkpoint choice is documented. Preceding shape, numerical and evidence-collection repairs were engineering changes, not reported hyperparameter-search trials. The completed fit's costs are in Appendix H; a complete cumulative cost over earlier preparation and unsuccessful attempts was not recorded and is not assigned zero.

Rebinding follows the saved pair-level criterion: both programs must match their respective expected behavior on the registered public-rule-derived inequivalence witness. It is not a proof of equivalence on all inputs. Before-adaptation witness successes, reaggregated from the saved original pair records, are 6/8, 0/8, 7/8 and 0/8 in arithmetic/minimal, arithmetic/normal, list/minimal and list/normal order; after adaptation they are 8/8 in each. Before-adaptation *full-audit* success on both members is instead 5/8, 0/8, 7/8 and 0/8. The two criteria must not be inferred from one another or from marginal 16-state counts. No program was executed again to produce this summary.

Q.5 Recorded runtime and resource interpretation

Table 38: Non-sensitive runtime specification of the common-base fit and subsequent frozen-base, delayed-demand, one-step and multi-round studies. Versions come from their bound environment inventory, not a new machine probe. Requirements are distinguished from saved measurements.

Item	Bound record or verification limit
Accelerator	One RTX 4090; nominal contract 48 GiB. Launcher checks device name and at least 44 GiB free, with a 44 GiB per-process allocation limit. Exact total/free-byte snapshot not retained.
CPU / operating system	Four PyTorch intra-op threads in the launcher; Linux isolation. CPU model/count, inter-op setting, distribution/kernel version and contemporaneous contention were not retained.
Python / libraries	Python 3.11.15; PyTorch 2.11.0+cu128; Transformers 5.12.1; PyTorch CUDA build 12.8. Host driver version was not retained.
Model numerics	Native LlamaForCausalLM, eager attention, BF16 base; float32 online factors and LoRA parameters. Fixed physical 8×512 frame; native chat template, thinking off, no KV cache.
Generation / replay	Temperature 1, no top-k/top-p; maximum 128 answer tokens including EOS, logical context 512; current-weight grammar support and log-probability replay.
Online optimizer	Rank 4; random nonzero A, zero B; AdamW 10^{-4}, default betas $(0.9, 0.999)$ and epsilon 10^{-8}, zero weight decay; gradient norm cap 1; projection to $[-16, 16]$.

The earlier output-format and state–program diagnostics precede the fixed native-frame adaptation; their physical-token totals do not imply that they used the later 8×512 path. Hardware details not retained for an earlier study are not filled from this table. The fixed frame aligned sampling, strict-prefix replay and batch CE in the implemented BF16 path; it is an implementation/numerical choice, not a mathematical requirement of PDSA. For that path each forward charges 4,096 physical positions, including masked padding and dummy rows. Generated-token counts measure a different quantity. Device capacity is also distinct from peak allocated/reserved CUDA bytes. Whole-invocation wall time, search timers and per-request ledger deltas cannot be converted into unmeasured per-method GPU time, energy or equal-FLOPs efficiency.

Q.6 A conditional first-update argmax bound

This is an explanatory consequence of the standard optimizer, not a measured cause or a new general theorem. Assume exact arithmetic, the first actual bias-corrected AdamW step from empty first/second moments, $B_0 = 0$, zero weight decay, no extra residual scaling, unchanged A_0, and no active parameter projection. Let g be the gradient actually used by the step, after any gradient

clipping. Coordinatewise,

$$\Delta B_{ij} = -\eta \frac{g_{ij}}{|g_{ij}| + \epsilon}, \qquad |\Delta B_{ij}| \le \eta.$$

This follows because the bias-corrected first and second moments on the first step are g and g^2. At fixed visible prefix s, let $u(s) = A_0 h_{\bar{\theta}}(s)$. Then

$$\|\Delta z(s)\|_\infty = \|\Delta B\, u(s)\|_\infty \le \eta \|u(s)\|_1.$$

Indeed, the absolute change in each logit is at most the sum of the absolute coordinate products. If the legal support is unchanged and the original top-one/top-two logit gap is $m(s)$, the gap can shrink by at most $2\eta\|u(s)\|_1$. Therefore

$$m(s) > 2\eta \|u(s)\|_1$$

is a sufficient condition for the argmax at this prefix to remain unchanged. A singleton legal support is unchanged trivially. Failure of the inequality does not guarantee a flip. Keeping an entire greedy sequence requires the appropriate condition along its prefix chain. Neither a single TV value nor $\eta = 10^{-4}$ alone provides the missing feature scale and margin. The actual mixed-precision forward, floating-point rounding, projection and later steps with momentum require their own analysis; the first-step bound is not automatically applicable to the multi-round trajectory. No stored scale/margin measurement here identifies a unique reason for the unchanged primary outcomes.

Q.7 Material availability

The executed checkpoint is the public `MiniCPM5-1B-SFT`. The pinned revision is `a60b37f1fc409c54e1e337b0723aaac6f92dfec0`. It is loaded locally with remote code and network fetching disabled. Qualification of current-fact grounding is not automatically qualification of compositional repair or another task family. The completed invocations bound an environment inventory, package versions, weight hashes and finite resource contract. This does not guarantee portability to a different numerical environment.

The source covers memory, DSL generation, feedback updates, independent restores and analysis; new-task-family confirmation remains outstanding. The LaTeX bundle rebuilds the document. The size-limited companion supplies the fitted common adapter, initial output policy and all 26 saved states matching the registered initial internal-policy parameters. It includes inputs, allocation, source, configuration and records for twelve studies, and inventories 106 omitted intermediate-checkpoint or calibration-cache files. The common adapter's 688,128 parameters in 32 FP32 tensors occupy 2,752,512 parameter bytes; its serialized file is larger. Safe CPU inspection verified tensor identities without constructing a model. Each snapshot's sampling RNG denotes its saved point, not a universal initial state. No weights or inputs were regenerated.

The core study's complete program texts, logical memberships, original public/audit cases and aggregate audit scores are included. Audit data and reference material remain separate from the public-only model input interface. Saved per-case public outcomes and module gradient summaries are provided where recorded; missing individual audit-case flags and global gradient norms are not reconstructed. File/static and saved-data checks validate identities, original public winners, saved-score minima and all 384 decomposition rows without a model, interpreter or statistical resampling. The asset index identifies older inherited-readout raw materials outside this export; this is not a full-paper independent reproduction.

Post-hoc pool scoring is complete; the later raw-pool recheck and fixed-skeleton parameter rebinding are unrun. Historical method and runner source is supplied, but the core comparison API remains inspection-only: neither a standalone neural rerun nor raw-program replay from this package has been validated. Foundation-model weights are obtained separately from the fixed source. Historical environment metadata is included as recorded, with missing fields left unknown. Bitwise neural sampling/training reproducibility on arbitrary hardware is not established. Table 25 summarizes these limits.

The historical table was checked against the existing diagnostic summary and its matching terminal record. No evaluation was rerun and no historical threshold was changed. The authoritative private mapping retains original identifiers, source commits, raw-artifact hashes, and descriptive public study names without renaming raw experiments. The first-stage sealed report remains a separate evidence snapshot.