WHAT PRETRAINING AND MIDTRAINING MAKE LEARNABLE FROM REWARDS?

Anonymous authors
Paper under double-blind review

ABSTRACT

A reward can identify a correct answer while leaving the computation needed for new inputs undetermined. We study what pretraining and midtraining contribute to this gap. In sequential state computation and contextual memory, we characterize mechanisms that agree on every training reward yet disagree on held-out answers. Task-independent source observations resolve this ambiguity. We then construct finite sampled-Adam paths from random initialization through source prediction and reward adaptation in the same parameters: prediction acquires execution or retrieval, and rewards learn how to use it. Experiments with pretrained Qwen2.5 checkpoints measure this division of labor through source midtraining. With first-operation source supervision, Sequential models acquire execution and show a strong reward-task-selection advantage over a private-random source control. Memory experiments show retrieval retention and task selection under an independent source intervention. Public tasks separate accuracy at reward entry, subsequent gain and final performance. Together, these results explain how source information becomes computation that rewards can use.

1 INTRODUCTION

Reinforcement learning can substantially improve reasoning in pretrained language models, as demonstrated by DeepSeek-R1 (DeepSeek-AI et al., 2025). What makes a model ready to learn from rewards? Earlier prediction stages shape downstream performance: domain- and task-adaptive training improve classification (Gururangan et al., 2020), and agent-native midtraining strengthens subsequent software-engineering training (Zeng et al., 2026). To choose training data and interpret these gains, we need to understand what each stage contributes. Prediction may supply information absent from rewards, build a computation that uses it, or already teach part of the task. These possibilities lead to different explanations of the same final score. We ask which information and computations make a new task rule learnable from rewards.

A terminal reward evaluates an answer without directly observing its intermediate mechanism. Consider a learner that composes unknown operations on a hidden state. Two transition rules can agree on every training reward yet demand opposite held-out answers: their difference cancels along training sequences and reappears on test sequences. More reward queries restricted to that support cannot distinguish them; observing individual transitions can. We study such *source prediction* during initial pretraining or *midtraining* of an existing model, followed by reward adaptation of the same parameters (Figure 1). A *world* pairs a reusable mechanism with a task rule; our theoretical source laws reveal the former while leaving the latter unspecified. Source prediction trains composition or retrieval; reward training then learns a task-specific use of that computation.

Exactly what rewards leave open. Prior theory explains how RL exploits rare task demonstrations in prediction data (Tsilivis et al., 2025); controlled experiments also show RL composing pretrained primitives into reusable strategies (Abdulsalam et al., 2026). We characterize what source observations must add when training rewards leave the mechanism ambiguous. In our Sequential family, a two-parameter family of transformations preserves all training rewards, while half reverse every held-out answer (Theorem 4.1). Local source observations distinguish these possibilities. Contextual memory gives a second exact characterization: after finitely many numerical source records, unobserved contexts impose an exact error floor that additional rewards cannot lower (Theorem 5.1).

A finite learning path through the same parameters. From specified random initializations, we construct finite sampled Adam paths through both stages with high-probability guarantees. Sequential source prediction teaches a transformer to execute transitions on its own generated prefixes; two

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

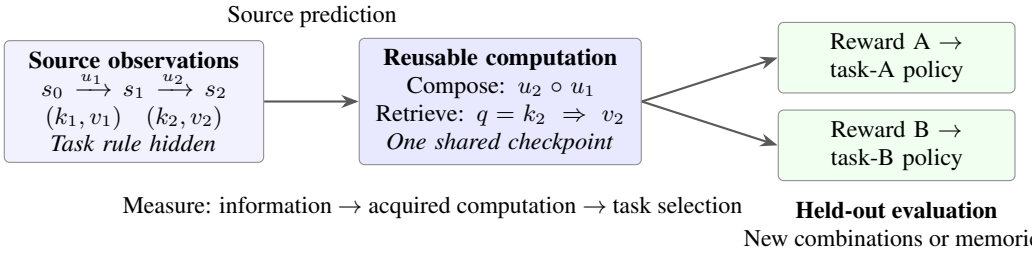


Figure 1: Source prediction prepares a shared computation for reward adaptation. States s_i and operations u_i illustrate composition; keys k_i, values v_i and query q illustrate retrieval. Rewards select task-specific use.

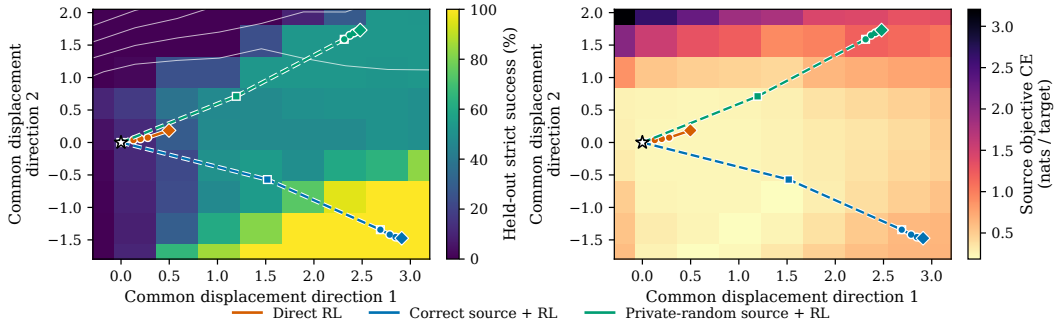


Figure 2: Measured Sequential parameter slice: task success (left) and source loss (right). Colors evaluate models in a common SVD plane; paths project actual checkpoints and need not share their scores. Star: shared start; dashed squares: source; solid circles: RL; diamonds: endpoints. White contours interpolate source loss. Evaluation and reconstruction checks: Section 6.3 and Figure 15.

REINFORCE updates using $2B_R$ terminal verifications then select the initial operation, giving held-out success at least $7/8$ (Theorem 4.2). In contextual memory, source Adam acquires retrieval and reward Adam learns an input-dependent rule, reaching at least $19/20$ within explicit budgets; freezing randomly initialized query/key attention under the same schedule keeps success at most $1/64$ above chance (Theorem 5.2). Widths, batches, step sizes and budgets are chosen without test information.

Measuring acquisition and task-specific use. Our Qwen2.5 experiments apply source midtraining to pretrained checkpoints and measure acquisition, retention and task-specific use separately. Across eight matched worlds, correct-source models acquire Sequential execution and Memory retrieval; action-target Sequential training reaches 82.61% success versus 44.15% for a private-random source control (Figures 3 and 4). Source-checkpoint probes reveal the computation acquired before rewards; paired reward branches test its use for opposite tasks. Memory replay preserves retrieval; an independent eight-world confirmation reaches 75.32% task success versus 49.86% with alternative retrieval (Section C.2). An independent single-world comparison shows correct-source preparation enabling rapid reward learning from zero task success at reward entry (Figure 15). The measured parameter slice in Figure 2 complements these actual-checkpoint curves by displaying source loss and task success across a common plane. On GSM8K and HotpotQA, accuracy at reward entry, reward-stage gain and final accuracy reveal how preparation changes subsequent learning.

Contributions.

- Exact information characterizations: which mechanisms training rewards leave indistinguishable, and how source observations resolve the ambiguity (Sections 4 and 5).
- Finite sampled-Adam paths that turn source observations into execution or attention and then into task learning, with a frozen-attention comparison isolating the role of acquired retrieval (Theorems 4.2 and 5.2).
- Measurements that distinguish acquiring a computation, retaining it and learning its task-specific use in Qwen2.5, together with the contribution of each training stage to public-task accuracy (Section 6). Section 2 connects the results to prior work; Section B develops their implications.

2 Related Work

Information supplied by observations and feedback. Learning an unknown mechanism depends on the observations and queries available to the learner. Query learning identifies automata through membership tests and counterexamples (Angluin, 1987; Weiss et al., 2019), while side information can resolve ambiguities left by relational measurements (Foster et al., 2018). A terminal reward oracle (a verifier), an expert label stream, and a supplied specification therefore define different learning problems (Balcan et al., 2026; Alshiekh et al., 2018). We characterize transformations of a mechanism and its task rule that preserve every training reward yet reverse held-out answers. Source observations distinguish these reward-equivalent worlds. The obstruction is informational: additional rewards from the same training support cannot remove it, independently of the optimizer. Extended comparisons appear in Section A.

Learning across source and adaptation stages. Intermediate supervised training and domain-adaptive prediction improve downstream learning under different information interfaces (Phang et al., 2018; Gururangan et al., 2020); agent-native midtraining uses software-engineering trajectories (Zeng et al., 2026). Tsilivis et al. (2025) analyze prediction followed by reinforcement learning on short and long demonstrations of a parity task, using full-chain correctness in their learning theorem. Wei & Kim (2026) study graph world-model pretraining and outcome-reward adaptation through gradient-flow analyses. Our source data reveal reusable transitions or retrieval structure while leaving the downstream task rule unspecified. The learning results follow finite sampled updates through both stages and evaluate new operation combinations or memory configurations. Empirical primitive-to-composition training (Abdulsalam et al., 2026) motivates a related distinction between exposure to components and learning to use them. Our experiments test source acquisition, execution and reward-dependent selection separately.

Acquisition and reuse of representations. Prediction can provide downstream features under suitable source–target relationships (Saunshi et al., 2020). Ren et al. (2024) show that correlated pretrained transitions can bypass an initially sample-intensive learning stage. Attention-learning analyses explain how data and optimization produce particular retrieval or traversal strategies (Ran-Milo et al., 2026); later fine-tuning can also distort useful features (Kumar et al., 2022). We derive the source-trained computation and track its use during reward updates. Both constructions follow finite sampled Adam through both stages: the sequential result controls execution on generated histories, and the contextual result follows acquired attention and head factors. These results connect the information in source observations to a representation that makes a new reward-defined rule learnable. The full comparison, including initialization, update algorithms and guarantee objects, appears in Appendix A.

3 Framework and Preliminaries

A world and its train/test split are fixed before sampling; the learner and initialization are chosen independently of that world. The learner first predicts task-independent source records, then adapts the same parameters with terminal rewards on fresh training prompts; it receives neither test feedback nor intermediate correctness labels. The neural results use bias-corrected Adam (Kingma & Ba, 2015), retaining parameters and resetting optimizer moments at the stage boundary. Exact objectives, updates and resource counts are in Theorems H.1 and I.1; the SGD path is in Section J.1.

The source stopping rule and reward update count are fixed before training; no test prompt influences stopping or checkpoint selection. We count source tokens or records, reward queries, generated actions and updates separately; the *verifier* returns the terminal reward.

Sequential mechanism. Over $\mathbb{F}_2$, let $Z = \mathbb{F}_2^{2r+1}$, $r \geq 1$, $m = |Z|$ and $q = 2m$. A state is (s, z) and a control in $\mathcal{U}$ is (b, u), with public coordinates z, u and private bit s. An unknown $m \times m$ table θ gives

$$\phi_{(b,u)}(s, z) = (s + b + \theta_{u,z}, z + u). \tag{3.1}$$

The alphabet contains disjoint state and control tokens plus PAD and EOS; their grammar and index map are known, while neural embeddings are randomly initialized. Write $u = (x, y, t)$ with $x, y \in \mathbb{F}_2^r$ and $t \in \mathbb{F}_2$, and $\ell(u) = t + x \cdot y \in \mathbb{F}_2$. In an H-operation task, the initial state is uniform and the displayed suffix $U_2, \ldots, U_H$, with $U_i = (b_i, u_i)$, is uniform conditional on $\sum_{i=2}^{H} \ell(u_i) = 0$ in training and one in test. The unknown first operation $\eta = (b_\eta, u_\eta)$ is the *task binding*: it specifies which use of the transition mechanism earns reward. A prompt $X = (S_0, U_2, \ldots, U_H)$ has target $y^*(X) = \phi_{U_H} \circ \cdots \circ \phi_{U_2} \circ \phi_\eta(S_0)$.

Source records expose single transitions in the order $S_0, U_2, \ldots, U_H, U_1, S_1, \ldots, S_H, \mathrm{EOS}$, with uniform S_0, an independent uniform U_1, a training-tag suffix, and states evolving in execution order under $T_U = aP_U + (1-a)\Pi$. Here P_U is the permutation kernel of ϕ_U, Π resets to a uniform state, and $a = 1 - c/H$ with $c = 1/64$ for Adam. Source prediction sees no downstream binding or reward. During adaptation the policy generates a first control, H states and EOS; the verifier checks grammar, public evolution and the final target, allowing a self-repaired private history. For a prompt X, $E(w, X)$ is valid endpoint success, $S(w, X)$ endpoint success with the correct first control, $M(w, X) = \Pr(A \in \mathcal{U}, Y_1 = \phi_\eta(S_0))$ the probability of the task-correct first state, with A the generated control and Y_1 the first generated state, and $V(w, X)$ a grammar or public-path violation. These are free-generation probabilities, and $S \le E$. Full distributions, verifier semantics and budgets are given in Theorems E.1 and H.1.

4 Information and a Finite Neural Learning Path

4.1 Rewards Leave an Exact Ambiguity

Fix u_η. Two worlds are *reward-equivalent* if every grammar-valid training candidate receives the same reward in both. Equivalence concerns every allowed reward query, so it constrains any learning algorithm.

Theorem 4.1 (Sequential reward equivalence; informal version of Theorem E.1)**.** *For $H \ge 3$, two worlds are reward-equivalent if and only if some $\kappa, d \in \mathbb{F}_2$ satisfy*

$$\theta'_{u,z} = \theta_{u,z} + \kappa + d\ell(u), \qquad b'_\eta = b_\eta + (H \bmod 2)\kappa + d\ell(u_\eta). \tag{4.1}$$

Their targets differ on every test prompt exactly when $d = 1$.

The κ shift changes every transition bit; d changes just tag-one operations. Even-tag training suffixes cancel the latter shift after the binding adjustment, whereas odd-tag test suffixes expose it. Conversely, short paths padded with zero controls force all tag-zero and all tag-one differences to agree, so (4.1) is exhaustive (Section E.2.2). On a reward-equivalent pair with $d = 1$, any reward-only learner has average test success at most $1/2$, regardless of its training-query budget.

Source records expose transitions but leave η unspecified. Let the initial observation O include source data and inherited weights. It records the world information available before reward learning. For a *known pair* differing by $d = 1$, let α be the overlap of the two laws of O. If each fresh reward prompt has odd parity with probability ε, the optimal equal-prior test error is exactly $\alpha(1-\varepsilon)^Q/2$ (Proposition E.3). Queries may adapt their candidates; prompts remain exogenous and charged. Thus rare informative prompts replace the unlimited-query obstruction by a finite-exposure barrier.

Three operations make the gap visible. Take $r = 1$, $H = 3$, $\theta = 0$, $\eta = (0, (0, 0, 0))$ and $\theta'_{u,z} = \ell(u)$. Starting from private bit zero with $b = 0$, the original mechanism remains at zero. Under θ', a suffix with tags $(1, 1)$ produces private bits $(0, 1, 0)$ after the three operations, so the training endpoint agrees; tags $(1, 0)$ produce $(0, 1, 1)$, reversing the test endpoint. Public coordinates are identical. A tag-one source transition has next-bit-one probability $(1-a)/2$ versus $(1+a)/2$ in the two worlds. Source records therefore reveal the private transition before the training-word cancellation.

At fixed confidence and $H = \Theta(\log q)$, a symbolic route trades $O(\log q)$ source tokens for $\Theta(q^2)$ queries. Table recovery instead leaves at most $q - 1$ binding tests; Section F gives the assumptions and computation costs.

4.2 Source Execution Enables Reward Learning

Executing a learned transition requires producing the right kind of token: a control, a state or EOS. A predictor can have accurate conditional transition bits yet emit EOS almost surely on a generated suffix (Theorem H.6). The source objective therefore includes every token position and a positive ridge penalty. Its canonical optimum controls the token-class probabilities together with the transition logits, yielding an executable policy on its own generated prefixes.

The neural learner uses causal attention and a bilinear feed-forward readout on full padded prefixes. Here e_l is the token embedding at lag l, e_0 the current-position embedding, R the number of heads, and $b_h(l)$ a learned positional bias, t_{att} the fixed temperature, and d_h the query/key width; $\odot$ denotes

elementwise multiplication. Its attention weights and logits are

$$\alpha_h(l) = \mathrm{softmax}_l\left(b_h(l)/t_{\mathrm{att}} + \langle Q_h e_0, K_h e_l\rangle/\sqrt{d_h}\right), \tag{4.2}$$

$$z_{\mathrm{att}} = e_0 + R^{-1/2}\sum_h O_h \sum_l \alpha_h(l) V_h e_l, \tag{4.3}$$

$$f_w(X) = W\left[z_{\mathrm{att}} + U((Bz_{\mathrm{att}} + b)\odot(Cz_{\mathrm{att}} + c)) + b_{\mathrm{out}}\right]. \tag{4.4}$$

The maps Q_h, K_h, V_h, O_h form each head; B, C, U, W and their biases give the bilinear readout. All embeddings, attention blocks, feed-forward blocks and output rows train from random initialization. The theorem also covers the specified normalized-readout member; dimensions, rates and its exact map are in Sections G.3 and H.5. Source next-token cross-entropy includes every target and a coupled positive L_2 penalty; an observable fresh-batch gradient criterion stops source Adam. Adam uses $(\beta_1, \beta_2) = (0.9, 0.999)$. After removing the penalty, retaining all parameters and resetting moments, two fresh-batch, on-policy REINFORCE–Adam updates use terminal reward, baseline zero and every sampled trajectory, including zero-reward and malformed ones.

Theorem 4.2 (Finite sampled-Adam learning; informal version of Theorem H.1)**.** *Fix $q = 2^{2r+2}$, $r \geq 1$, source reset scale $c = 1/64$, $\delta \in (0,1)$ and either network member above. With the explicit world-independent horizon, widths, batches, rates and stopping rule in Theorem H.1, run source Adam followed by exactly two reward updates of B_R trajectories each. For every fixed world, with probability at least $1-\delta$, source stopping succeeds and, simultaneously for every held-out X, the final iterate w_f and source iterate w_s obey*

$$\min_{J\in\{E,M,S\}} J(w_f, X) \geq \tfrac{7}{8}, \quad \max_{J\in\{E,M,S\}} J(w_s, X) \leq \tfrac{1}{q} + 2(H+2)\varepsilon_{\mathrm{logit}},$$
$$\min_{J\in\{E,M,S\}} [J(w_f, X) - J(w_s, X)] \geq \tfrac{3}{4}. \tag{4.5}$$

Also $V(w_f, X) \leq 9c/8 + 1/128$, with $Q = 2B_R$ reward queries. Here $\varepsilon_{\mathrm{logit}}$ is the prescribed uniform logit tolerance. The full resource vector is in Theorem H.1.

At source stopping, the model executes a supplied operation but guesses the task binding, giving success of order $1/q$. Reward samples distinguish the correct binding from wrong public paths and opposite private bits; two updates concentrate on the correct choice. Establishing this for Adam requires tracking the learned executor during those updates: the momentum and adaptive denominator depend on the same sampled trajectories that move the backbone. The proof couples the full sampled path to a reference binding-selection path, with explicit bounds for this dependence, malformed outputs and backbone motion. Thus the reward guarantee uses the computation acquired by source training in the same moving parameters (Section H.5).

A separate finite SGD path appears in Theorem J.1.

5 Learning the Representation That Rewards Use

Sequential rewards select one global first operation after source training has learned execution. A contextual task makes the rule depend on the input: reward learning needs a representation that retrieves the relevant memory. We again identify the missing information, then follow finite Adam updates through retrieval acquisition and task learning.

5.1 Contextual Memory and Its Observation Channels

There are $n_{\mathrm{ctx}} \geq 1$ contexts and two memory slots. With odd content dimension $p \geq 17$ and cue dimension $k = p - 2$, a world has a retrieval orientation $U_\ell \in \{I_2, P_{\mathrm{swap}}\}$ and contextual sign $s_\ell \in \{\pm 1\}$ for each context, with $U = \mathrm{diag}(U_1, \ldots, U_{n_{\mathrm{ctx}}})$, plus a shared cue vector $\tau \in \{\pm 1\}^k$.

More explicitly, a source record draws a uniform context ℓ, an independent Haar-uniform frame $R \in O(2)$ and independent uniform selected slot $J \in \{1,2\}$. Its keys are $\kappa_j = e_\ell \otimes Re_j$, query $x = U\kappa_J$, and values $V_j = (0_{n_{\mathrm{ctx}}}, G_j)$ with independent $G_j \sim N(0, I_p)$; the prediction target is $\mathrm{sign}(G_{J,1}G_{J,2})$. Random frames prevent a fixed slot shortcut. Only this future bit contributes source cross-entropy, while source matching is the model's probability of the correct slot. The downstream context signs and cue rule never enter these records.

A training reward prompt draws independent uniform context ℓ, query index $q \in \{1,2\}$ and cue $z \in \{\pm 1\}^k$. It has keys $e_\ell \otimes e_1, e_\ell \otimes e_2$, query $e_\ell \otimes e_q$, and opposing cue values $(e_\ell, 0, 0, \pm z)$. If

$U_\ell e_{j_\star} = e_q$, the correct action is $s_\ell(-1)^{j_\star - 1}\,\mathrm{sign}(\tau^\top z)$; the learner samples one action and observes only correctness. Test memories replace the training amplitudes $(1, -1)$ by independent uniform amplitudes $u_1 \in \{\pm 1\}$, $u_2 \in \{\pm \rho_{\mathrm{mem}}\}$, with $\rho_{\mathrm{mem}} = 1 - 2^{-\lceil \log_2(2\sqrt{k})\rceil}$. Thus every test memory configuration is new. Its label is $s_\ell\,\mathrm{sign}(u_{j_\star})\,\mathrm{sign}(\tau^\top z)$. Full distributions and numerical precision are in Theorems E.5 and I.1.

Let $\xi_\ell = 1$ for identity and -1 for swap. Training labels factor as $y_{\mathrm{train}} = s_\ell \xi_\ell v_q\,\mathrm{sign}(\tau^\top z)$ with $v_1 = 1, v_2 = -1$: rewards see the product $s_\ell \xi_\ell$, while the source reveals ξ_ℓ.

Theorem 5.1 (Contextual information; informal version of Theorem E.5)**.** *Two worlds are reward-equivalent on every training input exactly when, for a common $\varepsilon \in \{\pm 1\}$,*

$$\tau' = \varepsilon\tau, \qquad s'_\ell \xi'_\ell = \varepsilon s_\ell \xi_\ell \quad (\ell = 1, \dots, n_{\mathrm{ctx}}). \tag{5.1}$$

Orientations may differ independently, and flipping one reverses the equal-sign test-amplitude labels in its context. With N_{obs} iid records from the exact numerical source, taking the infimum over world-independent learners and all finite fixed iid reward-query budgets, the worst-world test error is

$$\mathcal{R}_{N_{\mathrm{obs}}} = \tfrac{1}{4}(1 - 1/n_{\mathrm{ctx}})^{N_{\mathrm{obs}}}. \tag{5.2}$$

For $n_{\mathrm{ctx}} = 1$, the power factor is one at zero records and zero after one or more. Source records alone leave minimax error $1/2$.

One source record identifies its context's orientation almost surely; a missed context has probability $(1 - 1/n_{\mathrm{ctx}})^{N_{\mathrm{obs}}}$, and the hidden orientation reverses labels on the equal-sign half of its test configurations. On that half, indistinguishable orientations force error $1/2$, giving the factor $\frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$. Reward labels identify the remaining cue rule up to its harmless common sign. A finite feedback table approaches the floor with an explicit query count (Theorem E.5). The finite-alphabet optimization theorem below has a distinct observation model; (5.2) applies to exact numerical source records.

5.2 Finite retrieval acquisition and task learning

The trainable model has factorized attention core M and bilinear head core C, initialized with independent Gaussian row factors. For query x, keys κ_j and values V_j it computes

$$\pi_j = \mathrm{softmax}_j(x^\top M \kappa_j), \quad h = \sum_{j=1}^{2} \pi_j V_j, \quad g = h\sqrt{\psi(\|h\|^2)}, \quad f = g^\top C g. \tag{5.3}$$

The policy takes action $+1$ with probability $\sigma(f) = (1 + e^{-f})^{-1}$. Here $M = d_{\mathrm{width}}^{-1} \sum_i q_i k_i^\top$, $C = d_{\mathrm{width}}^{-1} \sum_i o_i a_i b_i^\top$; every factor trains. The fixed normalization class is

$$\psi(u) = \sum_{j=1}^{m_{\mathrm{norm}}} \omega_j (u + \zeta_j)^{-\alpha_j}, \quad \omega_j > 0,\ \sum_j \omega_j = 1,\ 0 \le \alpha_j \le 1,\ 0 \le \zeta_j \le Z_{\max} < \infty. \tag{5.4}$$

Source cross-entropy predicts the selected Gaussian bit. Reward Adam uses sampled correctness with baseline zero, retaining zero-reward actions in the batch; parameters persist between stages while moments reset. The normalized query/key and head factors use the specified positive steps and denominator offsets in Theorem I.1.

Theorem 5.2 (Acquisition and transfer; informal version of Theorem I.1)**.** *Fix the task dimensions, any architecture in (5.4), a world (U, s, τ), $\delta \in (0, 1)$ and Adam parameters $0 \le \beta_1 < 1$, $0 < \beta_2 < 1$, $\beta_1^2 < \beta_2$. Choose width, batches, steps, offsets and input precision as in Theorem I.1. With probability at least $1 - \delta$ jointly: (i) correct-slot matching rises from at most 0.51 to at least 0.9 uniformly over contexts, frames and queries, and greedy source-bit accuracy exceeds 0.9; (ii) after reward training, sampled success is at least $19/20$ on every held-out context, query, cue and amplitude pair, at least $2/5$ above the source checkpoint; (iii) under the same schedule with initial query/key factors frozen in both stages but every head factor trained, sampled test success is at most $1/2 + 1/64$, with a paired gap greater than $2/5$. The guarantee also covers the specified finite coordinate alphabet with numerical decoding and real-valued optimizer states.*

The head receives its task information through attention-pooled values. Query/key rows form M and head rows form C; every supplied coordinate trains at a nonzero rate. At exactly uniform attention, opposing cue values cancel; source-learned alignment exposes the selected cue for transferable task learning. At source entry, random query/key factors give matching near chance; supervised prediction of the selected Gaussian sign creates positive head alignment and then a persistent drift in attention.

From source prediction to retrieval. Along the source reference path $M = \beta U$, positive head alignment $t > 0$ with the source bit gives $\partial_\beta L(t, \beta) < 0$ for every normalization member. Predicting the selected value therefore increases attention to its slot. The analysis first establishes that head alignment from random initialization, then controls the sampled factor dynamics around the reference path. Supplying an already aligned representation would bypass this acquisition step.

From retrieval to a contextual rule. Along the reward reference path, updates build context–cue coefficients proportional to $s_\ell \tau_i$ in C. With attention contrast $c > 0$, the pooled cue carries a direction for this update; at exactly uniform retrieval ($c = 0$), opposing cue values cancel. The reward reference has effective scale $w(c) = c\psi(1 + kc^2)$. Its boundary $c \geq 1/\sqrt{k}$ is preserved by these reference dynamics for every normalization mixture, even when further sharpening reduces $w(c)$ (Proposition I.2). The resulting context–cue interactions change the policy with the input, whereas a constant action remains at chance.

The frozen control retains its small random initial attention contrast while its head trains under the same finite schedule. The proof bounds the small initial contrast and its possible amplification by the head, yielding the stated near-chance upper bound. Sampled-path comparisons for both branches also control momentum, gradient–denominator dependence and source errors carried into reward training. Full width, precision and budget formulas are in Sections I.7 and I.8.

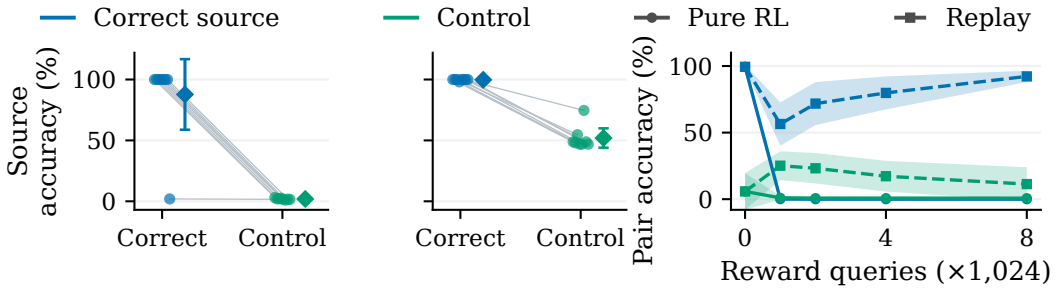
6 Experiments

We measure acquisition, task selection and retention under full-parameter source midtraining of pretrained Qwen2.5 models (Qwen et al., 2024). Evaluation is greedy; reward training samples at temperature one in the original experiments, with calibrated sampling in the independent Memory confirmation (Section C.2). World–seed pairs are independent units: A/B branches are averaged within world before computing 95% Student-t intervals and paired effects. Methods and complete comparisons appear in Sections C.1 and C.2; earlier experimental designs are described separately (Sections C.4 and C.13).

6.1 What computation does source training acquire?

Eight worlds per module compare correct source with a structural control on Qwen2.5-0.5B, after task-independent format preparation. Sequential uses 2^{22} observed tokens; controls randomize private bits while preserving public states. Objectives predict states and END/EOS, or additionally the record's random first operation, independent of downstream binding. Adding targets changes supervision, not observed tokens. Memory uses 2^{24} tokens, an all-token objective, and correct versus wrong retrieval on matched inputs.

With the first operation supplied, action-objective Sequential models generate 87.74% complete source trajectories versus 1.93% for controls (Figure 3a). Memory first-field accuracy is 99.76% versus 51.95%; joint original/value-flip accuracy is 99.41% versus 5.81%. Requiring both answers to be correct tests intervention-sensitive predictions without supplying a downstream rule; syntax or a constant bit cannot suffice.



(a) Sequential source (b) Memory source (c) Memory retention

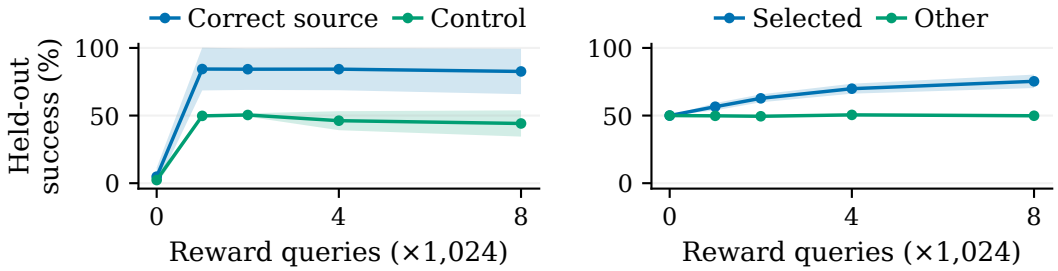
Figure 3: Source acquisition and retention across eight worlds. (a) Complete trajectories on 512 probes with first operation supplied (action objective). (b) First-field accuracy on 256 Memory records. Lines pair worlds; diamonds show means and 95% intervals. (c) Accuracy on 256 original/flip pairs, A/B averaged within world, with untruncated 95% bands. Replay adds 131,072 targets and 128 CE updates.

6.2 Do rewards select and retain the computation?

The original source checkpoints branch into opposite tasks A/B for 8,192 reward queries, with fresh Adam moments and a fixed learning rate 3×10^{-6}. Both verifiers score each output on the same 1,024 held-out evaluation records per world. Branches differ only in Sequential's private binding bit or all Memory context signs; mechanisms and prompt distributions agree. If J_{XY} is success of the task-X branch under verifier Y, task selectivity is $\Delta_{\text{task}} = \{(J_{AA} - J_{AB}) + (J_{BB} - J_{BA})\}/2$. Because the branches share a source checkpoint, this contrast measures whether changing the reward changes which task the model solves.

Sequential: source computation supports task selection. With first-operation supervision, correct-source success rises from 4.83% to 82.61% (Figure 4a), exceeding the control endpoint (Table 1). The predeclared task-selectivity advantage is 67.57 pp, interval $[37.04, 98.10]$, with two-test Holm-adjusted exact $p = .015625$ (Figure 13b). All sixteen branches per route make 128 nonzero reward updates at equal query budgets.

First-operation supervision targets the output position that begins a free rollout, without supplying a preferred task binding. The primary contrast tests source structure within this objective. The source-by-objective selectivity interaction is a separate, predeclared secondary comparison: 42.59 pp, t interval $[0.82, 84.36]$, exact $p = .08594$ (Table 4). The exact test leaves this secondary interaction unresolved.



(a) Sequential (b) Memory confirmation

Figure 4: Source preparation and reward learning. (a) Sequential action-objective routes, worlds 500–507. (b) Independent Memory confirmation, worlds 508–515: selected versus alternative retrieval after a shared correct-source start. Success uses 1,024 fixed held-out records; A/B are averaged within each world, with 95% intervals across eight worlds. Source costs differ between panels (Sections C.1 and C.2).

Memory: retaining and using retrieval. Source replay preserves the computation through reward adaptation: at 8,192 queries, correct-source replay retains 92.21% original/flip-pair accuracy versus 0% under pure RL (Figure 3c), with first-field accuracy 96.24% versus 49.66%. Each of 128 reward updates is followed by CE on 1,024 targets from that route's own source records, sharing Adam moments. The original predeclared task-selectivity effect is 0.67 pp, interval $[-1.22, 2.57]$ (Table 3); endpoint intervals also span zero (Table 1). These measurements separate retaining source behavior from using it for a reward task.

In an independent confirmation on eight new Memory worlds, task-relevant retrieval reaches 75.32% success versus 49.86% for alternative retrieval, from A/B-averaged entry scores of 50.00% (Figure 4b). The paired accuracy advantage is 25.46 pp (95% CI 20.10–30.81), positive in every world. The predeclared selectivity advantage is 50.92 pp (CI 40.20–61.63; exact $p = .0078125$), a separate single primary test. Both routes inherit the same correct 2^{24}-token source checkpoint, then receive matched training on public-projection and retrieval-vector targets. Reward adaptation uses calibrated sampling, vector/syntax replay without supervising source answer signs, and guarded steps. This comparison shows task-relevant retrieval supporting reward learning under this recipe; full methods, added source costs and all 32 endpoints appear in Section C.2.

Stage order and computation retention. Format learning and computation retention are distinct: an earlier study found that learning termination could coincide with a loss of source-probe accuracy (Figure 25 and Section C.14). This observation motivates format preparation before source acquisition. The resulting sequence lets probes distinguish what source training acquires, what adaptation retains, and which task the rewards select. Initialization and budget comparisons appear in Section C.13.

6.3 How does source training change the subsequent reward trajectory?

A separate Sequential world (508, Task A) compares direct RL, correct source and private-random source from one format checkpoint. Source routes use 2^{22} tokens with first-operation supervision; all routes use 8,192 reward queries and 128 nonzero RL updates. On 1,024 held-out inputs, correct-source success rises from 0% to 100% within 1,024 queries and remains there; direct RL and the private-random route finish at 51.76% and 51.27% (Figure 15). This independent world complements the eight-world A/B tests with a direct-RL comparison: source preparation enables rapid reward learning without already solving the task at entry. The original trajectories measure how preparation changes subsequent learning, while the parameter slice shows where those trajectories lie.

The common parameter plane (Figure 2) retains 91.96% of aggregate squared displacement. Its SVD basis is fitted jointly to source and reward checkpoints from all three routes. Its 81 grid models are evaluated on 256 task inputs and 128 source records. Task success uses strict output scoring; source cross-entropy is in nats per target. Every grid model uses the same evaluation records. Correct and private-random source move in different projected directions; direct RL stays nearer the origin in this projection. Colors score the slice models, while actual learning is read from the original-checkpoint curves: projected and original models can differ. Construction, residuals and reconstruction measurements are in Section C.3.

6.4 Where do public-task gains arise?

Five paired Qwen2.5-1.5B seeds compare direct RL, domain-source+RL and shuffled-source+RL on GSM8K (Cobbe et al., 2021) and HotpotQA (Yang et al., 2018). Direct RL skips added source training. Source uses 2^{22} tokens from OpenWebMath (Paster et al., 2023) or title-disjoint Wikipedia, without downstream labels; the control shuffles 64-token blocks. Reward adaptation uses 8,192 queries. On identical evaluation sets, GSM8K domain source lowers entry accuracy yet yields 5.97 ± 3.66 pp more reward-stage gain than direct RL, reaching nearly equal final accuracy. HotpotQA source raises entry accuracy, with 13.18 ± 1.50 pp less subsequent gain and lower final accuracy (Table 2). Budgets exclude inherited pretraining. Pilot selection, source-dose, answer-style and transfer comparisons appear in Section C.5 and Tables 11, 13 and 29.

Table 1: Task accuracy (%) in the original eight-world comparison (500–507), after 8,192 queries. Means ± 95% Student-t half-widths after A/B averaging, on 1,024 records; Δ is paired correct minus control (pp).

Setting	Correct	Control	Δ
Sequential			
State	27.99 ± 30.48	12.41 ± 15.57	$+15.58 \pm 19.42$
Action	82.61 ± 16.80	44.15 ± 9.71	$+38.46 \pm 18.69$
Memory			
Pure RL	48.16 ± 3.61	43.76 ± 14.75	$+4.40 \pm 15.78$
Replay	47.54 ± 3.44	42.30 ± 11.84	$+5.24 \pm 13.34$

Table 2: Public-task accuracy (%) at entry and after 8,192 queries; gain is final minus entry (pp). Means ± 95% Student-t half-widths over five paired seeds, on 1,319 GSM8K / 1,024 HotpotQA records. None denotes direct RL; its fixed-checkpoint entries have ± 0.

Source	Entry	Gain	Final
GSM8K			
None	8.57 ± 0	53.75 ± 4.14	62.32 ± 4.14
Domain	2.32 ± 0.39	59.73 ± 3.16	62.05 ± 2.93
Shuffled	8.58 ± 1.31	41.12 ± 28.34	49.70 ± 29.29
HotpotQA			
None	23.93 ± 0	23.67 ± 1.02	47.60 ± 1.02
Domain	35.10 ± 0.98	10.49 ± 1.74	45.59 ± 0.86
Shuffled	5.62 ± 0.96	38.83 ± 1.91	44.45 ± 1.13

7 Conclusion

Pretraining and midtraining can make a task learnable by revealing a mechanism that training rewards leave ambiguous and building the computation needed to use it. Our finite families connect this information gain to execution, retrieval and task learning along sampled-Adam paths from random initialization. With first-operation supervision, Sequential experiments demonstrate reward-task selection; Memory experiments connect retrieval retention to task use. Together with public-task comparisons, these results show why understanding learning from rewards requires tracking both the computation a model brings to the reward stage and how that stage uses it.

AI USE STATEMENT

AI tools assisted with literature search, model and claim formulation, drafting proof derivations, checking proof arguments in separate review passes, experiment design, code implementation and testing, and manuscript revision. These uses included substantive research assistance as well as editing. The authors are responsible for the mathematical statements, interpretations and reported results.

ETHICS STATEMENT

The theoretical tasks are synthetic. The language-model experiments use public mathematical and question-answering benchmarks and public text sources. They measure learning mechanisms rather than deploy a decision system. The main risks are overinterpreting small controlled models or benchmark-specific effects as guarantees for other settings; we state the model assumptions and report uncertainty and failure cases alongside the gains.

REPRODUCIBILITY STATEMENT

The appendix provides complete theorem statements, assumptions, resource budgets and proofs. Section C specifies models, source and reward budgets, world and seed units, controls, evaluation sets and statistical methods, with route-level results and learning curves. The accompanying anonymous code package provides experiment implementations and commands to regenerate figures from run outputs. The original eight-world comparison and independent Memory confirmation have separate methods and inference in Sections C.1 and C.2.

REFERENCES

Azwar Abdulsalam, Nishil Patel, and Andrew Saxe. RL Post-Training Builds Compositional Reasoning Strategies. arXiv:2607.07646v1, 2026. URL https://arxiv.org/abs/2607.07646v1.

Mohammed Alshiekh, Roderick Bloem, Rüdiger Ehlers, Bettina Könighofer, Scott Niekum, and Ufuk Topcu. Safe Reinforcement Learning via Shielding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. doi: 10.1609/aaai.v32i1.11797. URL https://ojs.aaai.org/index.php/AAAI/article/view/11797.

Sanae Amani, Christos Thrampoulidis, and Lin Yang. Safe Reinforcement Learning with Linear Function Approximation. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 243–253, 2021. URL https://proceedings.mlr.press/v139/amani21a.html.

Dana Angluin. Learning regular sets from queries and counterexamples. *Information and Computation*, 75(2):87–106, 1987. URL https://homepages.math.uic.edu/~lreyzin/papers/angluin87.pdf.

Yu Bai, Fan Chen, Huan Wang, Caiming Xiong, and Song Mei. Transformers as Statisticians: Provable In-Context Learning with In-Context Algorithm Selection. In *Advances in Neural Information Processing Systems*, volume 36, pp. 57125–57211, 2023. doi: 10.52202/075280-2498. URL https://papers.neurips.cc/paper_files/paper/2023/hash/b2e63e36c57e153b9015fece2352a9f9-Abstract-Conference.html.

Maria-Florina Balcan, Avrim Blum, Kiriaki Fragkia, Zhiyuan Li, and Dravyansh Sharma. Online Learnability of Chain-of-Thought Verifiers: Soundness and Completeness Trade-offs. arXiv:2603.03538v4, 2026. URL https://arxiv.org/abs/2603.03538v4.

Matej Balog, Alexander L. Gaunt, Marc Brockschmidt, Sebastian Nowozin, and Daniel Tarlow. DeepCoder: Learning to Write Programs. arXiv:1611.01989, 2016. URL https://arxiv.org/abs/1611.01989.

Ali Behrouz, Peilin Zhong, and Vahab Mirrokni. Titans: Learning to Memorize at Test Time. arXiv:2501.00663v1, 2024. URL https://arxiv.org/abs/2501.00663v1.

Enric Boix-Adsera, Omid Saremi, Emmanuel Abbe, Samy Bengio, Etai Littwin, and Joshua Susskind. When can transformers reason with abstract symbols? arXiv:2310.09753v2, 2023. URL https://arxiv.org/abs/2310.09753v2.

Steven Carr, Nils Jansen, Sebastian Junges, and Ufuk Topcu. Safe Reinforcement Learning via Shielding under Partial Observability. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 14748–14756, 2023. doi: 10.1609/aaai.v37i12.26723. URL https://ojs.aaai.org/index.php/AAAI/article/view/26723.

Siyu Chen, Heejune Sheen, Tianhao Wang, and Zhuoran Yang. Unveiling induction heads: Provable training dynamics and feature learning in transformers. arXiv:2409.10559v1, 2024. URL https://arxiv.org/abs/2409.10559v1.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training Verifiers to Solve Math Word Problems. arXiv:2110.14168, 2021. URL https://arxiv.org/abs/2110.14168.

Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified Adversarial Robustness via Randomized Smoothing. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 1310–1320, 2019. URL https://proceedings.mlr.press/v97/cohen19c.html.

Ofek I. Cohen, Lior Shani, Aviv Rosenberg, Ankur Samanta, Tal Wagner, and Yonathan Efroni. Frustratingly Simple Black-Box Adaptation of Language Models via Logit Bias. arXiv:2607.22837v1, 2026. URL https://arxiv.org/abs/2607.22837v1.

Edoardo Debenedetti, Ilia Shumailov, Tianqi Fan, Jamie Hayes, Nicholas Carlini, Daniel Fabian, Christoph Kern, Chongyang Shi, Andreas Terzis, and Florian Tramèr. Defeating Prompt Injections by Design. arXiv:2503.18813v2, 2025. URL https://arxiv.org/abs/2503.18813v2.

DeepSeek-AI et al. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. arXiv:2501.12948v1, 2025. URL https://arxiv.org/abs/2501.12948v1.

Funda Ergün, S. Ravi Kumar, and Ronitt Rubinfeld. On learning bounded-width branching programs. In *Proceedings of the Annual Conference on Computational Learning Theory*, 1995. URL https://homes.luddy.indiana.edu/fergun/PUBLICATIONS/colt95.pdf.

Jaime F. Fisac, Mo Chen, Claire J. Tomlin, and S. Shankar Sastry. Reach-Avoid Problems with Time-Varying Dynamics, Targets and Constraints. In *Proceedings of the 18th International Conference on Hybrid Systems: Computation and Control*, 2015. doi: 10.1145/2728606.2728612. URL https://people.eecs.berkeley.edu/~sastry/pubs/Pdfs%20of%202015/FisacReach2015.pdf.

Dylan Foster, Daniel Reichman, and Karthik Sridharan. Inference in sparse graphs with pairwise measurements and side information. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pp. 1810–1818, 2018. URL https://proceedings.mlr.press/v84/foster18a.html.

Hengyu Fu, Tianyu Guo, Yu Bai, and Song Mei. What can a Single Attention Layer Learn? A Study Through the Random Features Lens. arXiv:2307.11353, 2023. URL https://arxiv.org/abs/2307.11353.

Yonatan Geifman and Ran El-Yaniv. Selective Classification for Deep Neural Networks. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://papers.neurips.cc/paper_files/paper/2017/hash/4a8423d5e91fda00bb7e46540e2b0cf1-Abstract.html.

George Giapitzakis, Kimon Fountoulakis, Eshaan Nichani, and Jason D. Lee. On the Statistical Query Complexity of Learning Semiautomata: a Random Walk Approach. arXiv:2510.04115v1, 2025. URL https://arxiv.org/abs/2510.04115v1.

Vitoria Guardieiro, Avishree Khare, Adam Stein, and Eric Wong. Instruction Following by Principled Boosting Attention of Large Language Models. arXiv:2506.13734, 2025. URL https://arxiv.org/abs/2506.13734.

Chuan Guo, Juan Felipe Ceron Uribe, Sicheng Zhu, Christopher A. Choquette-Choo, Steph Lin, Nikhil Kandpal, Milad Nasr, Rai, Sam Toyer, Miles Wang, Yaodong Yu, Alex Beutel, and Kai Xiao. IH-Challenge: A Training Dataset to Improve Instruction Hierarchy on Frontier LLMs. arXiv:2603.10521v1, 2026. URL https://arxiv.org/abs/2603.10521v1.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8342–8360. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.740. URL https://aclanthology.org/2020.acl-main.740/.

Yaomengxi Han and Debarghya Ghoshdastidar. Transformers provably learn sparse XOR with polylogarithmic parameters. arXiv:2502.07553v2, 2026. URL https://arxiv.org/abs/2502.07553v2.

Josiah P. Hanna and Nicholas E. Corrado. When Can Model-Free Reinforcement Learning be Enough for Thinking? arXiv:2506.17124v1, 2025. URL https://arxiv.org/abs/2506.17124v1.

Cheng Huan and Hongwei Yuan. A mean-field analysis of multi-head self-attention under cross-entropy training. arXiv:2606.10469v1, 2026. URL https://arxiv.org/abs/2606.10469v1.

Ruiquan Huang, Yingbin Liang, and Jing Yang. How transformers learn regular language recognition: A theoretical study on training dynamics and implicit bias. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 25516–25549, 2025a. URL https://proceedings.mlr.press/v267/huang25s.html.

Yixiao Huang, Hanlin Zhu, Zixuan Wang, Jiantao Jiao, Stuart Russell, Somayeh Sojoudi, and Song Mei. Transformers Provably Learn to Internalize Chain-of-Thought. arXiv:2605.28600v1, 2026a. URL https://arxiv.org/abs/2605.28600v1.

Yu Huang, Zixin Wen, Aarti Singh, Yuejie Chi, and Yuxin Chen. Transformers Provably Learn Chain-of-Thought Reasoning with Length Generalization. In *Advances in Neural Information Processing Systems*, 2025b. URL https://yuxinchen2020.github.io/publications/transformers_CoT.pdf.

Yu Huang, Zixin Wen, Yuejie Chi, Yuting Wei, Aarti Singh, Yingbin Liang, and Yuxin Chen. On the Emergence of Implicit Curriculum in RLVR Learning Dynamics. arXiv:2602.14872v3, 2026b. URL https://arxiv.org/abs/2602.14872v3.

Sebastian Junges, Nils Jansen, and Sanjit A. Seshia. Enforcing Almost-Sure Reachability in POMDPs. In *Computer Aided Verification*, volume 12760 of *Lecture Notes in Computer Science*, pp. 602–625. Springer, 2021. doi: 10.1007/978-3-030-81688-9_28. URL https://link.springer.com/chapter/10.1007/978-3-030-81688-9_28.

Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In *International Conference on Learning Representations*, 2015. URL https://arxiv.org/abs/1412.6980.

Lingjing Kong, Xin Liu, Guangyi Chen, Martin Q. Ma, Xiangchen Song, Yuekai Sun, Mikhail Yurochkin, Taylor W. Killian, Ruslan Salakhutdinov, Kun Zhang, Eric P. Xing, and Zhengzhong Liu. From Reasoning Traces to Reusable Modules: Understanding Compositional Generalization in Language Model Reasoning. arXiv:2606.18089v2, 2026. URL https://arxiv.org/abs/2606.18089v2.

Ananya Kumar, Aditi Raghunathan, Robbie Jones, Tengyu Ma, and Percy Liang. Fine-Tuning can Distort Pretrained Features and Underperform Out-of-Distribution. arXiv:2202.10054, 2022. URL https://arxiv.org/abs/2202.10054.

Aviral Kumar, Vincent Zhuang, Rishabh Agarwal, Yi Su, John D Co-Reyes, Avi Singh, Kate Baumli, Shariq Iqbal, Colton Bishop, Rebecca Roelofs, Lei M Zhang, Kay McKinney, Disha Shrivastava, Cosmin Paduraru, George Tucker, Doina Precup, Feryal Behbahani, and Aleksandra Faust. Training Language Models to Self-Correct via Reinforcement Learning. arXiv:2409.12917, 2024. URL https://arxiv.org/abs/2409.12917.

Ali Larian, Qian Lin, Chang Zong Wu, and Daniel S. Brown. Multi-Modal, Multi-Environment Machine Teaching for Robust Reward Learning. arXiv:2607.08647v1, 2026. URL https://arxiv.org/abs/2607.08647v1.

Itay Lavie, Kirsten Fischer, Andrey Lekov, Frederic Van Maele, Zohar Ringel, and Moritz Helias. Phase transitions in attention: A bayesian theory of copy head emergence. arXiv:2606.12058v1, 2026. URL https://arxiv.org/abs/2606.12058v1.

Bingbin Liu, Jordan T. Ash, Surbhi Goel, Akshay Krishnamurthy, and Cyril Zhang. Transformers Learn Shortcuts to Automata. arXiv:2210.10749, 2022. URL https://arxiv.org/abs/2210.10749.

Hailiang Liu and Michael Herty. A mean-field theory of transformers: Well-posedness of the coupled data–parameter dynamics and global convergence of training. arXiv:2608.25055v1, 2026. URL https://arxiv.org/abs/2608.25055v1.

Bochen Lyu, Yiyang Jia, Xiaohao Cai, and Zhanxing Zhu. Transformers with RL or SFT provably learn sparse boolean functions, but differently. arXiv:2511.17852v3, 2026. URL https://arxiv.org/abs/2511.17852v3. ICML 2026 final version.

Alireza Modirshanechi, Benjamin Eysenbach, Peter Dayan, and Eric Schulz. Unifying Goal-Conditioned RL and Unsupervised Skill Learning via Control-Maximization. arXiv:2605.06145v1, 2026. URL https://arxiv.org/abs/2605.06145v1.

Eshaan Nichani, Alex Damian, and Jason D. Lee. How Transformers Learn Causal Structure with Gradient Descent. arXiv:2402.14735v2, 2024a. URL https://arxiv.org/abs/2402.14735v2.

Eshaan Nichani, Jason D. Lee, and Alberto Bietti. Understanding Factual Recall in Transformers via Associative Memories. arXiv:2412.06538v1, 2024b. URL https://arxiv.org/abs/2412.06538v1.

Keiran Paster, Marco Dos Santos, Zhangir Azerbayev, and Jimmy Ba. OpenWebMath: An Open Dataset of High-Quality Mathematical Web Text. arXiv:2310.06786, 2023. URL https://arxiv.org/abs/2310.06786.

Jason Phang, Thibault Févry, and Samuel R. Bowman. Sentence Encoders on STILTs: Supplementary Training on Intermediate Labeled-data Tasks. arXiv:1811.01088, 2018. URL https://arxiv.org/abs/1811.01088.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. URL https://arxiv.org/abs/2412.15115.

Nived Rajaraman, Audrey Huang, Miroslav Dudik, Robert Schapire, Dylan Foster, and Akshay Krishnamurthy. Learning to Reason with Curriculum II: Compositional Generalization. arXiv:2606.27721v1, 2026. URL https://arxiv.org/abs/2606.27721v1.

Yuval Ran-Milo, Yotam Alexander, Shahar Mendel, and Nadav Cohen. Outcome-Based RL Provably Leads Transformers to Reason, but Only With the Right Data. arXiv:2601.15158v4, 2026. URL https://arxiv.org/abs/2601.15158v4.

Yunwei Ren, Zixuan Wang, and Jason D. Lee. Learning and transferring sparse contextual bigrams with linear transformers. arXiv:2410.23438, 2024. URL https://arxiv.org/abs/2410.23438. NeurIPS 2024.

Clayton Sanford, Bahare Fatemi, Ethan Hall, Anton Tsitsulin, Mehran Kazemi, Jonathan Halcrow, Bryan Perozzi, and Vahab Mirrokni. Understanding Transformer Reasoning Capabilities via Graph Algorithms. arXiv:2405.18512v1, 2024a. URL https://arxiv.org/abs/2405.18512v1.

Clayton Sanford, Daniel Hsu, and Matus Telgarsky. Transformers, parallel computation, and logarithmic depth. arXiv:2402.09268v1, 2024b. URL https://arxiv.org/abs/2402.09268v1.

Sneheel Sarangi and Hanan Salam. Small LLMs Do Not Learn a Generalizable Theory of Mind via Reinforcement Learning. arXiv:2507.15788v1, 2025. URL https://arxiv.org/abs/2507.15788v1.

Yash Sarrof, Yupei Du, Katharina Stein, Alexander Koller, Sylvie Thiébaux, and Michael Hahn. On the Ability of Transformers to Verify Plans. arXiv:2603.19954v2, 2026. URL https://arxiv.org/abs/2603.19954v2.

Nikunj Saunshi, Sadhika Malladi, and Sanjeev Arora. A Mathematical Exploration of Why Language Models Help Solve Downstream Tasks. arXiv:2010.03648, 2020. URL https://arxiv.org/abs/2010.03648.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language Models Can Teach Themselves to Use Tools. arXiv:2302.04761, 2023. URL https://arxiv.org/abs/2302.04761.

Xiaolong Sun, Qichao Wang, Hangyu Li, and Liang Chen. Verifiable Memory: Learning Unified Memory Management with Local and Global Verifiers for Large Language Model Agents. arXiv:2608.03137v1, 2026. URL https://arxiv.org/abs/2608.03137v1.

Davoud Ataee Tarzanagh, Yingcong Li, Christos Thrampoulidis, and Samet Oymak. Transformers as Support Vector Machines. arXiv:2308.16898, 2023. URL https://arxiv.org/abs/2308.16898.

Nikolaos Tsilivis, Eran Malach, Karen Ullrich, and Julia Kempe. How Reinforcement Learning After Next-Token Prediction Facilitates Learning. arXiv:2510.11495v2, 2025. URL https://arxiv.org/abs/2510.11495v2.

Matteo Turchetta, Felix Berkenkamp, and Andreas Krause. Safe Exploration in Finite Markov Decision Processes with Gaussian Processes. In *Advances in Neural Information Processing Systems*, volume 29, 2016. URL https://papers.nips.cc/paper_files/paper/2016/hash/9a49a25d845a483fae4be7e341368e36-Abstract.html.

Bhavya Vasudeva, Puneesh Deora, Alberto Bietti, Vatsal Sharan, and Christos Thrampoulidis. Understanding Contextual Recall in Transformers: How Finetuning Enables In-Context Reasoning over Pretraining Knowledge. arXiv:2603.20969v1, 2026. URL https://arxiv.org/abs/2603.20969v1.

Dennis Volpano, Cynthia Irvine, and Geoffrey Smith. A Sound Type System for Secure Flow Analysis. *Journal of Computer Security*, 4, 1996. doi: 10.3233/JCS-1996-42-304. URL https://journals.sagepub.com/doi/10.3233/JCS-1996-42-304.

Eric Wallace, Kai Xiao, Reimar Leike, Lilian Weng, Johannes Heidecke, and Alex Beutel. The Instruction Hierarchy: Training LLMs to Prioritize Privileged Instructions. arXiv:2404.13208, 2024. URL https://arxiv.org/abs/2404.13208.

Po-Wei Wang, Priya Donti, Bryan Wilder, and Zico Kolter. SATNet: Bridging deep learning and logical reasoning using a differentiable satisfiability solver. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 6545–6554, 2019. URL https://proceedings.mlr.press/v97/wang19e.html.

Honghao Wei, Xin Liu, and Lei Ying. Safe Reinforcement Learning with Instantaneous Constraints: The Role of Aggressive Exploration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 21708–21716, 2024. doi: 10.1609/aaai.v38i19.30170. URL https://ojs.aaai.org/index.php/AAAI/article/view/30170.

Stanley Wei and Juno Kim. Provable Benefits of RLVR over SFT for Reasoning Models: Learning to Backtrack Efficiently. arXiv:2606.22938v1, 2026. URL https://arxiv.org/abs/2606.22938v1.

Gail Weiss, Yoav Goldberg, and Eran Yahav. Learning deterministic weighted automata with queries and counterexamples. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://papers.neurips.cc/paper/9062-learning-deterministic-weighted-automata-with-queries-and-counterexamples.pdf.

Ruibin Xiong, Yunchang Yang, Di He, Kai Zheng, Shuxin Zheng, Chen Xing, Huishuai Zhang, Yanyan Lan, Liwei Wang, and Tie-Yan Liu. On Layer Normalization in the Transformer Architecture. arXiv:2002.04745, 2020. URL https://arxiv.org/abs/2002.04745.

Anton Xue, Avishree Khare, Rajeev Alur, Surbhi Goel, and Eric Wong. Logicbreaks: A Framework for Understanding Subversion of Rule-based Inference. arXiv:2407.00075, 2024. URL https://arxiv.org/abs/2407.00075.

Beyazit Yalcinkaya, Niklas Lauffer, Marcell Vazquez-Chanlatte, and Sanjit A. Seshia. Provably Correct Automata Embeddings for Optimal Automata-Conditioned Reinforcement Learning. arXiv:2503.05042v2, 2025. URL https://arxiv.org/abs/2503.05042v2.

Kaiyu Yang, Aidan M. Swope, Alex Gu, Rahul Chalamala, Peiyang Song, Shixing Yu, Saad Godil, Ryan Prenger, and Anima Anandkumar. LeanDojo: Theorem Proving with Retrieval-Augmented Language Models. arXiv:2306.15626, 2023. URL https://arxiv.org/abs/2306.15626.

Tong Yang, Yu Huang, Yingbin Liang, and Yuejie Chi. Agentic Transformers Provably Learn to Search via Reinforcement Learning. arXiv:2606.00183v1, 2026. URL https://arxiv.org/abs/2606.00183v1.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. arXiv:1809.09600, 2018. URL https://arxiv.org/abs/1809.09600.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing Reasoning and Acting in Language Models. arXiv:2210.03629, 2022. URL https://arxiv.org/abs/2210.03629.

Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Yang Yue, Shiji Song, and Gao Huang. Does Reinforcement Learning Really Incentivize Reasoning Capacity in LLMs Beyond the Base Model? arXiv:2504.13837, 2025. URL https://arxiv.org/abs/2504.13837.

Ji Zeng, Dayuan Fu, Tiantian Mi, Yumin Zhuang, Yaxing Huang, Xuefeng Li, Lyumanshan Ye, Muhang Xie, Qishuo Hua, Zhen Huang, Mohan Jiang, Hanning Wang, Jifan Lin, Yang Xiao, Jie Sun, Yunze Wu, and Pengfei Liu. daVinci-Dev: Agent-native midtraining for software engineering. *arXiv preprint arXiv:2601.18418*, 2026. URL https://arxiv.org/abs/2601.18418.

Jin Peng Zhou, Kaiwen Wang, Jonathan Chang, Zhaolin Gao, Nathan Kallus, Kilian Q. Weinberger, Kianté Brantley, and Wen Sun. $Q\sharp$: Provably Optimal Distributional RL for LLM Post-Training. arXiv:2502.20548v2, 2025. URL https://arxiv.org/abs/2502.20548v2.

Appendix

A More Related Work

We organize prior work around three questions: what information each training stage supplies, how optimization turns that information into computation, and how rewards use the acquired computation.

A.1 Information supplied by observations and feedback

Identification with different observation channels. Learning an unknown mechanism depends on which queries and observations are permitted. Membership and equivalence queries support regular-language identification (Angluin, 1987); related methods learn weighted automata from queries and counterexamples (Weiss et al., 2019). Bounded-width branching programs and statistical-query learning of semiautomata provide further settings in which the observation model determines learnability (Ergün et al., 1995; Giapitzakis et al., 2025). Pairwise measurements with side information illustrate a complementary distinction: edge observations constrain relative labels, while additional vertex information helps recover the labels themselves (Foster et al., 2018). Our information results identify exactly which transition and task transformations leave all training rewards unchanged, and which source observations distinguish their held-out predictions. The resulting obstruction concerns indistinguishable observations even with unlimited reward queries; it is different from an optimizer's difficulty in finding an informative trajectory.

What a verifier supplies. The feedback interface matters beyond a binary reward's numerical range. Learning a chain-of-thought verifier online uses an expert label stream and has distinct soundness and completeness objectives (Balcan et al., 2026). Machine teaching can also choose among demonstrations and other feedback modalities (Larian et al., 2026). These settings supply different information from the terminal correctness feedback in our sequential task. In the contextual binary-action task, correctness together with the chosen action does reveal the binary label, so supervised adaptation is a valid alternative there. We account for this difference when comparing learning routes.

Specifications, safety, and external structure. Reach–avoid control specifies dynamics, target sets, and constraints (Fisac et al., 2015). Shielding uses a supplied specification and model information to restrict actions, including under partial observability (Alshiekh et al., 2018; Carr et al., 2023); almost-sure reachability in partially observable Markov decision processes (POMDPs) studies the corresponding structural conditions (Junges et al., 2021). Automata-conditioned reinforcement learning likewise uses an explicit automaton interface (Yalcinkaya et al., 2025). These comparisons identify what is supplied to a learner or controller; our sequential result concerns task utility and public-coordinate violations within a specified finite grammar.

Other reliability objectives require different observations and success events. Noninterference compares executions with different secret inputs (Volpano et al., 1996); selective classification couples conditional error with coverage (Geifman & El-Yaniv, 2017); randomized smoothing certifies predictions within a specified perturbation set (Cohen et al., 2019). Safe exploration also distinguishes zero-violation guarantees under suitable structural or startup information from bounds on accumulated constraint violation (Turchetta et al., 2016; Amani et al., 2021; Wei et al., 2024). These objectives each specify their own success event and verification requirement, distinct from terminal task success.

A.2 Learning across source and adaptation stages

Intermediate training and task information. Supervised training on intermediate labeled-data tasks (STILTs) places a supervised intermediate task between language-model pretraining and downstream adaptation (Phang et al., 2018). Domain- and task-adaptive pretraining continue prediction on unlabeled text from a relevant distribution (Gururangan et al., 2020). Agent-native midtraining instead uses trajectories preserving agent workflows, including observations from executed tools and tests (Zeng et al., 2026). Their common use of an intermediate stage does not make their information interfaces identical: supervised examples may reveal the downstream choice, whereas our source laws leave the task binding or contextual decision rule unspecified. The public-task experiments

use domain adaptation as an empirical comparison; the controlled tasks separate source mechanism information from downstream task information.

From source computation to reward learning. Tsilivis et al. (2025, Section 4, Theorems 1–2 and Appendices D.3–D.4) analyze next-token prediction followed by reinforcement learning on mixtures of short and long demonstrations of a single parity task. Their autoregressive linear model is trained with finite source stochastic gradient descent and self-taught reasoner (STaR) rounds. Their finite learning theorem uses a reward that checks the complete chain of thought, and explains how post-training increases the probability of long, correct computations. In our sequential construction, source observations describe local transitions without revealing the downstream task binding, and feedback checks the terminal answer. This changes both the missing information and what the feedback stage must learn.

Wei & Kim (2026, Sections 3.2–3.3 and 4.3–4.4) study a related world-model-to-reward sequence for graph backtracking. They establish a pretraining convergence result for a linear-softmax bigram model, then analyze post-training from a model assumed arbitrarily close to the world law. Their edge-state bigram analysis uses an outcome reward with a length penalty. Its gradient-flow and sign-policy-gradient-flow analysis compares outcome-reward learning with shortest-path supervised fine-tuning (SFT) through inference-time efficiency. Our sequential theorem follows finite sampled updates across both stages, from random initialization, and evaluates held-out operation words.

Reusable modules and compositional curricula provide further accounts of what stages contribute (Kong et al., 2026; Rajaraman et al., 2026). Empirically, Abdulsalam et al. (2026) study primitive rewrite pretraining followed by final-answer reinforcement learning (RL) and trace how valid composed procedures develop. Their distinction between exposure to primitives and usable procedures is particularly relevant to our computation interventions. Staged training can also change where a computation occurs: Huang et al. (2026a) analyze a structured curriculum for internalizing explicit chain-of-thought steps. Our question is the acquisition and downstream use of reusable computation across the two stages.

A.3 Acquisition and Reuse of Representations

Transferable prediction structure. Language-modeling performance can imply useful downstream features when the source and target tasks satisfy an appropriate relationship (Saunshi et al., 2020). Using linear attention and preconditioned projected proximal updates for sparse contextual bigrams, Ren et al. (2024) show that a correlated pretrained transition matrix supplies a signal that allows their learning algorithm to bypass its initial sample-intensive stage. Our contextual result instead derives the source-trained attention and head-factor distribution and follows it into learning a new reward-dependent rule with sampled Adam, including its two moment states. Representation reuse also depends on later updates: full fine-tuning can distort pretrained features and worsen out-of-distribution performance (Kumar et al., 2022). We therefore track both source acquisition and retention of the information needed during feedback.

Representing and learning sequential computation. Graph algorithms, parallel computation, and automata shortcuts characterize what transformer architectures can express and how depth or intermediate computation changes their resources (Sanford et al., 2024a;b; Liu et al., 2022). Huang et al. (2025b) analyze gradient-based learning and length generalization in structured state-tracking tasks. These results motivate explicit operation and history interfaces. Our sequential source theorem additionally establishes execution on generated prefixes outside the source trajectories, which is what makes the learned local computation available during reward rollouts.

Learning attention and memory. Margin-based token selection, induction-head dynamics, and latent causal structure provide different accounts of attention acquisition (Tarzanagh et al., 2023; Chen et al., 2024; Nichani et al., 2024a). Random-feature, sparse-parity, regular-language, mean-field, and Bayesian copy-head analyses study complementary representations and training regimes (Fu et al., 2023; Han & Ghoshdastidar, 2026; Huang et al., 2025a; Huan & Yuan, 2026; Liu & Herty, 2026; Lavie et al., 2026). Normalization can change gradient scales at initialization (Xiong et al., 2020). In our contextual model, the relevant interaction occurs across stages: attention pooling

changes correlation and signal scale, and the normalization family allows useful reward learning with different directions of attention-contrast change.

Memory is not a single observation interface. Associative-memory analyses study parametric factual storage (Nichani et al., 2024b); contextual recall combines pretraining knowledge with implicit attribute inference after fine-tuning (Vasudeva et al., 2026). Statistical in-context learning and abstract-symbol reasoning address other forms of adaptation and generalization (Bai et al., 2023; Boix-Adsera et al., 2023). Titans changes writable neural memory, while Verifiable Memory trains an agent's operations over long-term memory, active context, and episodic history using multiple verifier signals (Behrouz et al., 2024; Sun et al., 2026). Our contextual theorem concerns retrieval and a context–cue rule in a supplied numerical interface, and follows their acquisition through both training stages.

A.4 Reward learning, empirical comparisons, and capability interfaces

What rewards change. Outcome-reward learning can acquire a graph traversal strategy when the training distribution has sufficient mass on simple instances (Ran-Milo et al., 2026). That analysis follows expected-gradient flow of attention matrices with fixed values and prescribed initial structure, and also gives a discrete extension. Implicit curricula and learned search have also been analyzed through RL dynamics (Huang et al., 2026b; Yang et al., 2026), while Lyu et al. (2026) compare RL and supervised Boolean-policy learning with a supplied dependency mask and process rewards. Zhou et al. (2025) give a reference-policy-dependent resource analysis for Kullback–Leibler (KL)-regularized distributional value learning. These works clarify different entry conditions and learning objectives. Our paired constructions derive source-acquired computation and then study how feedback uses it, alongside the information that feedback alone cannot identify.

Empirical comparisons require the same distinctions. Yue et al. (2025) report improved small-budget success after RL without an expansion of large-budget coverage in their tested settings. Our equivalence result is of a different kind: it identifies held-out targets that rewards on the training distribution cannot reliably distinguish across reward-equivalent worlds. Self-correction training supplies a sequence of attempts and rewards (Kumar et al., 2024); thought-action models study when initial policies make internal computation useful (Hanna & Corrado, 2025); logit-bias adaptation studies a different route for using a fixed model (Cohen et al., 2026). Goal-conditioned control also depends on whether success means arrival at a deadline, first arrival, or occupancy (Modirshanechi et al., 2026). Our experiments distinguish access to successful trajectories, acquisition of the required computation, and reward-dependent adaptation, rather than treating any one of these measurements as all three.

Task-specific computation and verification. Several capability families require additional resources that are explicit in their original settings. SATNet incorporates a differentiable satisfiability solver (Wang et al., 2019); DeepCoder uses learned guidance for program search (Balog et al., 2016); LeanDojo provides an executable theorem-proving and retrieval environment (Yang et al., 2023). Toolformer and ReAct study tool calls and environment interaction (Schick et al., 2023; Yao et al., 2022). These are distinct sources of computation and feedback from our finite transition and contextual tasks. Plan-verification theory, rule-based inference attacks, and attention interventions similarly specify particular symbolic or architectural interfaces (Sarrof et al., 2026; Xue et al., 2024; Guardieiro et al., 2025).

Instruction-hierarchy training and external enforcement address different parts of agent reliability (Wallace et al., 2024; Guo et al., 2026; Debenedetti et al., 2025). Transfer across theory-of-mind tasks must also be measured rather than inferred from in-distribution success (Sarangi & Salam, 2025).

B Discussion

Rewards select among computations that exist. The information results and training constructions explain complementary parts of the same mechanism. Source observations distinguish mechanisms that training rewards leave indistinguishable. The Adam constructions turn this information into execution or retrieval, then use sampled rewards to learn its task-specific use: two updates in the sequential construction and an explicit finite budget in the contextual one. The acquired computation

applies to held-out inputs, allowing the task rule to transfer. This gives a concrete question to ask of a reward-stage gain: which acquired computation supports the new task rule? The measurements in Section 6 address each part of this question.

Three interventions, three questions. The no-source bound in contextual memory concerns world-independent learners with the specified observations. The frozen-attention run holds the training schedule fixed and freezes the initial random query/key factors in both stages while training the head. Direct language-model RL omits additional source midtraining and starts from inherited pretrained weights. Each intervention isolates one question: information access, representation learning, or continued source training. Section F and Table 32 compare the neural path, which reaches the binding through the model's own sampled trajectories at $2B_R$ verifier calls, with an explicit controller that needs at most $q - 1$ verifications, and with supervised and test-time alternatives.

Initial information and the stage of source prediction. The information results specify what the learner observes. An inherited checkpoint that already reveals the hidden mechanism can distinguish worlds even when their training rewards coincide; the no-source bounds assume a world-independent learner and initialization. Likewise, the source-only bound uses source laws that leave the downstream rule unspecified. The finite Adam constructions start from the stated random distributions, whose initial geometry lets the proofs establish head alignment and control the training trajectory. The Qwen experiments study continued source prediction from inherited weights, making source midtraining the intervention. The constructions and experiments thus study acquisition from two distinct starting points: specified random factors and inherited language-model weights. On public tasks, domain-related source text without downstream labels is a different information interface from the rule-independent source laws of the controlled families.

From acquired computation to task use. Source-checkpoint probes, retention measurements and paired reward branches separate acquired computation from its task-specific use. The historical output decompositions illustrate this distinction: a valid public path or output format can coexist with chance-level private-task behavior (Sections C.8 and C.12). The original eight-world Sequential comparison shows reward-task selection after execution is acquired; its Memory comparison shows that replay preserves intervention-sensitive retrieval (Section C.1). The independent Memory confirmation connects retrieval preparation to task learning: under a matched adaptation recipe, task-relevant retrieval reaches 75.32% success versus 49.86% for alternative retrieval (Section C.2). Both confirmation routes inherit correct source information and receive the same added training budget. The original comparison uses a two-test Holm family; the confirmation has a separate single primary test and additional source preparation. Together, these experiments show how stage-specific measurements connect the computation acquired from source data to its retention and reward-dependent use.

Entry accuracy, reward gain and final accuracy. Final accuracy combines two effects of source training: accuracy at reward entry and gain during reward adaptation. Our public-task comparisons show that these effects can move in different directions. Reporting all three scores locates the improvement in the training sequence; the controlled-task probes identify the computation available at each stage. This connects public-task evaluation to the paper's central question: which computations does source prediction make available for rewards to use?

C Experimental details and supplementary results

World/training seeds are the independent units throughout; shared source trajectories, evaluation prompts and checkpoints are repeated measurements within those units. Item-level results, configurations and analysis code accompany the paper.

C.1 Output supervision and source replay

We use eight fresh worlds (seeds 500–507) for each module, Qwen2.5-0.5B, and a shared task-independent format stage followed by source training and reward adaptation. The format stage has 128 updates for Sequential and 32 for Memory, at learning rate 10^{-5}. Source training also uses 10^{-5}: Sequential receives 2^{22} observed tokens in 512 updates, and discrete Memory receives 2^{24} in 2,048 updates. There is no format stage after source training.

Sequential crosses correct versus public-preserving/private-randomized source with state/END/EOS targets versus first-operation/state/END/EOS targets. The new target is an already observed random operation, independent of the private task binding; initial-state and suffix inputs are not prediction targets. Memory crosses correct versus wrong-retrieval source with pure reward adaptation versus source replay. Both retrieval routes use the all-token source objective. Replay samples only complete records previously observed by that route, with an independent random stream, and supplies exactly 1,024 source prediction targets after each reward update. It shares the branch's Adam state and learning rate, 3×10^{-6}, with reward training. Each replay branch thus uses an additional 131,072 source targets and 128 cross-entropy updates.

Each source checkpoint initializes two reward-trained branches, one for task A and one for task B. Relative to A, B flips the Sequential binding's private bit, or negates all Memory context signs; mechanism and prompt distribution remain fixed. Every branch receives 8,192 verifier queries in 128 batches of 64, at fixed learning rate 3×10^{-6}. We score unmodified generated outputs with the strict verifier. At 0, 1,024, 2,048, 4,096 and 8,192 queries, all routes and both task-trained models use the same fixed 1,024 held-out evaluation records within a world. Source probes are repeated at these points. There are 16 format stages, 48 source stages and 128 reward branches; shared ancestors do not create additional independent samples. Across both modules the reward-training budget is 1,048,576 queries.

The two primary paired effects compare correct and matched-control task selectivity under first-operation supervision for Sequential and under source replay for Memory. We first average the task-A and task-B contributions within each world, then report the eight world-level paired differences, a 95% Student-t interval, and a two-sided exact sign-flip test. Exact sign-flip inference assumes sign symmetry of paired world effects under the null. Holm correction applies jointly to these two primary tests. The target-objective and replay interactions are secondary comparisons. These samples and this test family are separate from the earlier eight-world acquisition study.

Resources, implementation and evaluation units. Sequential state-only and action-plus-state objectives share observed source records, but adding operation targets increases optimized target positions. Observed tokens, optimized prediction targets and optimizer calls are therefore distinct resources. Memory pure/replay A/B branches reuse the same route-specific source ancestor. Replay uses that ancestor's recorded labels, excludes partial boundary and held-out records, and preserves the reward-prompt random stream. A replay branch has 128 RL calls and 128 source CE calls, versus 128 RL calls for pure training. Both modules use BF16 model evaluation with FP32 master parameters and Adam moments in this eight-world comparison. Sampled completions use the generation backend; response log probabilities for updates are recomputed by the trainer. The constructions establish finite learning paths in the specified architectures and schedules; these Qwen experiments measure acquisition and reward-dependent use in pretrained language models.

The 1,024 evaluation items in a world are fixed across routes, tasks and checkpoints. Sampling can repeat a prompt: Sequential worlds 501 and 502 each contain 1,023 distinct prompt texts among 1,024 item IDs; the remaining worlds contain 1,024. These repeats are identical across paired routes and are retained with their original sample multiplicities. The independent statistical unit is the world, not an item, A/B branch or checkpoint. Source probes use 512 Sequential records with first operation supplied, and 256 Memory records with matched relevant-value interventions. These probes measure execution of the source computation on held-out records.

Primary and secondary effects. Table 3 reports the complete predeclared two-test family; Table 4 reports every predeclared secondary contrast. The Sequential primary effect compares correct and private-random source within the action objective. The source-by-objective interaction measures how adding action supervision changes the source-control selectivity advantage. Its Student-t interval is positive, while the unadjusted exact sign-flip test gives $p = .08594$ over the eight world effects.

Absolute success and task selection. Correct-source Sequential action models finish at 78.09% and 87.13% on tasks A/B, versus 38.39% and 49.90% under private-random source. Averaging A/B within world gives 82.61% versus 44.15%, a paired difference of 38.46 pp (95% interval $[19.78, 57.15]$). These absolute-success summaries complement the task-selectivity primary test. Here the source advantage is measured against the private-random source control.

Table 3: Predeclared source-control effects on task selectivity (percentage points), eight paired worlds per module. Two-sided exact sign-flip tests; Holm adjustment over both primary tests.

Module	Effect	95% CI	Exact p	Holm p
Sequential	67.57	[37.04, 98.10]	0.007812	0.015625
Memory	0.67	[-1.22, 2.57]	0.640625	0.640625

Table 4: All predeclared secondary contrasts at 8,192 queries (percentage points). Eight paired world differences; 95% Student-t intervals and unadjusted exact sign-flip tests.

Module	Comparison	Metric	Effect	95% CI	Exact p
Seq.	Correct: action–state	Select.	43.55	[1.47, 85.64]	0.08594
Seq.	Correct: action–state	Success	54.62	[23.75, 85.50]	0.01562
Seq.	Control: action–state	Select.	0.96	[-0.09, 2.02]	0.04688
Seq.	Control: action–state	Success	31.74	[13.15, 50.32]	0.01562
Seq.	Source interaction	Select.	42.59	[0.82, 84.36]	0.08594
Seq.	Source interaction	Success	22.88	[-1.44, 47.20]	0.08594
Mem.	Correct: replay–pure	Select.	-0.06	[-0.52, 0.40]	1.00000
Mem.	Correct: replay–pure	Success	-0.62	[-5.54, 4.31]	0.62500
Mem.	Control: replay–pure	Select.	-0.59	[-2.45, 1.27]	1.00000
Mem.	Control: replay–pure	Success	-1.45	[-6.13, 3.22]	0.56250
Mem.	Source interaction	Select.	0.52	[-1.40, 2.45]	0.87500
Mem.	Source interaction	Success	0.84	[-6.20, 7.87]	0.68750

Table 5: Original eight-world task success: reward entry, within-world reward gain, and final score at 8,192 queries. A/B branches are averaged within each world; means and 95% Student-t half-widths over eight worlds. Entry and final are percentages; gain is percentage points.

Module	Route	Entry	Gain	Final
Sequential	Correct/state	0.20 ± 0.46	27.80 ± 30.37	27.99 ± 30.48
Sequential	Control/state	0.00 ± 0.00	12.41 ± 15.57	12.41 ± 15.57
Sequential	Correct/action	4.83 ± 6.72	77.78 ± 17.72	82.61 ± 16.80
Sequential	Control/action	2.32 ± 3.39	41.83 ± 11.12	44.15 ± 9.71
Memory	Correct/pure	49.70 ± 0.71	-1.54 ± 3.80	48.16 ± 3.61
Memory	Control/pure	41.41 ± 11.93	2.35 ± 9.53	43.76 ± 14.75
Memory	Correct/replay	49.70 ± 0.71	-2.16 ± 2.82	47.54 ± 3.44
Memory	Control/replay	41.41 ± 11.93	0.90 ± 5.18	42.30 ± 11.84

Correct-source Memory replay retains the acquired source response: original/flip-pair accuracy is 92.21%, compared with 0% under pure RL, with a paired retention-difference interval of $[87.93, 96.50]$ pp. First-field accuracy is 96.24% versus 49.66%. Correct source exceeds wrong retrieval in final task accuracy by 4.40 pp under pure RL (95% interval $[-11.38, 20.18]$) and 5.24 pp with replay ($[-8.10, 18.58]$). These endpoint intervals and the primary task-selectivity result leave the task advantage unresolved at this budget. Replay preserves the measured source computation, while its task-specific use remains a separate outcome.

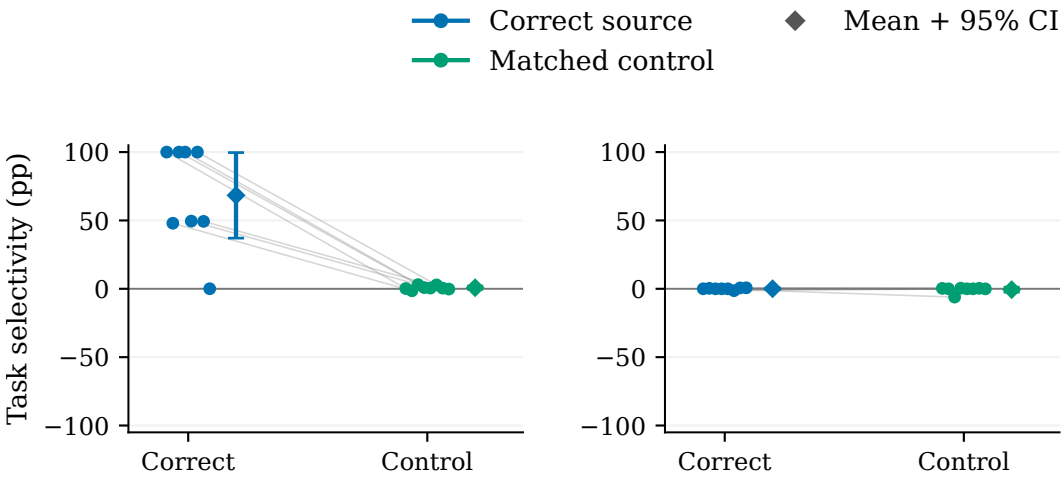


Figure 5: Correct and matched-control task selectivity at 8,192 reward queries: (a) Sequential action objective; (b) Memory source replay. Small dots are eight worlds, gray lines pair worlds, and diamonds give route means with untruncated 95% intervals. Primary inference uses paired differences, not overlap of route intervals.

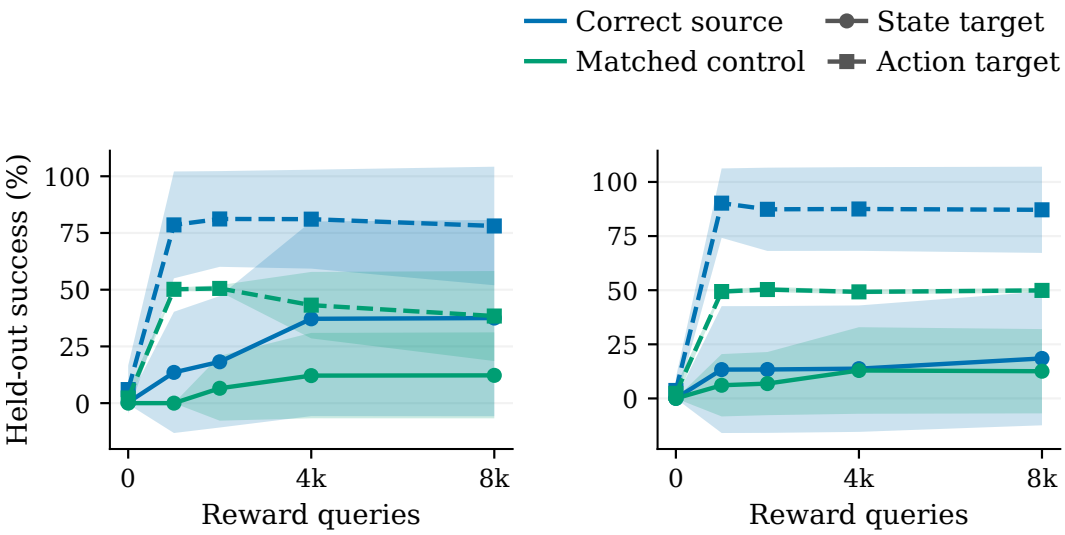


Figure 6: Sequential learning curves for (a) task A and (b) task B, with correct/private-random source crossed with state/action-plus-state targets. Each point averages eight worlds on 1,024 fixed evaluation items per world; bands are untruncated 95% Student-t intervals.

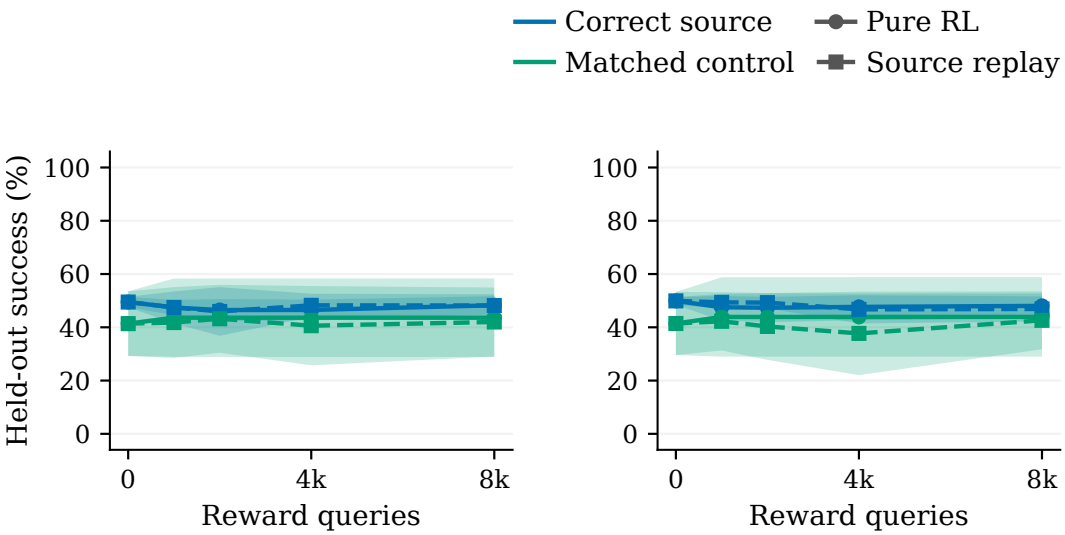


Figure 7: Memory learning curves for (a) task A and (b) task B. Correct/wrong retrieval is crossed with pure reward training/source replay; the same ancestors and evaluation items are paired. Means and untruncated 95% intervals over eight worlds.

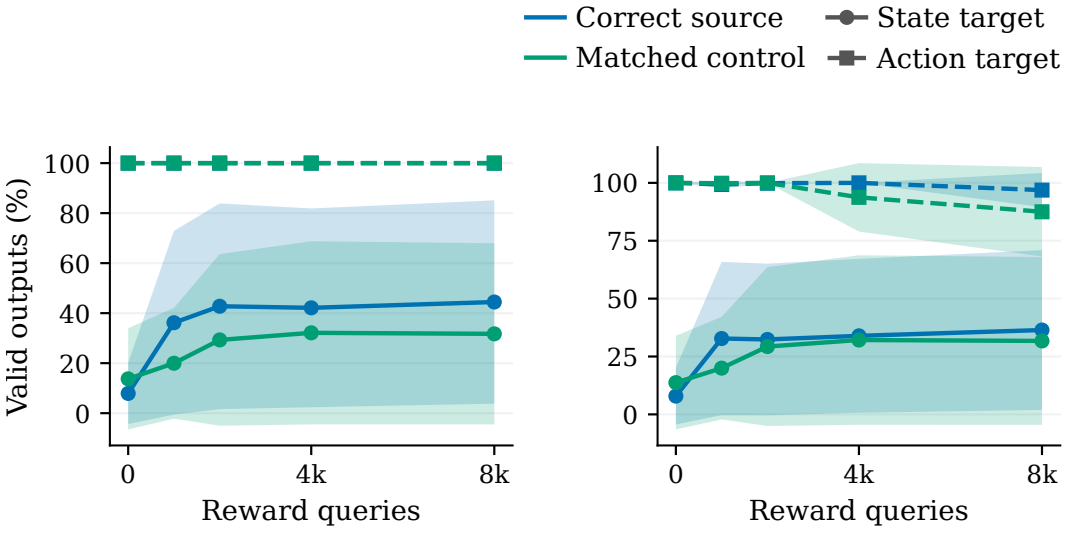


Figure 8: (a) Legal first-operation output and (b) complete syntax on the fixed held-out reward inputs. Four source routes, A/B averaged within world, with 95% intervals across eight worlds. Correct syntax and task correctness are separate metrics.

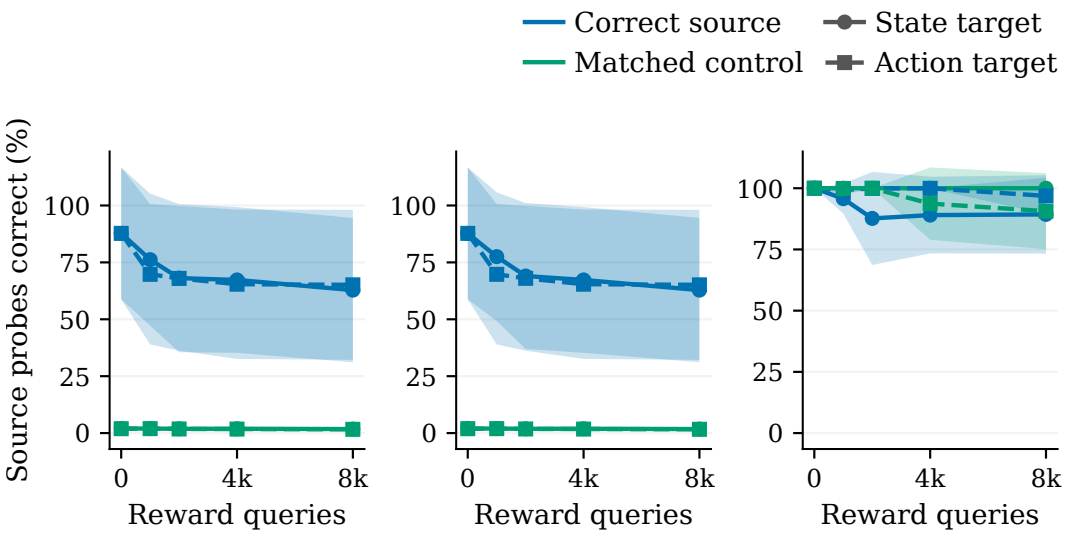


Figure 9: (a) Strict complete source trajectory, (b) required state-prefix content and (c) termination, on 512 fixed source probes per world with the first operation supplied. A/B are averaged within world; bands are 95% intervals over eight worlds.

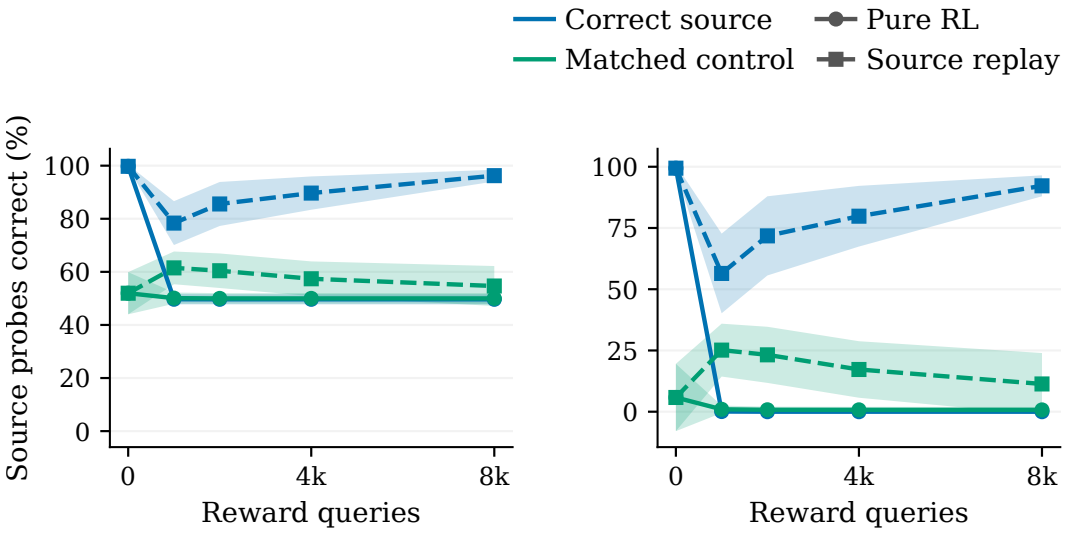


Figure 10: (a) Correct first field on 256 source records and (b) joint correctness on 256 original/relevant-flip pairs. A/B are averaged within each of eight worlds; bands are untruncated 95% intervals. Pure/replay branches share ancestors; replay contributes 131,072 additional source targets.

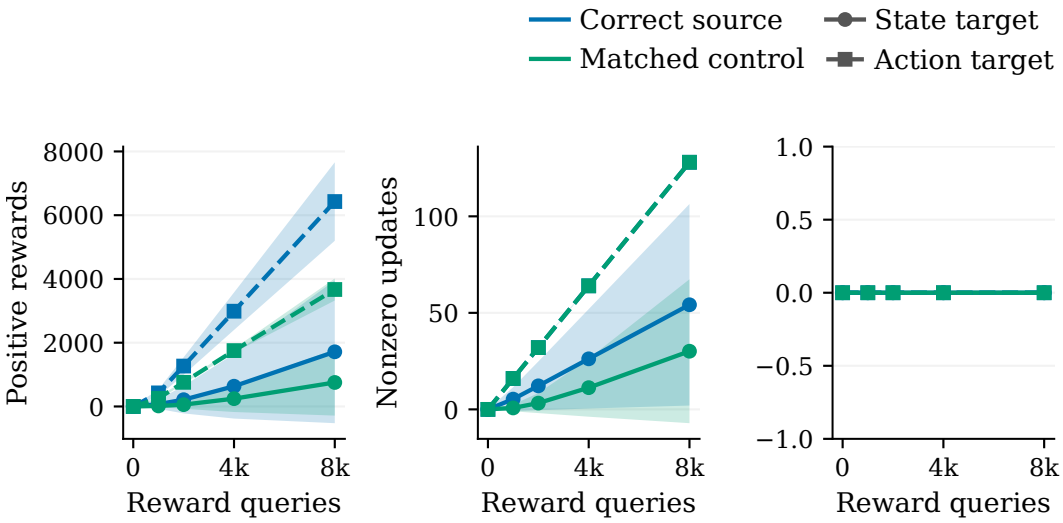


Figure 11: Sequential training diagnostics: (a) positive training rewards, (b) nonzero RL updates and (c) nonzero source CE updates (zero for these routes). Cumulative counts average A/B within world, then eight worlds, with 95% intervals. Logged calls are distinguished from measured nonzero updates.

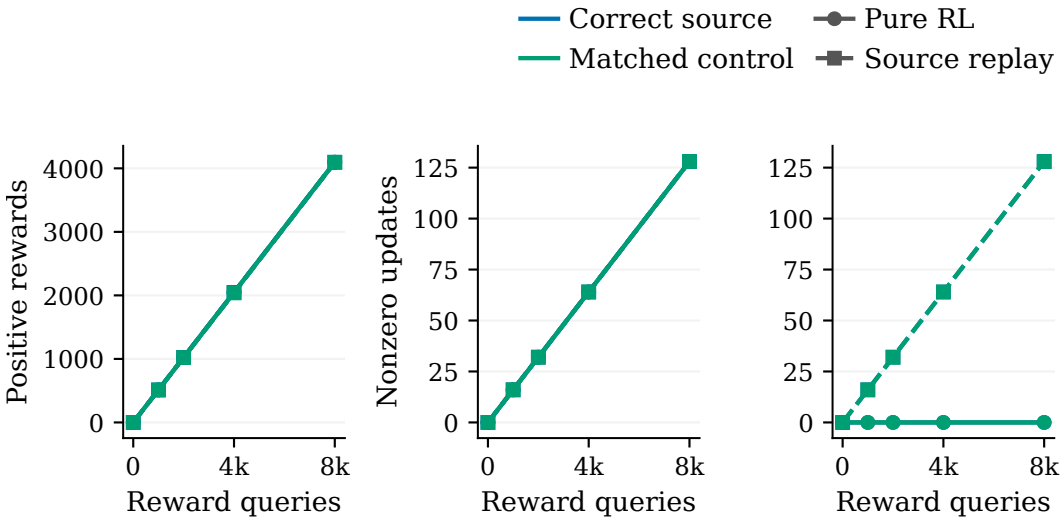
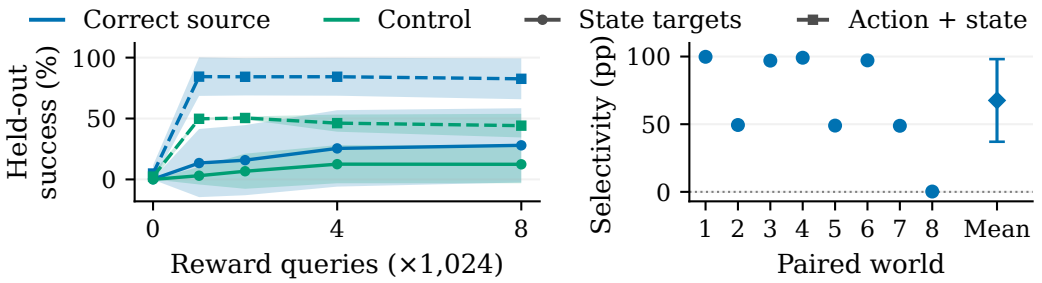


Figure 12: Memory training diagnostics: (a) positive training rewards, (b) nonzero RL updates and (c) nonzero source CE updates. A/B are averaged within world, with 95% intervals over eight worlds. CE calls do not increment the verifier-query count.



(a) Reward learning (b) Source-selectivity effect

Figure 13: Complete original Sequential comparison. (a) State-only and action-plus-state source objectives, correct and private-random source, on 1,024 fixed held-out records. (b) Action-objective source-selectivity effects for all eight worlds, with their mean and 95% interval. A/B are averaged within world. World-level results correspond to Tables 3 and 4.

C.2 Independent Memory confirmation: retrieval that rewards can use

We test task-relevant retrieval on eight new worlds, 508–515, using Qwen2.5-0.5B and the discrete Memory representation. Each world has four reward branches: selected/other retrieval crossed with opposite tasks A/B. The two interventions start from the same correct-source checkpoint and receive matched additional training; A/B then branch from the same intervention endpoint. The experiment asks whether preparing the relevant retrieval computation supports reward-task selection under a fixed adaptation recipe.

Source interventions. After format preparation and 2^{24} correct-source tokens, each route observes 131,072 additional records in 2,048 batches of 64. The selected route predicts the sign of the sum of the 15 content coordinates in the task-relevant memory slot; the other route predicts the same public projection of the other slot. Prompts are identical across routes. These targets depend on the retrieval mechanism, without the downstream contextual signs or cue-rule coefficients. Both routes therefore inherit correct source information; the comparison changes which retrieval computation the additional stage trains.

A second source stage uses 65,536 record exposures in 1,024 batches of 64, adding a retrieval-vector objective to next-token CE. For context $c \in \{1, \dots, 4\}$, the target $z \in \mathbb{R}^{60}$ places the retrieved 15-coordinate value in block c and sets the other blocks to zero. The route is inferred from observed source input/bit pairs. With final-prefix hidden state h, a fixed, data-independent 60×896 orthonormal-row projection P (NumPy seed 970506) predicts $P(h/\operatorname{RMS}(h))$; the auxiliary loss is $\|P(h/\operatorname{RMS}(h)) - z\|^2/15$, averaged over records and added to CE with coefficient one. Here $\operatorname{RMS}(h) = \sqrt{\max(\|h\|^2/d, 10^{-12})}$. The projection is a training buffer, absent from generated answers. These stages train all model parameters at nominal learning rate 10^{-5}; the vector stage uses the step safeguard below.

Reward adaptation and representation retention. Each branch uses 8,192 fresh on-policy reward queries: 128 batches of 64, REINFORCE baseline $1/2$, and nominal Adam learning rate 3×10^{-6}. Sampling and scoring use the same full-vocabulary autoregressive policy. After each reward update, 256 previously observed prefixes supply the same retrieval-vector loss and a syntax objective. Each prefix is paired with both supplied signs; CE supervises only the three public suffix tokens after each sign. Thus this step uses 1,536 CE targets and 15,360 vector components, without CE supervision of the answer sign. Adam moments are reset at stage boundaries and shared by reward and retention updates within each branch.

At the intervention endpoint, temperature is calibrated using 256 unlabeled task-training prompts to mean conditional sign entropy 0.5 bit, then fixed throughout adaptation. If the original next-token probabilities of the two sign tokens are p_+, p_-, write $m = p_+ + p_-$ and $d = \log(p_+/p_-)$. The processed probabilities are

$$p'_+ = m\,\sigma(d/T), \qquad p'_- = m\,[1 - \sigma(d/T)],$$

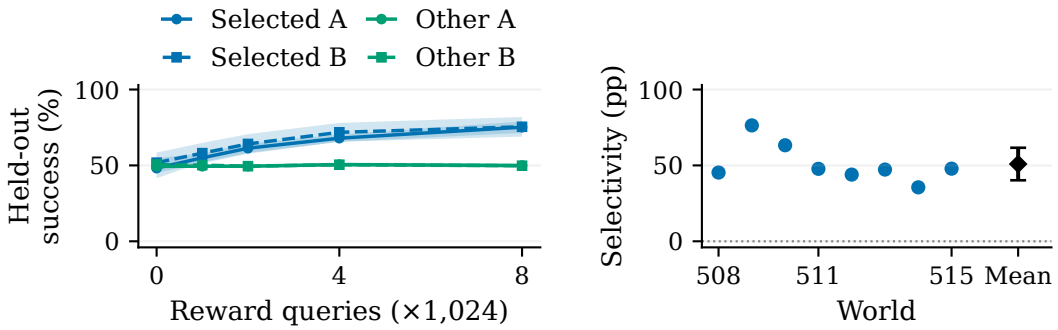
with all other token probabilities unchanged. Only the first sign position is transformed. The temperatures range from 3.7381 to 4.8753; training samples, sequence log probabilities and greedy evaluation use the same processing. Calibration uses no reward labels or held-out examples.

For each guarded Adam proposal, the displacement is scaled by the first accepted value in $\{1, 10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 0\}$. Acceptance requires finite values, loss at most its pre-update value plus 10^{-7}, and mean first-position full-vocabulary $D_{\mathrm{KL}}(\pi_{\mathrm{old}} \| \pi_{\mathrm{new}}) \leq .01$ on the update prompts. The guard scales parameter displacement; Adam's updated moments are retained. Every branch makes 128 actual reward updates; 121–128 of its 128 retention attempts change parameters. These settings define the complete recipe used for both source interventions.

Budgets. The shared warm start uses 16,777,216 observed source tokens. Per intervention, the public-projection stage adds approximately 20.56–20.60 million observed tokens in 131,072 records; the vector stage adds 65,536 record exposures. Per reward branch, retention adds 196,608 CE targets and 1,966,080 vector components in 128 attempts, separately from 8,192 reward queries. The 32 branches use 262,144 new training queries in total. These costs are additional to, and separately recorded from, the original experiment in Section C.1. The contrast evaluates task-relevant versus alternative retrieval under the complete recipe; individual components are not separately ablated.

Evaluation and paired inference. Each branch is evaluated greedily at 0, 1,024, 2,048, 4,096 and 8,192 reward queries on 1,024 fixed held-out records and 256 separate training-distribution records. Both task verifiers score each output. A/B scores are averaged within world before taking world means or differences; the eight worlds are the independent units. The single predeclared primary effect is selected-minus-other task selectivity Δ_{task} as defined in Section 6. Its exact two-sided sign-flip test enumerates all 2^8 sign assignments and remains separate from the earlier two-test Holm family. Intervals use Student's t with seven degrees of freedom.

Selected retrieval reaches 75.32% held-out success, compared with 49.86% for other retrieval; both routes enter reward training at 50.00% after averaging A/B within world. The paired accuracy difference is 25.46 pp (95% CI 20.10–30.81). All eight world differences are positive. The primary selectivity difference is 50.92 pp (95% CI 40.20–61.63), exact $p = 2/256 = .0078125$. All 32 branches have 100% format accuracy on their final evaluation records. With valid binary answers and opposite task labels, $J_{AB} = 1 - J_{AA}$ and $J_{BA} = 1 - J_{BB}$, so the selectivity difference is exactly twice the accuracy difference. Figure 14 and Table 6 display every world and both task branches.



(a) Separate A/B branches

(b) Selected minus other

Figure 14: Independent Memory confirmation. (a) Separate task-A and task-B success curves for selected and other retrieval; means and 95% Student-t bands over eight worlds, using 1,024 held-out records per branch. (b) Each point is a world's selected-minus-other task-selectivity effect after averaging A/B; black diamond and error bar give the mean and 95% interval. All fixed evaluation points and all worlds are retained.

Table 6: Independent Memory confirmation: all 32 branch endpoints after 8,192 reward queries. Success is a percentage of 1,024 fixed held-out records per branch. Differences first average A/B within world; all eight worlds are included.

World	Selected A	Selected B	Other A	Other B	Accuracy Δ	Selectivity Δ
508	73.34	75.10	52.44	50.68	22.66	45.31
509	83.69	91.31	49.32	49.32	38.18	76.37
510	80.66	81.84	49.80	49.41	31.64	63.28
511	70.90	73.05	47.56	48.63	23.88	47.75
512	73.54	71.09	50.10	50.59	21.97	43.95
513	73.83	72.27	48.83	50.00	23.63	47.27
514	73.83	66.31	52.54	52.05	17.77	35.55
515	71.68	72.66	48.73	47.75	23.93	47.85

C.3 Independent Sequential parameter-plane measurement

This descriptive comparison uses the prespecified world 508 and Task A, separately from the formal worlds 500–507. Three routes share a Qwen2.5-0.5B checkpoint after 128 task-independent format updates. Direct RL skips added source training. Correct source and the public-preserving/private-random control each use 2^{22} observed tokens and 512 source updates, predicting the record's first operation, subsequent states and END/EOS. The input initial state and suffix operations are not supervised targets under this objective. Each reward branch uses learning rate 3×10^{-6}, batch size 64 and 128 updates, totaling 8,192 verifier queries. The three branches make 128 nonzero updates each and receive 3,874, 7,668 and 3,786 positive training rewards, respectively. Additional source cost is separate from the common inherited checkpoint and equal reward budgets.

Actual learning and evaluation units. Figure 15 reports greedy strict success on 1,024 fixed held-out records, balanced between suffix lengths four and six, at 0, 1,024, 2,048, 4,096 and 8,192 reward queries. Source-endpoint task success is distinct from a source-computation probe: the correct-source model has 0% task success before RL and 100% after 1,024 queries. Final advantages are 48.24 percentage points over direct RL and 48.73 over the private-random source control. The bands use 2,000 item-bootstrap resamples stratified by suffix length, with shared indices across routes and checkpoints. These intervals describe item uncertainty conditional on the prespecified world and trained run; they collapse at all-correct or all-incorrect endpoints. This single-world comparison is analyzed separately from the eight-world tests.

One common plane and complete measurement grid. Let ϑ_0 denote the shared format checkpoint. We use 17 distinct checkpoints: the origin, both source routes at 2^{20} and 2^{22} tokens, and all three reward routes at 1,024, 2,048, 4,096 and 8,192 queries. Shared source-to-reward boundaries use the same weights and are counted once. An uncentered rank-two SVD of displacements $\vartheta_i - \vartheta_0$ across all these states gives orthonormal directions v_1, v_2. A checkpoint is projected to coordinates $\alpha_{ij} = \langle \vartheta_i - \vartheta_0, v_j \rangle$ and reconstructed as $\widetilde{\vartheta}_i = \vartheta_0 + \alpha_{i1} v_1 + \alpha_{i2} v_2$. Grid models use the same affine plane and are cast to the evaluation model's dtype.

We evaluate the 81 points of the fixed 9×9 grid, all 17 original checkpoints and their 17 reconstructions. Every measurement uses the same 256 held-out inputs (128 of each suffix length) and 128 fixed correct-source records. Source loss is the summed negative log-likelihood over 3,856 supervised first-operation/state/END/EOS targets divided by that target count, in nats per target. The colored cells in Figure 2 are measured grid values; only the white source-loss contours interpolate between grid values.

Geometric coverage and behavioral fidelity. The plane retains 91.96% of $\sum_i \|\vartheta_i - \vartheta_0\|^2$. For the direct-RL endpoint alone, the relative residual $\|\vartheta_i - \widetilde{\vartheta}_i\| / \|\vartheta_i - \vartheta_0\|$ is 89.63%, leaving only 19.67% of its squared displacement in the plane. Aggregate coverage thus masks route-wise reconstruction loss. On the common 256 inputs, actual/reconstructed success is 0.00%/100.00% at the correct-source endpoint, 0.00%/52.73% at the private-random source endpoint, and 53.91%/19.14% at the direct-RL endpoint. The two source-plus-RL endpoints agree at 100.00% and 52.73%, respectively. These denominators differ from the full 1,024-record learning curves.

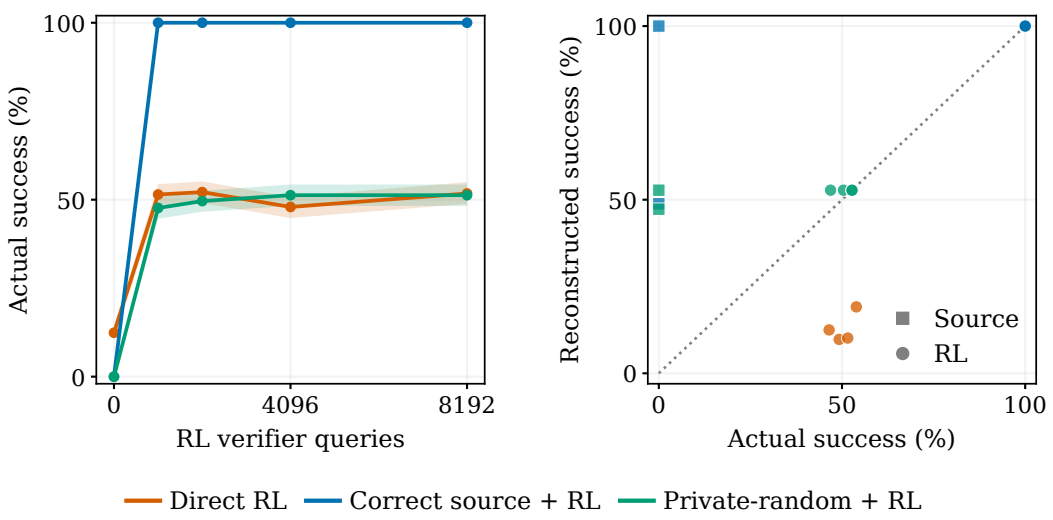


Figure 15: Actual reward learning and projection fidelity in world 508. Left: original-checkpoint success on 1,024 fixed records; bands are conditional 95% item-bootstrap intervals. Right: original versus reconstructed checkpoint success on the same 256 slice-evaluation inputs. Squares mark source checkpoints and circles mark RL checkpoints; overlapping points are retained. The diagonal denotes equal scores. The source routes receive 2^{22} additional tokens before RL; direct RL skips this stage. All three routes share format preparation and use 8,192 reward queries.

Because the basis includes later RL checkpoints, reconstructing a source checkpoint introduces directions learned during reward adaptation. The grid describes the behavior of reconstructed models in this shared plane; the original-checkpoint curves measure the actual training behavior. Together they relate the measured slice to learning on task A, while the paired A/B experiments measure task selectivity.

C.4 Eight-world source acquisition and reward adaptation

This comparison separates learning a source computation from using it under a new reward. Each module uses eight independently generated worlds (seeds 300–307), two source routes and two reward tasks A/B per route. Sequential uses its original representation, comparing the correct source with a source that preserves every public state and randomizes the private bits. Memory uses discrete signed-permutation keys and ± 1 values, comparing correct retrieval with wrong retrieval under the same all-token objective. Inputs and source exposure are matched within each world; A/B branches share the same source checkpoint.

Training and evaluation. All runs use Qwen2.5-0.5B Base and full-parameter Adam. A task-independent format stage precedes source training: 128 updates for Sequential and 32 for Memory, batch 64, learning rate 10^{-5}. Sequential then observes 2^{22} source prediction tokens in 512 updates, optimizing state tokens and the existing END/EOS targets; Memory observes 2^{24} tokens in 2,048 updates, optimizing all tokens. Source learning rate is 10^{-5} for both. No subsequent format stage is applied. Each source checkpoint branches directly into tasks A/B, with fresh Adam moments, fixed learning rate 3×10^{-6} and 8,192 verifier queries (128 batches of 64 sampled completions). The strict reward and output parser are unchanged. Greedy evaluation uses 1,024 fixed held-out prompts at 0, 1,024, 2,048, 4,096 and 8,192 queries; both task verifiers score every output. All 64 reward branches finish the full budget, totaling 524,288 training queries.

Source computation. The source-stage diagnostics in Figure 16 use 512 Sequential prompts per world, with the first operation given and horizons five and seven equally represented. A complete answer contains all correct states and END in exactly the required positions. Memory uses 256 original source records and 256 corresponding relevant-value interventions per world; a pair is correct only when both first-field answers are correct. These are source-distribution probes without the

downstream task binding. Across eight worlds, correct minus control differences are 87.50 percentage points for Sequential complete trajectories (95% paired t interval $[62.89, 112.11]$), 44.48 for Memory original records ($[31.92, 57.05]$), and 88.28 for Memory intervention pairs ($[60.57, 115.99]$). These source diagnostics are distinct from the two primary reward-task comparisons below. The t intervals are untruncated; their endpoints can exceed the bounded metric range.

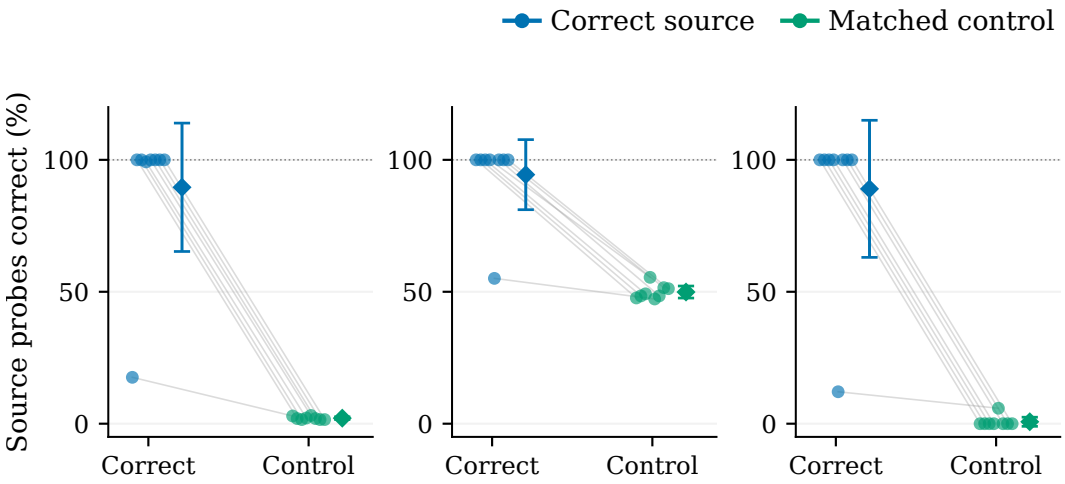


Figure 16: Historical source acquisition, worlds 300–307. (a) Sequential complete trajectories on 512 probes with first operation supplied; (b) Memory first-field correctness and (c) original/value-flip pair correctness, each on 256 records or pairs. Correct source and matched structural controls are paired within world; points are worlds, diamonds means with untruncated 95% Student-t intervals. This protocol is separate from worlds 500–507 in Section C.1.

Task selection. For each route and world, selectivity is $\frac{1}{2}[(J_{AA} - J_{AB}) + (J_{BB} - J_{BA})]$, where J_{XY} is the success of the task-X branch under verifier Y. The two primary effects subtract the matched-control selectivity from correct-source selectivity at 8,192 queries. We average eight world-level differences, use Student-t intervals with seven degrees of freedom, enumerate all 2^8 sign flips for two-sided tests, and apply Holm correction to these two hypotheses. This two-test family is separate from the earlier comparisons. Neither interval excludes zero (Table 7 and Figure 18). Sequential world effects are $[0, 0, 50, 0, -1.123, 0, 0, -1.172]$ percentage points, so its positive mean is concentrated in one world. Full learning curves appear in Figure 17.

Table 7: Correct-source minus matched-control task selectivity after 8,192 reward queries, eight worlds per module. Effects and 95% paired Student-t intervals are in percentage points; p is the exact two-sided sign-flip value and p_H applies Holm correction to the two comparisons.

Module	Effect	95% interval	p	p_H
Sequential	5.96	$[-8.92, 20.85]$	1.000	1.000
Memory	-0.87	$[-2.10, 0.35]$	0.125	0.250

Computation during reward adaptation. Figure 19 repeats the fixed source probes through the reward stage. A/B probe scores are averaged within each world before forming the eight-world mean and interval. Correct-source Sequential complete trajectories change from 89.60% to 88.94%, with END placement at 100% throughout. Twenty-six of the 32 Sequential reward branches receive no positive reward and make no parameter updates, so most branches retain the source checkpoint unchanged. Correct-source Memory first-field accuracy changes from 94.38% to 50.34% and relevant-pair accuracy from 89.01% to 0%, while its source-probe public-format accuracy stays at 100%. All 32 Memory branches make 128 nonzero updates. Thus computation acquired on source inputs and its expression after reward adaptation are separately observable properties.

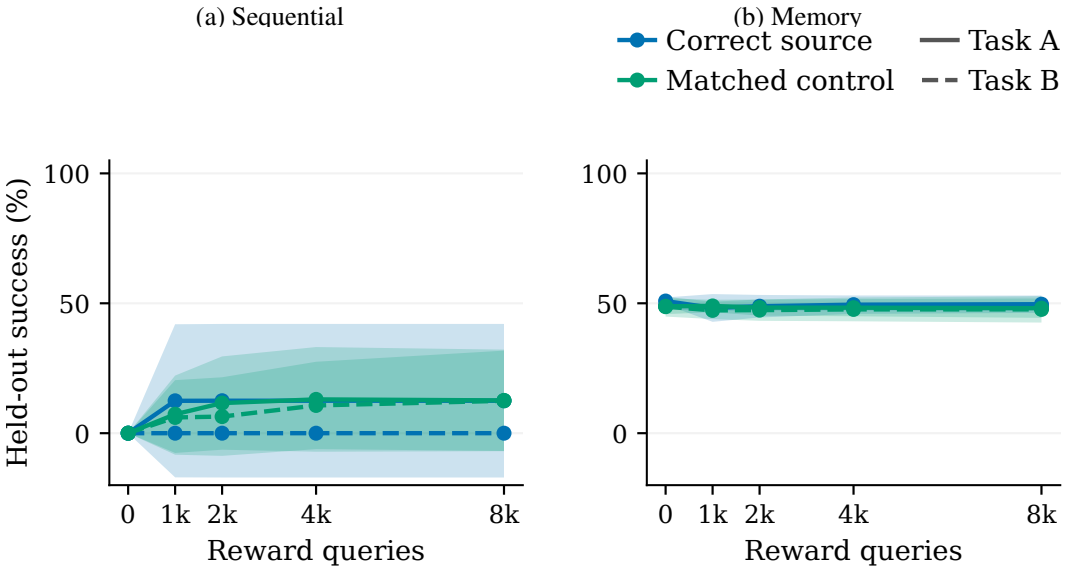


Figure 17: Held-out success during reward adaptation, eight worlds per route. Colors distinguish correct source and the matched control (random private states for Sequential; wrong retrieval for Memory); solid and dashed lines denote tasks A and B. Markers are world means on 1,024 prompts per world at each query budget, with 95% Student-t bands. Training samples at temperature one; evaluation is greedy. Intervals are shown without clipping to the success-rate range.

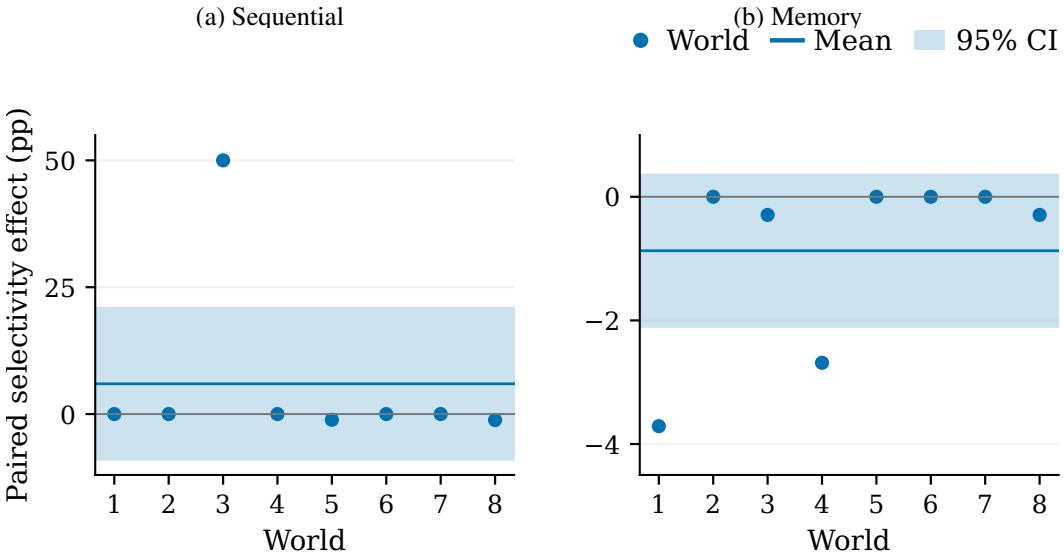


Figure 18: Correct-source minus matched-control task selectivity at 8,192 queries. Each dot is one complete world-level A/B contrast; worlds 1–8 correspond to seeds 300–307. The blue line is their mean, the shaded band its 95% Student-t interval, and the gray line marks zero. Units are percentage points. Both comparisons include all eight worlds; vertical scales differ between panels.

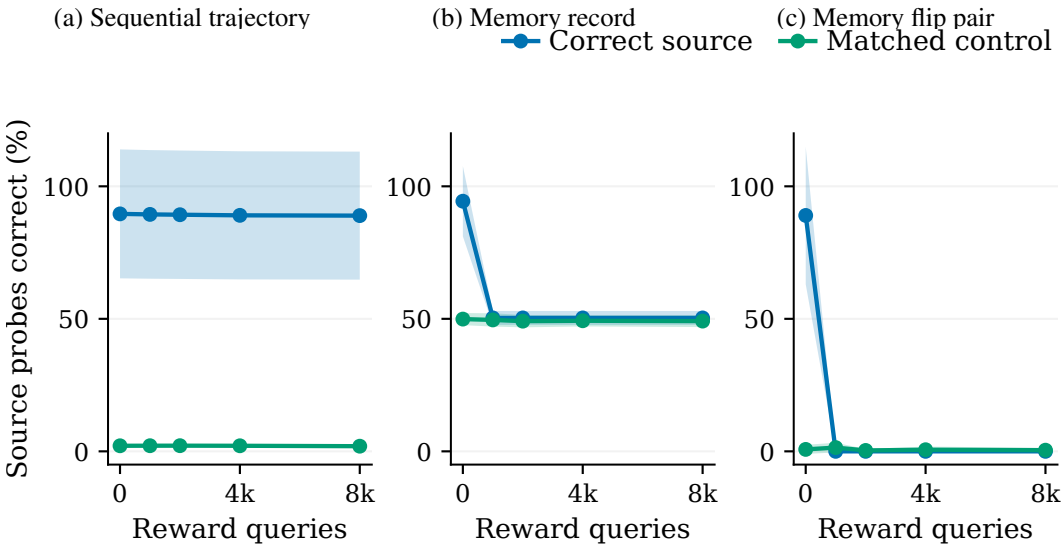


Figure 19: Source-computation probes through reward adaptation, using the same metrics and matched controls as Figure 16. Lines join the five measured query budgets; shaded bands are 95% Student-t intervals over eight worlds after averaging A/B within each world. Shared ancestors are not counted as additional independent samples. Sequential probes supply the first operation; Memory pairs require both the original and relevant-value-flipped bit to be correct.

The following subsections describe the earlier controlled-task and public-task comparisons; the eight-world source-acquisition design above has its own source objectives and budgets.

C.5 Earlier execution, task-selection and interface comparisons

These earlier comparisons identify which part of learning changes with source preparation. They use their own objectives, budgets and checkpoints, described below.

Execution and task choice. The endpoint decomposition separates a valid public trajectory from its private answer, while opposite-reward branches test task-specific selection (Figure 20 and Sections C.8 and C.12). This distinguishes learning to execute from learning which execution the task requires.

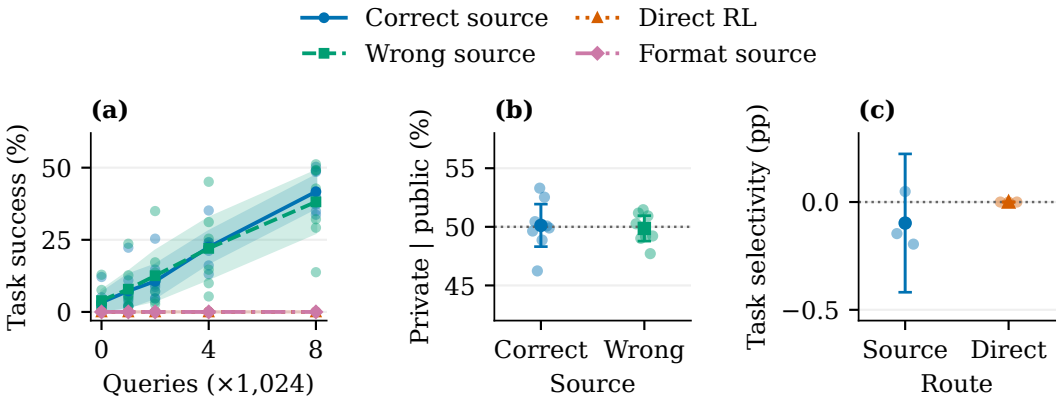


Figure 20: Historical Sequential execution and private-task adaptation, using 2^{20} source prediction tokens. (a) Greedy success on 1,024 held-out prompts per world: eight worlds for correct source, wrong source and direct RL, and three for format control. (b) Final private-bit accuracy among outputs with valid public execution and the correct public endpoint, computed within each world before averaging; direct and format routes have no eligible outputs. (c) Shared-source A/B branches in three worlds: paired task selectivity Δ_{task} in percentage points (pp). Small translucent dots show individual worlds; connected markers in (a) and large markers in (b,c) show world means. Shading and error bars are 95% Student-t intervals across worlds; horizontal jitter separates overlapping points. Dotted lines mark 50% private accuracy in (b) and zero task selectivity in (c).

Acquisition and stage order. Memory source probes test the retrieved-value rule directly (Section C.8). The separate two-world study measures state content, termination and value-flip responses before and after format training (Section C.14). Together these diagnostics locate the computation available at each stage.

Initialization, budgets and learning routes. Section C.13 compares source dose, inherited versus random initialization, model scale, continued prediction, answer supervision and test-time selection. Section C.12 gives the output-level decomposition and sampled-decoding comparison, with the full denominators and uncertainty.

C.6 Task definitions and observation permissions

For sequential execution, a state is $(s, z) \in \{0,1\} \times \mathbb{F}_2^3$, and an operation (b, u) maps it to $(s \oplus b \oplus \theta(u, z), z \oplus u)$. Here $\oplus$ is addition modulo two. Writing $u = (t, x, y)$, let $a(u) = t \oplus xy$. Displayed suffixes have length $L \in \{4, 6\}$; the actual execution horizon is $H = L+1$, including the hidden initial operation. Training and validation draw independently from suffixes with $\bigoplus_j a(u_j) = 0$; held-out suffixes have parity one. The paired mechanism replaces $\theta(u, z)$ by $\theta(u, z) \oplus a(u)$ and compensates the hidden initial binding. The terminal training-reward function is unchanged. Evaluation has 512 prompts at each length.

A source record serializes the initial state, the operation suffix, the independently uniform first operation, and the resulting states. At each transition a reset occurs with probability $(1/64)/H$, replacing the state uniformly. Every source token is a next-token prediction target, including random fields and EOS. In textual outputs, END is an explicit field terminator, distinct from the tokenizer's end-of-sequence marker. The verifier checks output syntax, the public z trajectory, and the final state; correct intermediate answer bits are not required. Source data do not contain the downstream hidden binding.

The contextual memory experiment uses four hidden orientations and a 15-dimensional shared cue rule, as in the contextual model. Source records contain random keys, a transformed query, Gaussian values, and the raw future bit, with no task-dependent signs. Target memory amplitudes separate training from four equally weighted test configurations $(\pm 1, \pm 0.875)$. Equal-sign and opposite-sign subsets are reported separately. Language-model inputs use textual KEY, QUERY, VALUE, and BIT fields, with numeric values rendered to eight significant digits. The language-model objective predicts all serialized fields; the contextual theorem trains on the future-bit position of typed numerical input.

The information panel uses eight paired worlds and nested source record counts $0, 1, 2, 4, 8, 16, 32, 64$. The contextual curve $\frac{1}{4}(\frac{3}{4})^N$ is the analytic expected error floor; the measured coverage statistic is the number of contexts actually observed. The paired-world construction is verified by enumeration.

C.7 Models, training and evaluation

The models are Qwen2.5-0.5B Base and Qwen2.5-1.5B Base, with all parameters trainable; Table 8 lists the comparisons and their independent seeds. Source training uses 8,192 prediction targets per update, with saved states at $2^{18}, 2^{19}, 2^{20}$ targets. Feedback uses one fresh batch of 64 sampled completions per REINFORCE update, for 8,192 verifications in total. Its loss is the negative batch mean of reward times the sum of sampled completion log probabilities, including EOS, with baseline zero; zero-reward batches still advance Adam moments. There is no reward filtering, Kullback–Leibler (KL) penalty, weight decay or gradient clipping. Parameters and Adam moments are stored in bfloat16 with 32-bit residual, normalization and probability computations; sampled completions are produced by SGLang, with trainer-side response log probabilities recomputed for the update. Contexts are limited to 2,048 tokens and continuations to 128 new tokens. Sampling uses temperature one without top-k or top-p truncation. Feedback begins from the source parameters with reset moments.

Learning rates are selected on two pilot worlds disjoint from the evaluated seeds, by source validation cross-entropy over $\{10^{-5}, 3 \cdot 10^{-5}, 10^{-4}\}$ and then by validation feedback on the allowed training support over $\{10^{-6}, 3 \cdot 10^{-6}, 10^{-5}\}$, with smaller rates breaking ties. The selected rates (Table 9) are shared by all routes within a task, model and initialization. Evaluation occurs at $0, 1024, 2048, 4096, 8192$ training queries. Each checkpoint has 1,024 fixed paired test prompts, evaluated with both greedy and sampled decoding; raw outputs, success, public-path and format errors, and truncation are saved. BBH uses every item from the two three-object tasks, a zero-shot option-only prompt, greedy decoding, and a strict option parser at the base, source, and final reward checkpoints.

Table 8: Experimental comparisons. Source-only measurements reuse source checkpoints; the fixed external base model contributes one checkpoint.

Comparison	Independent seeds and scope
Direct RL, correct source, paired wrong source	8 sequential and 5 contextual, 0.5B
Random initialization and 1.5B transfer	3 sequential seeds per three-arm panel
Source dose	5 sequential seeds, 0.5B, shared source
Alternative learning routes	3 sequential LM seeds
Alternate downstream task choice	3 sequential seeds, shared source states
Format-only source then RL	3 sequential seeds
BBH checkpoint evaluation	Base, source, feedback; 3 source/feedback seeds

Table 9: Adam learning rates used in the historical controlled and public-task comparisons, selected on pilot worlds by source validation cross-entropy and training-support validation feedback. Source and reward stages reset optimizer moments.

Task / model / initialization	Source rate	Reward rate
Contextual memory / 0.5B / pretrained	3×10^{-5}	10^{-5}
Sequential / 0.5B / pretrained	3×10^{-5}	10^{-6}
Sequential / 0.5B / random	10^{-4}	10^{-6}
Sequential / 1.5B / pretrained	3×10^{-5}	10^{-6}
GSM8K / 1.5B / pretrained	10^{-5}	10^{-5}
HotpotQA / 1.5B / pretrained	3×10^{-5}	10^{-5}

C.8 Alternative routes and controls

Continued source prediction consumes the next 2^{20} targets with the source optimizer state continued. Interleaved source updates add one 1,024-target source cross-entropy update after each reward update, drawing only from observed source records. Answer-only SFT generates its own prefix and supervises just the terminal answer and END; it receives no expert intermediate states, and answer labels are counted separately from verifications. Best-of-N reuses nested sets of $N = 1, 2, 4, 8$ candidates from a fixed source model with test-time verification. The classical comparator estimates transitions from the same complete observed source records and searches the hidden binding with allowed training feedback. Source balancing changes sampling over initial public contexts at the same token budget.

Format exposure without a shared transition rule. The format control uses three sequential Qwen2.5-0.5B worlds. It reuses the source input law: the initial state, suffix, independent first control, vocabulary, field positions, sequence lengths and END marker. Every output-state token is instead an independent uniform element of the legal state space. The generator reads neither the target transition table nor the downstream binding. Thus it preserves the format and the state marginal distribution while removing any persistent state–operation relation. The mechanism-sensitive contrast is $G_{\text{source}} - G_{\text{format}}$.

Shared-source, opposite-task feedback branches. For three sequential worlds, write the hidden first operation as $\eta = (b, u)$. Task A uses the original η and task B uses $(b \oplus 1, u)$, with identical θ. The source generator does not read η and the displayed prompts omit it. Every public coordinate is unchanged, while the terminal private bit flips under the same suffix, so the two success sets are disjoint. Task-B direct RL and correct-source RL branches are paired with the task-A branches; correct-source branches share the exact source checkpoint. Both tasks' scoring functions are applied offline to the same outputs at zero and final queries; Table 19 reports the cross-scored rates and Table 20 the exact eligible counts $C = N_A + N_B$ behind them.

The paired selectivity is -0.10 percentage points with interval $[-0.42, 0.22]$ (Figure 20 and Tables 19 and 20). Together with the conditional private accuracy in Section C.12, this identifies shared public execution without a measurable preference for the branch's reward task.

Memory-rule acquisition probes. Per-update answer frequencies locate input-independent behavior, and training label frequencies distinguish it from a majority-label task. Two matched source runs, with the same source token budget, learning rate and Adam schedule as the historical continuous-numeric Memory runs, are probed before and after source training: 32 fixed held-out source records each generate three input domains and three value interventions, giving 288 observations per checkpoint. The domains use the original random key frames, basis keys, or basis keys with discrete nuisance coordinates; the first two content coordinates retain the source parity rule. Negating the selected slot's first coordinate reverses the future bit, whereas the same intervention on the unselected slot preserves it. For every input we record the log-likelihoods of the two valid bit completions (including the common terminator), their difference, two-candidate accuracy, and loss by source field. A correct relevant-value response together with invariance to irrelevant values indicates rule acquisition.

Table 10: Source-rule diagnostics from two matched memory source runs. Intervention triples share an original record. Fractions report correct cases over the displayed denominator; CE denotes cross-entropy.

Measurement	Run A	Run B
Two-candidate completion accuracy	46/96	49/96
Original records correct	14/32	17/32
Prediction reverses on relevant-value flip	1/32	0/32
Original and relevant-flip pair both correct	0/32	0/32
Query-field CE, before	0.2045	0.2114
Query-field CE, after	0.0430	0.0397
Bit-field CE per token, after	0.3483	0.3869

Each group of 96 cases consists of 32 records and their relevant and irrelevant interventions. The answer field contains two predicted tokens; normalizing the two completion likelihoods gives binary cross-entropies of 0.6949 and 0.7705. The lower query-field loss shows learning of predictable source content, but the pair test shows that these models do not reliably express the composed retrieval-and-sign-product response. After feedback, all five correct-source runs produce a constant valid action, with 49.18% greedy correctness on balanced task labels; wrong-source runs share this endpoint (Table 21). Comparing source and feedback checkpoints separates predictable-field learning from input-dependent task responses.

C.9 Uncertainty

Means and 95% Student-t intervals are computed across world/training seed means, with individual seed values retained; paired differences use the same seeds. The sampled sequential correct-source-versus-direct and correct-source-versus-wrong-source endpoint contrasts additionally use exact two-sided sign-flip tests with Holm correction at 0.05 (Table 24). The sign-flip tests assume sign symmetry of paired effects under the null. Holm adjustment uses a fixed family size of three, with $p = 1$ assigned to the remaining slot; the adjusted values for the two reported contrasts are 0.0234 and 0.672. Secondary comparisons report effect sizes and intervals. Separately, 5,000 paired item-bootstrap resamples quantify uncertainty conditional on the trained models.

C.10 Public-task adaptation and feedback

Public-task source midtraining uses domain-related text without downstream labels; it does not instantiate the exact rule-independence assumption of the theoretical source laws. Reported budgets cover additional source tokens and reward queries, excluding inherited pretraining. The public-task panel uses five paired Qwen2.5-1.5B seeds for each of GSM8K and HotpotQA, with three reward routes: direct, domain-source and shuffled-source learning. Each source trajectory predicts 2^{22} tokens and retains 2^{20} and 2^{22} checkpoints. Shuffling acts within each document's 64-token blocks, preserving the target-token multiset. Source and feedback learning rates use the same grids and selection rule as the controlled experiments, with two pilot seeds per task; selection uses source validation cross-entropy and training-distribution validation rewards.

The numeric answer parser compares final GSM8K answers; HotpotQA uses normalized exact match for reward and official answer F1 as an additional measurement. Mathematical continuations have at most 512 new tokens, and question-answering continuations at most 256. HotpotQA prompts are admitted by tokenized length at most 4,096; all 7,405 development candidates qualify, and a fixed 1,024-question subset is evaluated. All 1,319 GSM8K test questions are evaluated. Source article titles are disjoint from downstream training and evaluation article titles, and question and document identities, dataset versions and tokenization are fixed before training.

Reward queries are evaluated at $0, 1024, 2048, 4096, 8192$. The three intermediate points use 512 paired prompts; the pre-RL and final evaluations use the full 1,319 or 1,024 items, and learning curves use the common 512 prompts at every checkpoint. Each task also has three answer-only SFT seeds initialized from the corresponding domain-source state, supervising only the answer and EOS

with 8,192 answer labels and the same update batch. Fixed-model best-of-16 uses the same 256 test prompts for base and domain-source models. Public transfer uses 100 GSM-Symbolic templates with ten fixed variants each, and every item in the two three-object BBH tasks; symbolic variants are repeated observations of a template. Public results show every seed, means and 95% Student-t intervals over seed-level outcomes and paired gains; conditional item-bootstrap intervals preserve item pairing, and GSM-Symbolic resamples templates. Figure 21 shows the entry scores, endpoints and paired gain differences summarized in Table 2, and Figure 23 the greedy learning curves on the common 512 prompts, which locate where along the 8,192 queries each route's gain accrues.

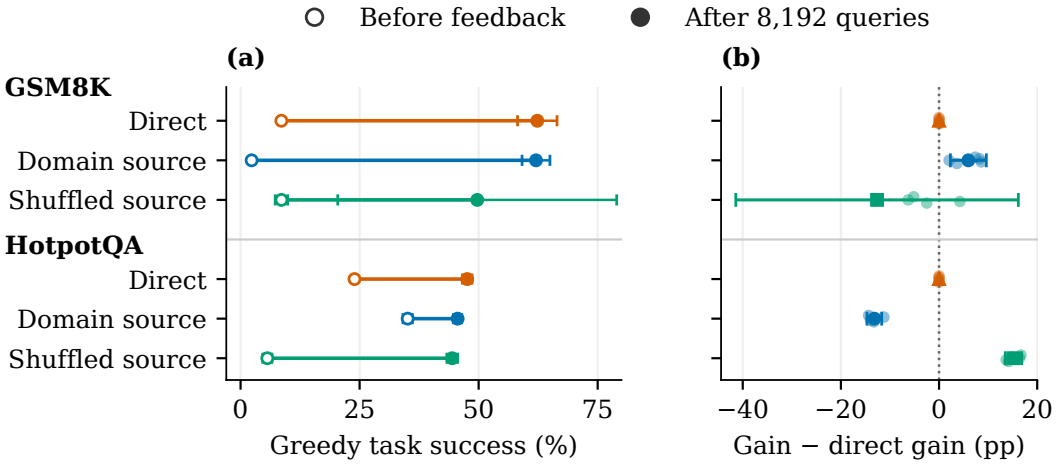


Figure 21: Public-task starting scores and reward gains, five paired training seeds. (a) Open and filled markers show mean greedy success before and after RL; segments connect the same route. (b) Within-seed reward gain minus direct RL's gain, in percentage points: small translucent dots are seeds, large markers are means, and the dotted line is zero. Horizontal bars are 95% Student-t intervals. Each seed and route uses the same full endpoint sets: 1,319 GSM8K and 1,024 HotpotQA prompts. Direct starting scores evaluate one fixed pretrained checkpoint.

C.11 Secondary public-task metrics and external transfer

Success, numeric correctness and F1 are higher-is-better; format errors and truncation are lower-is-better. Uncertainty describes training-seed variation.

Table 11: Public-task secondary greedy measurements. Percentages apply to correctness, F1 and error rates; generated length is tokens. RL uses five seeds and answer-only SFT three. F1 is answer token-overlap F1; format and truncation rates use all generated answers. Entries are seed means with 95% Student-t half-widths.

Task	Method	Decoding	Metric	Mean and 95% CI
GSM8K	Direct RL	greedy	numeric correct	62.3 ±4.1
GSM8K	Direct RL	greedy	Format error	0.5 ±0.2
GSM8K	Direct RL	greedy	Length (tokens)	108.1 ±8.8
GSM8K	Direct RL	greedy	truncated	0.3 ±0.1
GSM8K	Correct source	greedy	numeric correct	62.0 ±2.9
GSM8K	Correct source	greedy	Format error	2.2 ±4.2
GSM8K	Correct source	greedy	Length (tokens)	149.8 ±101.5
GSM8K	Correct source	greedy	truncated	1.6 ±3.4
GSM8K	Shuffled source	greedy	numeric correct	49.7 ±29.3
GSM8K	Shuffled source	greedy	Format error	17.8 ±47.5
GSM8K	Shuffled source	greedy	Length (tokens)	177.8 ±210.9
GSM8K	Shuffled source	greedy	truncated	17.7 ±47.6

Continued on next page

Table 11: Public-task secondary greedy measurements. Percentages apply to correctness, F1 and error rates; generated length is tokens. (continued)

Task	Method	Decoding	Metric	Mean and 95% CI
GSM8K	Answer SFT	greedy	numeric correct	12.4 ± 0.4
GSM8K	Answer SFT	greedy	Format error	0.0 ± 0.0
GSM8K	Answer SFT	greedy	Length (tokens)	4.5 ± 0.0
GSM8K	Answer SFT	greedy	truncated	0.0 ± 0.0
HotpotQA	Direct RL	greedy	Answer F1	60.4 ± 1.6
HotpotQA	Direct RL	greedy	Format error	0.0 ± 0.0
HotpotQA	Direct RL	greedy	Length (tokens)	5.3 ± 0.3
HotpotQA	Direct RL	greedy	truncated	0.0 ± 0.0
HotpotQA	Correct source	greedy	Answer F1	58.4 ± 1.1
HotpotQA	Correct source	greedy	Format error	0.0 ± 0.0
HotpotQA	Correct source	greedy	Length (tokens)	5.6 ± 0.2
HotpotQA	Correct source	greedy	truncated	0.0 ± 0.0
HotpotQA	Shuffled source	greedy	Answer F1	57.4 ± 1.4
HotpotQA	Shuffled source	greedy	Format error	0.0 ± 0.0
HotpotQA	Shuffled source	greedy	Length (tokens)	5.5 ± 0.4
HotpotQA	Shuffled source	greedy	truncated	0.0 ± 0.0
HotpotQA	Answer SFT	greedy	Answer F1	69.0 ± 0.5
HotpotQA	Answer SFT	greedy	Format error	0.0 ± 0.0
HotpotQA	Answer SFT	greedy	Length (tokens)	4.9 ± 0.2
HotpotQA	Answer SFT	greedy	truncated	0.0 ± 0.0

Table 12: HotpotQA greedy results within each indicated question-type subset. Success is normalized exact match; rates are percentages with seed-level 95% Student-t half-widths.

Method	Decoding	Type	Metric	Mean and 95% CI
Direct RL	greedy	bridge	Answer F1	60.5 ± 1.1
Direct RL	greedy	bridge	success	46.2 ± 0.4
Direct RL	greedy	comparison	Answer F1	60.2 ± 3.5
Direct RL	greedy	comparison	success	52.9 ± 3.7
Correct source	greedy	bridge	Answer F1	58.3 ± 1.1
Correct source	greedy	bridge	success	44.2 ± 0.9
Correct source	greedy	comparison	Answer F1	58.9 ± 2.4
Correct source	greedy	comparison	success	51.2 ± 2.6
Shuffled source	greedy	bridge	Answer F1	57.3 ± 1.3
Shuffled source	greedy	bridge	success	43.0 ± 1.0
Shuffled source	greedy	comparison	Answer F1	57.7 ± 2.4
Shuffled source	greedy	comparison	success	50.0 ± 2.4
Answer SFT	greedy	bridge	Answer F1	70.0 ± 1.1
Answer SFT	greedy	bridge	success	52.9 ± 2.2
Answer SFT	greedy	comparison	Answer F1	65.0 ± 3.2
Answer SFT	greedy	comparison	success	56.9 ± 1.8

Table 13: External-transfer and verification-selection results. Bo16 denotes best-of-16 selection with test-time verification. ± gives training-seed 95% Student-t half-widths; brackets give item/template-bootstrap intervals for a fixed model. BBH logic/tracking are three-object logical deduction/object tracking.

Training	Method	Evaluation	Checkpoint	Seeds	Success (%)
GSM8K	Base/Bo16	GSM8K	base	1	$44.5\,[38.3, 50.8]^{\dagger}$
GSM8K	Correct source/Bo16	GSM8K	source	5	22.1 ± 4.9
GSM8K	Direct RL	BBH logic	feedback	5	40.1 ± 1.6
GSM8K	Direct RL	GSM-Symbolic	feedback	5	54.5 ± 3.2
GSM8K	Direct RL	BBH tracking	feedback	5	11.4 ± 17.6
GSM8K	Correct source	BBH logic	feedback	5	30.9 ± 21.0
GSM8K	Correct source	GSM-Symbolic	feedback	5	53.0 ± 3.5
GSM8K	Correct source	BBH tracking	feedback	5	13.2 ± 22.3
GSM8K	Correct source	BBH logic	source	5	40.7 ± 1.2
GSM8K	Correct source	GSM-Symbolic	source	5	2.1 ± 0.4
GSM8K	Correct source	BBH tracking	source	5	30.8 ± 3.3
GSM8K	Shuffled source	BBH logic	feedback	5	27.1 ± 21.3
GSM8K	Shuffled source	GSM-Symbolic	feedback	5	43.6 ± 26.5
GSM8K	Shuffled source	BBH tracking	feedback	5	11.4 ± 18.5
GSM8K	Shuffled source	BBH logic	source	5	39.5 ± 2.9
GSM8K	Shuffled source	GSM-Symbolic	source	5	6.3 ± 0.8
GSM8K	Shuffled source	BBH tracking	source	5	32.8 ± 2.6
HotpotQA	Base/Bo16	HotpotQA	base	1	$45.7\,[39.5, 52.0]^{\dagger}$
HotpotQA	Correct source/Bo16	HotpotQA	source	5	50.7 ± 1.3
HotpotQA	Direct RL	BBH logic	feedback	5	41.5 ± 1.2
HotpotQA	Direct RL	BBH tracking	feedback	5	35.5 ± 0.8
HotpotQA	Correct source	BBH logic	feedback	5	41.0 ± 1.6
HotpotQA	Correct source	BBH tracking	feedback	5	36.1 ± 1.1
HotpotQA	Correct source	BBH logic	source	5	39.0 ± 1.8
HotpotQA	Correct source	BBH tracking	source	5	38.2 ± 1.3
HotpotQA	Shuffled source	BBH logic	feedback	5	26.8 ± 10.7
HotpotQA	Shuffled source	BBH tracking	feedback	5	34.0 ± 2.8
HotpotQA	Shuffled source	BBH logic	source	5	5.8 ± 2.3
HotpotQA	Shuffled source	BBH tracking	source	5	35.2 ± 2.0
Shared	Base	GSM-Symbolic	base	1	$5.8\,[3.5, 8.5]$
Shared	Base	BBH logic	base	1	$41.6\,[35.6, 48.0]$
Shared	Base	BBH tracking	base	1	$34.4\,[28.4, 40.4]$

A dagger marks a fixed-base best-of-16 estimate. Bracketed base intervals bootstrap prompts (or GSM-Symbolic templates) conditional on the one fixed model; GSM-Symbolic uses equal template means over ten variants per template.

C.12 Endpoint decomposition and complete controlled-task results

For the sequential decomposition, let N be all evaluated prompts, P syntactically valid outputs with a consistent public path, C those outputs that also have the correct terminal public coordinate, and K those in C with the correct private bit. We report P/N, C/N and K/C; the last quantity is undefined when $C = 0$. Confidence intervals use per-seed ratios, and exact counts are shown. Cross-scoring yields disjoint task-verifier counts $N_A + N_B = C$. Memory validity is the fraction producing a legal bit and terminator; its conditional correctness uses only valid completions. All gains are computed from the two feedback evaluations on identical prompts. Greedy fixed-base scores repeat across training seeds because they evaluate one inherited checkpoint.

Public execution and private-task accuracy. Sequential evaluation uses held-out suffixes of lengths four and six. Each source route receives 2^{20} prediction tokens, then 8,192 reward queries: 128 updates of 64 sampled completions. Eight paired worlds compare correct source, coherently wrong source and direct RL. The wrong source is a consistent source law with the same public transition rule but a private transition table different from the target mechanism. Correct-source models obtain 242–584 positive training rewards per world, whereas direct RL and a format control that randomizes every output state obtain none. Consistent source structure, not exposure to the output format, is what gives rewards successful trajectories to reinforce.

Correct-source greedy success rises from 3.30% to 41.67% and wrong-source success from 3.94% to 38.15% (Table 23). The endpoint decomposition, which splits greedy success into a valid public path, a correct public endpoint and a correct private bit, says what this shared improvement is made of. Correct-source outputs have valid public paths on 96.22% of prompts and correct public endpoints on 83.25%, while private-bit accuracy conditional on correct public execution is 50.12% for the correct source and 49.87% for the wrong source (Table 17 and Figure 20). The gain is acquired public execution. The error partition over all prompts (Table 22) identifies the private bit as the largest correct-source error category, 41.58% of all prompts, against 12.96% for a wrong public endpoint and under 4% for malformed or path errors, so the decomposition locates the remaining error in private-state computation and its task-specific use. A classical transition-table learner given the same source records and training feedback reaches 100% in three tested worlds (Table 16), so the observations carry the private information, and the decomposition measures how much of it a neural route has turned into computation.

Format and reward-bearing trajectories. In the historical Sequential direct-RL runs, all 1,024 greedy outputs per world are malformed. The format-source control instead produces syntactically valid answers on 97.75% of prompts on average, but none has a consistent public trajectory (Table 22). Thus a zero public-path rate can reflect either invalid syntax or invalid execution. Historical Memory direct-RL outputs likewise contain no valid bit-and-terminator answer. Across these eight Sequential direct, three format-source and five Memory direct runs, each full 8,192-query training budget contains zero positive rewards and all 128 reward-loss values are zero. The zero endpoints therefore accompany an absence of reward-weighted learning signal. Tables reporting decomposition and cross-task counts reuse these outputs; they are not additional independent runs.

Table 14: Greedy evaluation on identical before/after prompts. Start and end are success percentages; gain is percentage points. Entries are means with 95% Student-t half-widths across paired seeds. Start is the feedback-entry evaluation on the same prompts; direct public-task starts evaluate one fixed base checkpoint.

Training route	Seeds	Before (%)	After (%)	Gain (pp)
Sequential				
Correct source	8	3.30 ±3.27	41.67 ±5.56	38.38 ±5.21
Direct RL	8	0.00 ±0.00	0.00 ±0.00	0.00 ±0.00
Wrong source	8	3.94 ±3.48	38.15 ±10.68	34.20 ±9.87
Format source	3	0.00 ±0.00	0.00 ±0.00	0.00 ±0.00
Memory				
Correct source	5	51.02 ±2.53	49.18 ±1.43	−1.84 ±3.15
Direct RL	5	0.00 ±0.00	0.00 ±0.00	0.00 ±0.00
Wrong source	5	49.90 ±1.46	49.18 ±1.43	−0.72 ±1.35
GSM8K				
Direct RL	5	8.57 ±0.00	62.32 ±4.14	53.75 ±4.14
Domain source	5	2.32 ±0.39	62.05 ±2.93	59.73 ±3.16
Shuffled source	5	8.58 ±1.31	49.70 ±29.29	41.12 ±28.34
HotpotQA				
Direct RL	5	23.93 ±0.00	47.60 ±1.02	23.67 ±1.02
Domain source	5	35.10 ±0.98	45.59 ±0.86	10.49 ±1.74
Shuffled source	5	5.62 ±0.96	44.45 ±1.13	38.83 ±1.91

Table 15: Sequential source dose on five matched worlds. Greedy success after 8,192 reward queries; values are percentages with 95% Student-t half-widths. Source checkpoints share their training trajectories.

Source prediction tokens	Worlds	Final success (%)
2^{18}	5	0.00 ± 0.00
2^{19}	5	0.00 ± 0.00
2^{20}	5	41.99 ± 8.92

Table 16: Greedy sequential endpoints for initialization, scale and alternative-route controls after 8,192 reward queries. Entries are seed means with 95% Student-t half-widths; the classical learner uses the same source records and permitted training feedback.

Setting	Route	Worlds	Success (%)
0.5B, pretrained, 2^{20} source tokens	Balanced source	3	39.71 ± 32.80
	Interleaved source updates	3	38.57 ± 24.05
	Classical table learner	3	100.00 ± 0.00
0.5B, random initialization, 2^{20} source tokens	Correct source	3	0.00 ± 0.00
	Wrong source	3	0.00 ± 0.00
	Direct RL	3	0.00 ± 0.00
1.5B, pretrained, 2^{20} source tokens	Correct source	3	26.43 ± 35.44
	Wrong source	3	22.27 ± 30.76
	Direct RL	3	0.00 ± 0.00

Table 17: Sequential greedy endpoints. Public path requires valid syntax and a consistent public trajectory. Correct public additionally requires the target terminal public coordinate. Private accuracy conditions on this latter event; – denotes a zero denominator in every seed. The first two rates use all prompts; the last is the within-seed conditional ratio. Entries are seed means and 95% Student-t half-widths.

Route	Seeds	Public path (%)	Correct public (%)	Private | public (%)
Correct source	8	96.22 ± 8.56	83.25 ± 11.38	50.12 ± 1.81
Wrong source	8	98.13 ± 2.87	76.77 ± 21.79	49.87 ± 1.09
Direct RL	8	0.00 ± 0.00	0.00 ± 0.00	–
Format source	3	0.00 ± 0.00	0.00 ± 0.00	–

Sampled decoding. Under sampled decoding of the same checkpoints (one sampled answer per prompt), source-trained sequential models also outperform direct RL, by 5.48 ± 1.68 points at the endpoint, while the correct-source reward-stage gain is 0.31 ± 0.55 percentage points against 38.38 ± 5.21 under greedy decoding. The output decompositions in this subsection use greedy decoding (Figure 22 and Table 24).

C.13 Alternative routes, selection and external transfer

Source dose and access to reward. The budget comparison evaluates the 2^{18}, 2^{19} and 2^{20} checkpoints of five source trajectories. At 2^{18} and 2^{19} source tokens, no feedback run receives a positive reward; at 2^{20} every run does, and greedy success reaches 41.99% (Table 15 and Figure 24). Access to successful trajectories therefore switches on within a factor of two in source budget.

Initialization, model scale and adaptation. In the three-world comparison at 2^{20} source tokens, no random-initialized route attains sequential success. The contrast with pretrained initialization shows that inherited weights matter under this budget and protocol. At 1.5B parameters, correct-source and wrong-source routes reach 26.43% and 22.27% while direct RL stays at zero; source sampling

Table 18: Exact denominators for the sequential greedy decomposition. N is all prompts, P valid public paths, C correct public endpoints, and K correct private bits among C. The conditional accuracy is K/C separately in each seed; $C = 0$ is missing.

Route	Seed	N	P	C	K
Correct source	0	1024	1012	695	345
Correct source	1	1024	1024	1021	515
Correct source	2	1024	1024	824	381
Correct source	3	1024	1024	773	412
Correct source	4	1024	1024	1017	497
Correct source	5	1024	1024	771	405
Correct source	6	1024	726	715	358
Correct source	7	1024	1024	1004	501
Wrong source	0	1024	957	772	388
Wrong source	1	1024	1024	1024	524
Wrong source	2	1024	1024	610	299
Wrong source	3	1024	1022	274	141
Wrong source	4	1024	1024	1021	502
Wrong source	5	1024	1024	646	329
Wrong source	6	1024	940	918	438
Wrong source	7	1024	1024	1024	504
Direct RL	0	1024	0	0	0
Direct RL	1	1024	0	0	0
Direct RL	2	1024	0	0	0
Direct RL	3	1024	0	0	0
Direct RL	4	1024	0	0	0
Direct RL	5	1024	0	0	0
Direct RL	6	1024	0	0	0
Direct RL	7	1024	0	0	0
Format source	3	1024	0	0	0
Format source	4	1024	0	0	0
Format source	5	1024	0	0	0

Table 19: Greedy cross-scoring of the same saved outputs, three paired worlds. A/B success sets are disjoint. The conditional A score uses the union of the two sets as denominator; it is undefined for direct runs with no eligible outputs. Entries are seed means with 95% Student-t half-widths.

Route	Reward task	Test A (%)	Test B (%)	A | A or B (%)
Correct source	A	42.77 ± 12.42	40.59 ± 21.93	51.57 ± 5.88
Correct source	B	41.83 ± 14.49	39.45 ± 22.79	51.73 ± 6.34
Direct RL	A	0.00 ± 0.00	0.00 ± 0.00	–
Direct RL	B	0.00 ± 0.00	0.00 ± 0.00	–

Table 20: Exact cross-task greedy counts. $C = N_A + N_B$ is the common eligible denominator: valid public execution with the correct public endpoint. Each world uses 1,024 prompts shared across its A/B branches and scorers; task B changes only the required private bit.

Route	Seed	N	C	N_A	N_B
Correct source	3	1024	773	412	361
Correct source	4	1024	1017	497	520
Correct source	5	1024	771	405	366
Source / task B	3	1024	824	439	385
Source / task B	4	1024	988	482	506
Source / task B	5	1024	685	364	321
Direct RL	3	1024	0	0	0
Direct RL	4	1024	0	0	0
Direct RL	5	1024	0	0	0
Direct / task B	3	1024	0	0	0
Direct / task B	4	1024	0	0	0
Direct / task B	5	1024	0	0	0

Table 21: Contextual greedy endpoints. Correctness conditional on a valid bit is undefined if no valid answer is produced. Constant counts runs producing a single valid action on all 1,024 prompts. A balanced input-independent bit rule has 50% population accuracy. Entries are seed means with 95% Student-t half-widths.

Route	Seeds	Valid (%)	Correct (%)	Correct | valid (%)	Constant runs
Correct source	5	100.00 ± 0.00	49.18 ± 1.43	49.18 ± 1.43	5
Wrong source	5	100.00 ± 0.00	49.18 ± 1.43	49.18 ± 1.43	5
Direct RL	5	0.00 ± 0.00	0.00 ± 0.00	–	0

Table 22: Sequential greedy endpoint error partition (percent of all prompts). Malformed, invalid public path, wrong public endpoint, wrong private bit and success are mutually exclusive and exhaustive. Private-bit failures here use all prompts as denominator; conditional accuracy uses only correct-public outputs. Entries are seed means with 95% Student-t half-widths.

Route	Malformed	Path error	Public error	Private error	Success
Correct source	3.72 ± 8.58	0.06 ± 0.14	12.96 ± 11.00	41.58 ± 6.16	41.67 ± 5.56
Wrong source	1.81 ± 2.83	0.06 ± 0.10	21.36 ± 22.37	38.62 ± 11.18	38.15 ± 10.68
Direct RL	100.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
Format source	2.25 ± 5.25	97.75 ± 5.25	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00

Table 23: Paired differences of reward-stage gains: correct source minus the indicated route. Only shared seeds enter each contrast; entries are seed-paired mean differences in percentage points with 95% Student-t half-widths.

Task	Evaluation	Comparator	Seeds	Difference (pp)
sequential	greedy	Direct RL	8	38.38 ± 5.21
sequential	greedy	Wrong source	8	4.17 ± 8.95
sequential	greedy	Format source	3	41.96 ± 10.16
contextual	greedy	Direct RL	5	-1.84 ± 3.15
contextual	greedy	Wrong source	5	-1.11 ± 2.06
GSM8K	greedy	Direct RL	5	5.97 ± 3.66
GSM8K	greedy	Shuffled source	5	18.61 ± 30.16
HotpotQA	greedy	Direct RL	5	-13.18 ± 1.50
HotpotQA	greedy	Shuffled source	5	-28.34 ± 2.42

Table 24: Sampled-decoding robustness check on the same checkpoints and prompts as the greedy results (one sampled answer per prompt at temperature one). Start and end are success percentages; gain is percentage points; entries are seed means with 95% Student-t half-widths. The two sequential endpoint contrasts use exact two-sided sign-flip tests; adjusted p-values use a fixed three-hypothesis Holm family, of which the two measured endpoint contrasts are shown and the third hypothesis is entered conservatively at $p = 1$.

Training route	Seeds	Before (%)	After (%)	Gain (pp)
Sequential				
Correct source	8	5.18 ±1.71	5.48 ±1.68	0.31 ±0.55
Direct RL	8	0.00 ±0.00	0.00 ±0.00	0.00 ±0.00
Wrong source	8	5.16 ±1.59	5.73 ±1.72	0.56 ±0.77
Format source	3	0.00 ±0.00	0.00 ±0.00	0.00 ±0.00
Memory				
Correct source	5	48.26 ±1.16	49.18 ±1.43	0.92 ±1.41
Direct RL	5	0.00 ±0.00	0.00 ±0.00	0.00 ±0.00
Wrong source	5	48.16 ±2.66	49.18 ±1.43	1.02 ±2.39
GSM8K				
Direct RL	5	5.13 ±0.22	45.75 ±3.96	40.62 ±4.04
Domain source	5	1.62 ±0.25	45.17 ±5.19	43.55 ±4.99
Shuffled source	5	2.68 ±0.91	40.89 ±13.87	38.21 ±13.87
HotpotQA				
Direct RL	5	13.59 ±0.90	45.27 ±1.02	31.68 ±0.83
Domain source	5	19.55 ±1.97	43.71 ±1.48	24.16 ±1.82
Shuffled source	5	2.99 ±0.71	42.23 ±1.05	39.24 ±1.13
Sequential endpoint contrasts (difference in pp, p, p_{Holm})				
Correct source − Direct RL		5.48 ±1.68;	$p = 0.0078$;	$p_{\text{Holm}} = 0.0234$
Correct source − Wrong source		−0.24 ±0.51;	$p = 0.336$;	$p_{\text{Holm}} = 0.672$

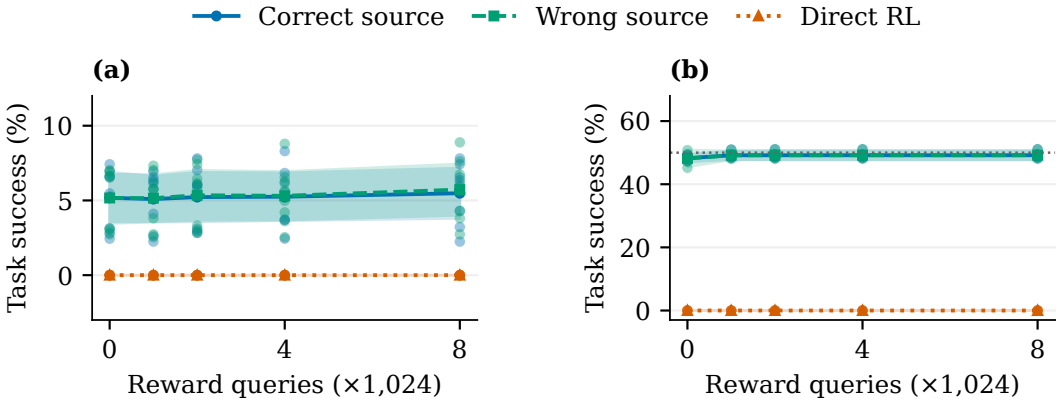


Figure 22: Sampled evaluation of (a) Sequential, eight worlds, and (b) Memory, five worlds. Each checkpoint uses one answer per prompt at temperature one on 1,024 prompts per world. Curves and large markers show world means, small translucent dots individual worlds, and shaded bands 95% Student-t intervals. The dotted 50% line in (b) marks binary chance. Figure 20 and Tables 14 and 17 report greedy evaluation of these historical checkpoints.

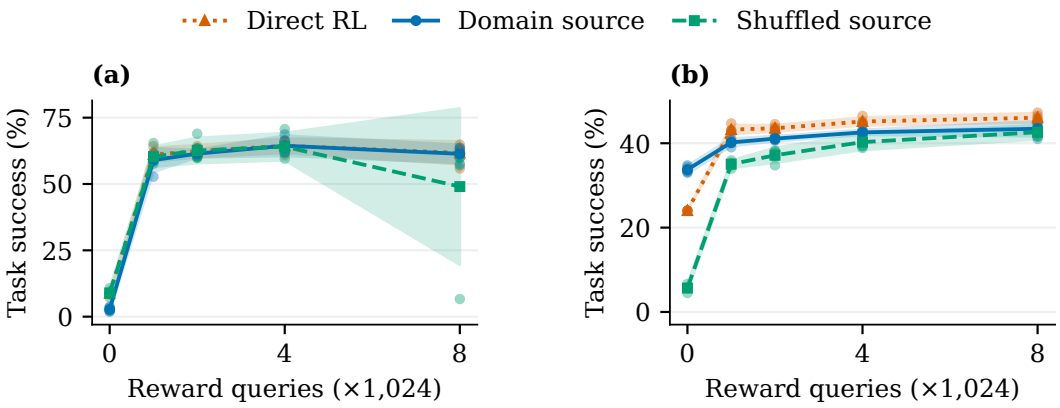


Figure 23: Greedy learning curves for (a) GSM8K and (b) HotpotQA on the same 512 prompts at every checkpoint, five training seeds. Curves and large markers are seed means, small translucent dots are individual seeds, and bands are 95% Student-t intervals. The full endpoint comparison in Figure 21 and Table 2 uses the larger matched evaluation sets.

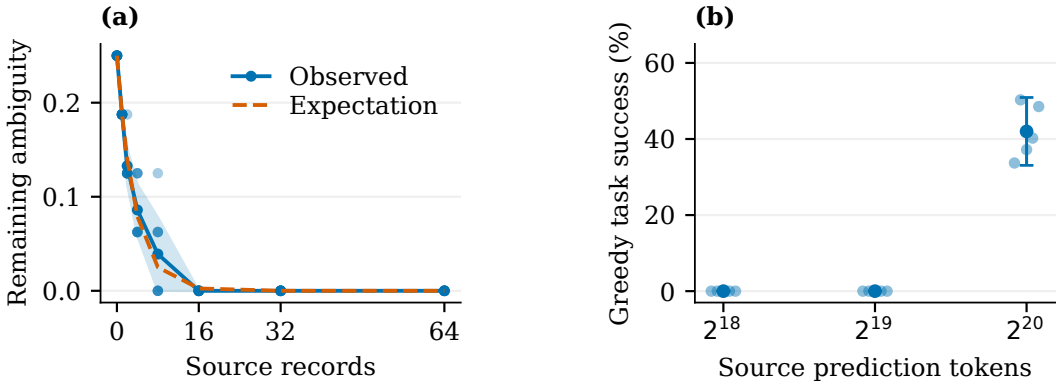


Figure 24: Two effects of source observations on distinct horizontal scales. (a) Remaining contextual ambiguity versus the number N of source records in eight worlds; the dashed curve is the analytic expectation $\frac{1}{4}(\frac{3}{4})^N$. (b) Greedy Sequential success after 8,192 reward queries versus source prediction tokens, five matched worlds. Solid curves and large markers are world means; small translucent dots are individual worlds. Shading in (a) and error bars in (b) are 95% Student-t intervals.

balanced over initial public contexts, at the same token budget, reaches $39.71 \pm 32.80\%$ over three worlds, with substantial between-world uncertainty (Table 16). Continued prediction and answer-only supervised fine-tuning (SFT) do not reproduce the reward-stage improvement, whereas interleaving source updates with RL retains it at 38.57% (Tables 16 and 25). Fixed-source best-of-N selection raises success from 4.95% at $N = 1$ to 26.73% at $N = 8$ by spending test-time verifier queries instead of training queries (Table 26). Source-only, doubled-source and answer-only SFT scores on the same three worlds are 3.55%, 1.40% and zero (Table 25); their separate source and answer-label costs are listed with the endpoints.

Public SFT yields 12.36% on GSM8K and 53.74% on HotpotQA over three seeds (Table 28). Transfer to GSM-Symbolic gives 53.0% after source+RL versus 54.5% after direct RL; BBH logic and object-tracking results vary by task and route (Table 13). Controlled sequential checkpoints change little on the same external tasks (Table 27). On separate 256-question sets, best-of-16 selection scores 22.1% after domain source versus 44.5% for the fixed base on GSM8K, and 50.7% versus 45.7% on HotpotQA (Table 13).

Table 25: Greedy sequential source-only and alternative-adaptation endpoints.

Route	Seeds	Source tokens	Answer labels	Success (%)
Source only, all worlds	8	1,048,576	0	3.30 ± 3.27
Source only, paired worlds	3	1,048,576	0	3.55 ± 3.76
Continue source prediction	3	2,097,152	0	1.40 ± 3.40
Answer-only adaptation	3	1,048,576	8,192	0.00 ± 0.00

All evaluations use 1,024 identical prompts within a world. The paired-world rows use three matched worlds. All routes have zero reward-training queries. Answer-only adaptation additionally uses its recorded answer-prediction targets, separately from source observations. Intervals are 95% Student-t intervals across seed means.

Table 26: Verification-selected sampling from the fixed sequential source checkpoint. Uncertainty is the 95% Student-t half-width across training seeds.

Candidates	Seeds	Prompts	Test queries/seed	Success (%)
1	3	1,024	1,024	4.95 ± 6.29
2	3	1,024	2,048	8.72 ± 9.86
4	3	1,024	4,096	15.92 ± 16.01
8	3	1,024	8,192	26.73 ± 24.57

Each row selects a correct answer, when present, among the first N sampled candidates. All rows use nested prefixes of the same eight candidates per prompt, with no reward-stage parameter updates. Intervals summarize three seed-level means.

Table 27: Greedy external transfer after controlled sequential training of Qwen2.5-0.5B.

Checkpoint	Independent units	Prompts	Success (%)
Logical deduction (250 prompts)			
Base checkpoint	1 fixed model	250	30.40 [24.80, 36.00]
After source prediction	3 training seeds	250	34.40 ±2.63
After source + reward	3 training seeds	250	34.27 ±3.19
Object tracking (250 prompts)			
Base checkpoint	1 fixed model	250	21.20 [16.40, 26.40]
After source prediction	3 training seeds	250	28.13 ±14.10
After source + reward	3 training seeds	250	27.60 ±12.69

Source and feedback uncertainty is a 95% Student-t interval over three training seeds. Base brackets give the 95% item-bootstrap interval for the fixed pretrained model.

Table 28: Greedy answer-only supervised adaptation on public tasks.

Task	Seeds	Prompts/seed	Answer labels/seed	Success (%)
GSM8K	3	1,319	8,192	12.36 ±0.38
HotpotQA	3	1,024	8,192	53.74 ±1.38

Each Qwen2.5-1.5B run starts from its domain-source checkpoint and receives 8,192 terminal-answer labels, with no reward-training queries. Source training uses 2^{22} prediction tokens. Success is exact answer correctness; intervals are 95% Student-t intervals over three training seeds, not prompt intervals.

C.14 Source-acquisition probes and public source dose

This subsection gives the two-world format-stage study discussed in Section 6 and the 2^{20}-token public source-dose comparison.

Public source-dose comparison. The 2^{20}-token source route uses Qwen2.5-1.5B, the same source learning rates as the 2^{22} route and 8,192 training verifications per feedback run. Within each seed, the direct, 2^{20} and 2^{22} routes share identical evaluation prompts and the same ordered sequence of 8,192 training prompts. Tables 29 and 30 report start, endpoint, gain and seed-paired contrasts.

Table 29: Public source-dose endpoints. Five paired training seeds; the same 1,319 GSM8K or 1,024 HotpotQA prompts at both endpoints and across routes. Rates are percentages; gains are percentage points. Uncertainty is a seed-level 95% Student-t half-width. Direct starting scores evaluate one fixed pretrained checkpoint.

Source tokens	Before RL (%)	After RL (%)	Gain (pp)
GSM8K; 8,192 reward queries			
None (direct)	8.57 ±0.00	62.32 ±4.14	53.75 ±4.14
2^{20}	3.11 ±0.57	61.76 ±3.15	58.65 ±3.67
2^{22}	2.32 ±0.39	62.05 ±2.93	59.73 ±3.16
HotpotQA; 8,192 reward queries			
None (direct)	23.93 ±0.00	47.60 ±1.02	23.67 ±1.02
2^{20}	35.55 ±1.44	46.23 ±1.28	10.68 ±2.61
2^{22}	35.10 ±0.98	45.59 ±0.86	10.49 ±1.74

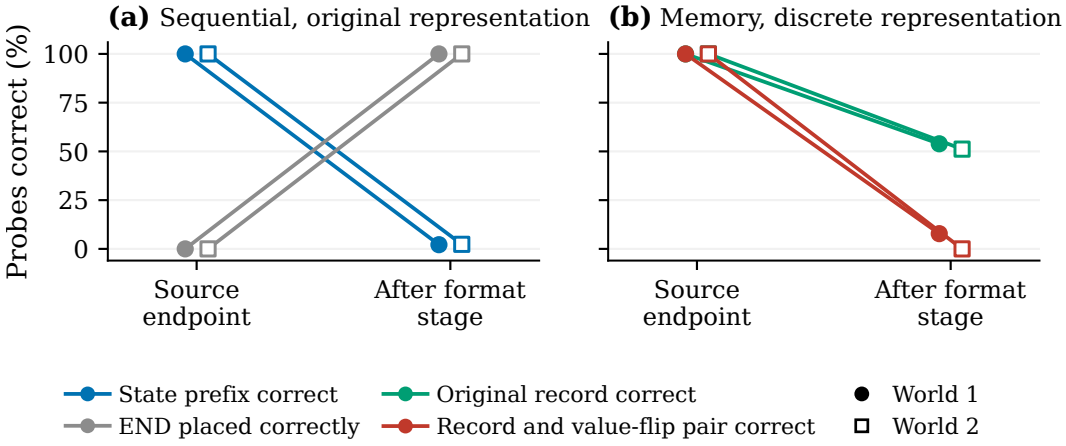


Figure 25: Acquired computation before and after the answer-format stage, two worlds at 2^{24} source tokens. (a) Sequential, original representation with state targets: correct first-H state prefixes and correct END placement, out of 512 probes with the first operation supplied. (b) Memory, discrete representation with the all-token objective: correct first-bit answers on original records and on both members of record/value-flip pairs, each out of 256. Circles and squares are the two worlds, slightly offset where they coincide. Colors identify metrics; lines join the same world and metric across stages.

Table 30: Seed-paired public source-dose contrasts, in percentage points (pp). Each effect is calculated within seed before forming the five-seed mean and 95% Student-t half-width.

Contrast	Endpoint difference (pp)	Gain difference (pp)
GSM8K		
2^{20} minus direct	-0.56 ± 4.87	4.90 ± 5.28
2^{22} minus direct	-0.27 ± 3.63	5.97 ± 3.66
2^{20} minus 2^{22}	-0.29 ± 5.64	-1.08 ± 6.38
HotpotQA		
2^{20} minus direct	-1.37 ± 0.37	-12.99 ± 1.67
2^{22} minus direct	-2.01 ± 0.88	-13.18 ± 1.50
2^{20} minus 2^{22}	0.64 ± 1.12	0.20 ± 1.82

Acquisition design. Two source worlds per representation compare learning rates 10^{-5} and 3×10^{-5} at 2^{22} prediction tokens, and the better rate per representation is trained to 2^{24} tokens with matched objective controls. Sequential compares state targets with private-bit-weighted state targets (private weight four, other state targets weight one). Memory compares all-token prediction with future-bit prediction, the latter excluding END and end-of-sequence (EOS) from its source loss. Each source update observes 8,192 prediction-token positions. Alongside the original representations, an alternative Sequential representation adds an explicit length boundary and a local-to-complete-trajectory curriculum, and an alternative Memory representation uses discrete keys and values; alternative tasks are reported separately from the original continuous-numeric Memory task.

An answer-format stage precedes and follows source training. Its loss supervises complete public-format answers, including termination, on records whose private fields are independent random labels, so it reveals neither the target mechanism nor the task binding. Sequential uses 128 format updates per stage and Memory 32, with batch 64 and learning rate 10^{-5}; stage transitions reset Adam moments.

Separating state content from termination. Each Sequential endpoint has 512 source-validation prompts, split equally between horizons $H = 5$ and $H = 7$, with the random first operation supplied and no task binding. The complete-answer metric requires exactly the expected states followed by

END; a content diagnostic instead checks the first H generated states without modifying the output. Seven of the eight original-representation 2^{22} endpoints already generate at least 511/512 correct state prefixes. At 2^{24}, all four original source models and three of four curriculum models generate every state prefix correctly, continuing to the 128-token limit rather than emitting END. After the format stage, every 2^{24} Sequential model places END correctly on all 512 prompts, while state-prefix counts fall to 6–14/512 (Table 31). The two metrics separate termination, learned in the format stage, from state-prefix accuracy, which falls sharply after this stage.

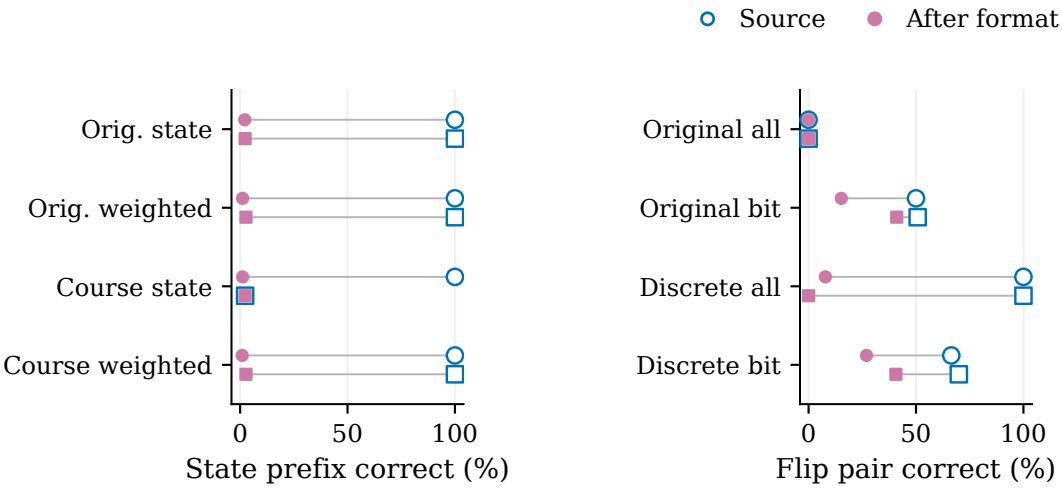


Figure 26: Two-world acquisition probes before and after the format stage at 2^{24} source tokens. Left: exact Sequential state prefixes, excluding termination, on 512 probes. Right: both members correct in each of 256 Memory original/relevant-flip pairs. Open blue markers are source endpoints; filled purple markers follow the format stage. Circles and squares denote the two worlds, and links pair the same world and representation. Legend abbreviations: "Orig." is the original representation and "Course" the Sequential curriculum representation; "state" uses state-token targets and "weighted" upweights private state targets; "all" and "bit" are the Memory all-token and future-bit objectives. The Sequential diagnostic uses source-validation support with the first operation given. Figure 25 shows the original-representation Sequential state-target and discrete Memory all-token conditions, adding termination and first-bit accuracy to the two diagnostics shown here.

Memory interventions and the format interface. Each Memory world has 256 original source records and paired relevant-value flips; a successful pair requires the correct first bit on both inputs. At 2^{24} tokens, original-representation all-token source accuracy averages 49.61% with no successful pairs, and future-bit accuracy averages 75.20% with 50.39% successful pairs. Discrete all-token source models answer all originals and relevant pairs correctly and already satisfy the public format checks. After the format stage, these means fall to 52.54% and 3.91%; discrete future-bit means change from 85.35%/68.16% to 64.45%/33.79%. For a binary private field predicted with probability p, independent uniform labels give expected cross-entropy $-\frac{1}{2}\log p - \frac{1}{2}\log(1-p)$, minimized at $p = 1/2$: supervising random private content pulls the model toward an input-independent private prediction, which is consistent with the observed loss of intervention sensitivity after the format stage and motivates placing format preparation before source acquisition and excluding such random-label supervision afterwards.

Table 31: Acquisition-probe counts at 2^{24} source tokens. Sequential columns count correct first-H state prefixes out of 512; Memory columns count correct original bits and relevant-flip pairs out of 256. Each cell shows before/after format. Before the format stage no Sequential model emits END; after it, every model places END correctly on all 512 prompts.

Representation	Objective	World	Content before	Content after	Pair before/after
Sequential; content /512					
Original	State	1	512	11	–
Original	State	2	512	12	–

Table 31: Acquisition-probe counts (continued).

Representation	Objective	World	Content before	Content after	Pair before/after
Original	Weighted	1	512	6	–
Original	Weighted	2	512	14	–
Curriculum	State	1	512	6	–
Curriculum	State	2	12	12	–
Curriculum	Weighted	1	512	5	–
Curriculum	Weighted	2	512	14	–
Memory; content and pairs /256					
Original	All-token	1	130	130	0/0
Original	All-token	2	124	132	0/0
Original	Future-bit	1	192	145	128/39
Original	Future-bit	2	193	182	130/105
Discrete	All-token	1	256	138	256/20
Discrete	All-token	2	256	131	256/0
Discrete	Future-bit	1	220	152	170/69
Discrete	Future-bit	2	217	178	179/104

D READING MAP AND THEOREM INTERFACES

The information results first specify what observations can distinguish. The optimization results then construct finite training paths under their stated architectures and schedules. Complete statements appear next to their proofs. Shared network facts are proved once; each optimizer checks its own parameter choices and failure allocations.

Question	Complete statement and proof
Sequential reward equivalence	Theorem E.1 and Section E.2.2
Contextual source information	Theorem E.5 and Section E.5
Shared initialization and network bounds	Section G
Sequential finite Adam path	Theorem H.1 and Section H.5
Contextual finite Adam path	Theorem I.1 and Section I.7
SGD alternative	Theorem J.1 and Section J
Information and computation costs	Section F
Experimental protocols and results	Section C

Here $q = 2m$ counts sequential states and controls, H is the operation horizon, and $N = 2H + 2$ counts tokens in a source record. Within shared network bounds, N is the padded context length, R the attention-head count, $k_{\mathrm{head}} = d/R$ its width, and D_{feat} the feature dimension. Optimizer sections explicitly instantiate these quantities; their local rate and radius symbols do not change the task binding. Contextual dimensions are n_{ctx} and $k = p - 2$. A source record, prediction target, verifier query and optimizer update are distinct resources.

E INFORMATION SUPPLIED BY SOURCE AND REWARD OBSERVATIONS

In either family a world fixes a reusable mechanism and a downstream rule. Initial information O includes any source sample and inherited weights; reward adaptation produces a transcript of prompts, sampled outputs and terminal feedback. The decision is an output on a held-out input. In Sequential the mechanism is θ, the rule is η, and source reveals noisy transitions; in Contextual they are U and (s, τ), and source reveals selected-value predictions. Reward equivalence concerns the complete permitted feedback interface, while learnability by the specified network additionally requires the finite optimization arguments below. World-independent initialization is part of each no-source lower-bound experiment; arbitrary pretrained weights must instead be counted in O.

E.1 MECHANISMS WITH IDENTICAL TRAINING REWARDS

Use the sequential state space, operation tags, training and held-out supports, and terminal verifier of Section 3. A world fixes the transition table θ and initial operation $\eta = (b_\eta, u_\eta)$. All sums below are in $\mathbb{F}_2$.

Theorem E.1 (Mechanisms with identical training rewards)**.** *Let $H \geq 3$ and fix u_η. Two worlds are reward-equivalent if and only if, for some $\kappa, d \in \mathbb{F}_2$,*

$$\theta'_{u,z} = \theta_{u,z} + \kappa + d\,\ell(u), \qquad b'_\eta = b_\eta + (H \bmod 2)\,\kappa + d\,\ell(u_\eta). \tag{E.1}$$

Their targets differ on every test prompt exactly when $d = 1$.

The proof is given in Section E.2.2.

E.2 Exact identification from source and terminal observations

This appendix proves the statistical results. Throughout, $L = H - 1$ is the suffix length, $m = |Z|$, and all additions of bits or coordinates are in $\mathbb{F}_2$. Matrix norms and inner products are over the reals. The matrix arguments are analysis of a binary-equation learner; they do not grant it an additional oracle.

E.2.1 Optimal query order with a small source budget

The iid-prompt restriction makes recovery quantitative. A symbolic learner first tests public binding coordinates with random endpoint bits. Once that coordinate is identified, one terminal reply reveals one exact binary equation in the unknown transition table and binary binding coordinate. After solving these equations, the learner uses its stored source records to recover κ, d.

Canonical execution starts with η and applies the true transition at every step; δ below is the allowed failure probability.

Theorem E.2 (A matched low-source query regime)**.** *Let $L = H - 1 \geq \lceil 2^{16} \ln(2m) \rceil$ and $0 < \delta < 1$. The explicit fresh-iid learner with budgets (E.18) and (E.20) has, for each fixed world, probability at least $1 - \delta$ of correct canonical execution on every test prompt. Its endpoint success, target first-state retention and task utility then equal one. It uses*

$$N_{\mathrm{pre}} \leq 32a^{-2} \ln(6/\delta) + 2H + 2, \tag{E.2}$$

$$Q = O\left(m^2 + \left(m + \frac{m^2}{L} \right) \ln(m/\delta) + m \ln(1/\delta) \right), \tag{E.3}$$

with nQ generated decisions, HQ prompt tokens and zero gradient updates. At fixed $\delta = 1/16$, fixed $c \in (0, 1/64]$, $a = 1 - c/H$ and $L = \lceil 2^{16} \ln(2m) \rceil$, this is $O(\log m)$ raw source tokens and $O(m^2)$ queries. For $m \geq 512$, any learner with n_{rec} source records and worst-case expected endpoint risk at most $1/16$ must satisfy

$$Q \geq \frac{5}{896} m^2 - n_{\mathrm{rec}} H - 1, \tag{E.4}$$

even when granted the binding and adaptive chosen training prompts. Hence the query coordinate is minimax-optimal in this fixed-confidence, logarithmic-source-budget and logarithmic-horizon regime.

The low-error structure is the key to the upper bound. For the binary entrywise difference f between two transition tables, form a signed transition matrix $A(z, z') = (-1)^{f(z'+z,z)}/m$, indexed by public states z, z' and its tagged counterpart. Their powers describe the probability of agreement along a random suffix. Writing ε for training-endpoint disagreement probability, near-perfect agreement forces one matrix close to a signed rank-one form. After sign rounding, only $O(\varepsilon m^2/L)$ transition defects and $O(\varepsilon m)$ vertex-sign choices remain at error ε. The resulting small hypothesis count makes $O(m^2)$ iid equations sufficient when $L = \Theta(\log m)$. A separate short-witness argument rules out arbitrarily small nonzero error before applying this rounding estimate.

The binding prefix has its own sampling law. Retaining its factor throughout the suffix approximation controls accuracy on this distinguished row. Finally, bounded source sign statistics recover the two residual bits. The source is independent of the feedback calculation because it is stored but not used during that calculation. The lower bound packs $2^{\Theta(m^2)}$ separated mechanisms and counts at most $n_{\mathrm{rec}} H + Q$ information bits. The proofs and complete computation/storage charges are in Section E.2.

The logarithmic source-token charge includes the length of a mandatory whole record. At fixed confidence and fixed noise level, decoding the two ambiguity parameters needs only a constant-order number of informative transitions. Thus record length and the information needed for the two residual bits make distinct contributions to the token count.

E.2.2 THE COMPLETE OBSERVABLE QUOTIENT

Fix the public binding coordinate u_η. Write $f(u,z) = \theta'_{u,z} + \theta_{u,z}$ and $\beta = b'_\eta + b_\eta$. At suffix-start coordinate z, equality of all training endpoints is equivalent to

$$\beta + f(u_\eta, z + u_\eta) + \sum_{j=1}^{L} f(v_j, z_{j-1}) = 0, \qquad z_j = z_{j-1} + v_j, \quad \sum_j \ell(v_j) = 0. \tag{E.5}$$

Consequently every admissible length-L path starting at z has the same cost. Its all-zero path has cost $Lf(0,z)$. For $\ell(u) = 0$, compare $(u, 0, \dots, 0)$ with $(0, u, 0, \dots, 0)$. This gives $f(0,z) = f(0, z+u)$. The set $\ell^{-1}(0)$ spans Z: it contains the coordinate vectors in the x and y components and $t + x_1 + y_1$, whose tag is zero. Thus $f(0,z) = \kappa$ is constant. Comparing a single tag-zero control followed by zeros with the all-zero path gives $f(u,z) = \kappa$ for every tag-zero control.

For tag-one controls u, v, the path $(u, v, 0, \dots, 0)$ has tag zero. Comparing with the all-zero path gives $f(u,z) = f(v, z+u)$. At any fixed vertex all tag-one outgoing costs therefore agree; call their value $a(z)$. The same equality gives $a(z) = a(z+u)$ for each tag-one u. The tag-one controls contain $t, t + x_i, t + y_i$ and span Z, so $a(z) = \lambda$ is constant. Hence

$$f(u,z) = \kappa + d\ell(u), \qquad d = \kappa + \lambda. \tag{E.6}$$

Substitution into (E.5) forces $\beta = H\kappa + d\ell(u_\eta)$. Conversely these differences cancel on every tag-zero suffix and produce the difference d on every tag-one suffix. This argument requires only $L \geq 2$.

Endpoint equality is equivalent to equality of the full binary verification oracle once u_η agrees. Invalid grammar or public-coordinate paths are rejected in both worlds. On any given prompt one can instead choose a valid public-coordinate path and either final state bit. Exactly the correct endpoint is accepted, independently of the intermediate state bits. These two candidates distinguish unequal endpoints. This proves the claimed equivalence for the actual verifier, including repaired trajectories.

E.2.3 FRESH PROMPTS GIVE EXACT BINARY EQUATIONS

First identify u_η. For each of its m possible values, use at most $k = \lceil \log_2(3/\delta) \rceil$ fresh iid training prompts. Output a grammar-compliant candidate with the required public coordinates for that coordinate and a fresh fair final state bit. A wrong coordinate never succeeds; the correct coordinate succeeds with probability $1/2$ independently on each trial. Stop at the first success, or declare failure after the fixed budget. The failure probability is at most $2^{-k} \leq \delta/3$, and the query cost is at most $Q_{\mathrm{id}} = mk$.

Thereafter choose $A = (0, u_\eta)$ and a compliant trajectory with final state bit zero. Its binary reward reveals the true endpoint bit: reward one means bit zero, and reward zero means bit one. Subtract the known initial state bit and the displayed control bits. The remaining label is

$$b_\eta + \theta_{u_\eta, z_0} + \sum_{i=2}^{H} \theta_{u_i, z_{i-1}}, \qquad z_1 = z_0 + u_\eta, \quad z_i = z_{i-1} + u_i. \tag{E.7}$$

Each observation is a known binary row in $m^2 + 1$ unknown bits. The rows are iid because prompts remain fresh. Gaussian elimination returns any consistent representative. It remains to show that a controlled number of these particular path rows eliminates every difference outside the two-bit kernel. Counting the number of unknown bits alone does not establish that assertion.

E.2.4 A COARSE SEPARATION BOUND

Define the signed convolution

$$M_\ell(z,w) = m^{-1}(-1)^{\ell(w+z)}. \tag{E.8}$$

At Walsh frequency (α, β, τ) its eigenvalue is the average of $(-1)^{t + x\cdot y + \alpha\cdot x + \beta\cdot y + \tau t}$. It vanishes when $\tau = 0$. When $\tau = 1$, the sum over y forces $x = \beta$ and leaves $2^{-r}(-1)^{\alpha\cdot\beta}$. Thus $\|M_\ell\| = \rho = 2^{-r} \leq 1/2$.

For any cost f, either two length-nine blocks have the same start, end and tag but opposite costs, or

$$f(u,z) = \kappa + h(z) + h(z+u) + d\ell(u). \tag{E.9}$$

To see this, comparisons $(0,u)$ versus $(u,0)$ detect nonconstant self-loop costs. Without such a witness, subtract their constant κ. Comparisons (u,u) versus $(0,0)$ then detect unequal forward and reverse costs. If those also agree, use the undirected lifted graph on $(z,e) \in Z \times \mathbb{F}_2$, whose u-edge leads to $(z+u, e+\ell(u))$. A direct edge gives any public displacement; the three-edge loop (x_1, y_1, x_1+y_1) has zero displacement and odd tag. The graph therefore has diameter at most four. A breadth-first spanning tree has depth at most four. A non-tree edge violating a potential produces a fundamental closed walk of length at most nine with odd cost. Pad this walk and the zero walk to length nine. If no such violation occurs, the graph has a potential $G(z,e)$. Since costs do not depend on e, the difference $G(z,e+1) + G(z,e)$ is constant along edges, hence is a constant d. Writing $G(z,e) = h(z) + de$ proves (E.9). All earlier short witnesses can likewise be padded because their endpoints agree.

Place either witness block at the beginning of the suffix and fix its public starting vertex. At least one suffix control remains when $L \geq 10$. Conditioning the entire suffix tag to zero leaves the probability of each fixed block and starting vertex exactly m^{-10}. Couple the remaining controls identically. The endpoint-difference parities of the paired words disagree, so at least one is wrong. The error is at least m^{-10}; the two block probabilities already pay for the pairing, with no additional factor $1/2$.

For (E.9), the full endpoint difference is

$$\beta + H\kappa + h(z_0) + h(z_H) + d\ell(u_\eta). \tag{E.10}$$

The conditional final public-coordinate law is $m^{-1} + M_\ell^L(z_0 + u_\eta, z_H)$. Every entry is at least $m^{-1} - 2^{-rL} \geq (2m)^{-1}$ for $L \geq 10$. A nonconstant h therefore gives error at least $(2m)^{-1}$. A constant h gives a kernel difference or error one. In particular every nonkernel hypothesis has training disagreement at least m^{-10}.

E.2.5 Counting the low-error differences

Assume $L \geq \lceil 2^{16} \ln(2m) \rceil$. For $0 < \varepsilon \leq 1/800$, the number of difference hypotheses with error at most ε is bounded by

$$16 \exp\left\{ 160\varepsilon \left(m + \frac{m^2}{L} \right) \ln(m+1) \right\}. \tag{E.11}$$

Moreover every nonkernel difference has error at least

$$\varepsilon_* = \min\left\{ \frac{L}{24m^2}, \frac{1}{52m}, \frac{1}{800} \right\}. \tag{E.12}$$

Here is the quantitative argument. Put $A(z,w) = (-1)^{f(w+z,z)}/m$ and $A_c(z,w) = (-1)^{\ell(w+z)} A(z,w)$. Both have Frobenius norm one. Keep the binding factor $v(z) = (-1)^{\beta + f(u_\eta, z+u_\eta)}$. Then

$$1 - 2\varepsilon = m^{-1} \langle v, (A^L + A_c^L)\mathbf{1} \rangle. \tag{E.13}$$

Select the branch of larger operator norm a_0. It satisfies $a_0^L \geq (1-2\varepsilon)/2 \geq 1/4$, so $\delta_A = 1 - a_0^2 \leq 2\ln 4/L$. Its best rank-one singular approximation has squared residual δ_A. Round its left and right singular vectors to signs h, k. Each incorrectly rounded matrix entry contributes at least $1/m^2$ to this residual. The rounded matrix $P(z,w) = h(z)k(w)/m$ thus differs in at most $\delta_A m^2$ signs and in Frobenius norm by at most $2\sqrt{\delta_A}$. The complementary tagged branch has norm at most $\rho + 2\sqrt{\delta_A} \leq 5/8$. Consequently $a_0^L \geq 1 - 2\varepsilon - (5/8)^L$. For a nonkernel difference the preceding coarse bound gives $\varepsilon \geq m^{-10}$, whereas $(5/8)^L \leq m^{-20}$. Taking logarithms yields

$$\delta_A \leq 12\varepsilon/L, \qquad \#\{f \neq f_0\} \leq 12\varepsilon m^2/L, \tag{E.14}$$

where f_0 is the rounded separable cost, with the chosen tag branch restored. Zero singular-vector coordinates cause no problem: any wrong sign still has the stated residual cost. Kernel differences may be counted separately below the coarse separation.

Retain the *original* factor v and replace only the suffix costs by f_0. Before tag conditioning every suffix edge is uniform on m^2 pairs; conditioning costs at most a factor two. A union bound over L

edges and (E.14) bounds the change of suffix parity by 24ε. The auxiliary predictor therefore has error at most 25ε. This replacement does not require a bound on the distinguished binding row.

Write $g = kh$, with pointwise multiplication. The exact rank-one identity is

$$P^L \mathbf{1} = h\,\overline{kh}^{\,L-1}\bar{k}. \tag{E.15}$$

The complementary signed matrix of P has norm ρ. Thus auxiliary agreement is at most $|\bar{k}|\,|\bar{g}|^{L-1} + \rho^L$. Since $\rho^L \le \varepsilon$, the elementary logarithmic bounds from agreement at least $1 - 50\varepsilon$ imply

$$\operatorname{minority}(k) \le 26\varepsilon m, \qquad \operatorname{minority}(g) \le 102\varepsilon m/L. \tag{E.16}$$

Choose the branch, the two global signs and β in at most 16 ways. The remaining choices are bounded by

$$16\left[\sum_{i\le\lfloor 26\varepsilon m\rfloor}\binom{m}{i}\right]\left[\sum_{j\le\lfloor 102\varepsilon m/L\rfloor}\binom{m}{j}\right]\left[\sum_{d\le\lfloor 12\varepsilon m^2/L\rfloor}\binom{m^2}{d}\right]. \tag{E.17}$$

Using $\sum_{j\le t}\binom{n}{j} \le (n+1)^t$ for integral $t \ge 0$ gives (E.11); the coefficient 160 is conservative. If $\varepsilon < \varepsilon_*$, all three integer radii are zero. Then k, g are constant, there are no defects, and f is a constant plus a tag. Its binding bit gives either the kernel or error one. This proves (E.12).

E.2.6 Finite query and source schedules

Let $D = m + m^2/L$, $B = 2 + \lceil \log_2(52m^2) \rceil$ and $\gamma = \delta/3$. Use

$$Q_{\mathrm{rank}} = \left\lceil \max\left\{ \begin{matrix} 800\big[(m^2+1)\ln 2 + \ln(2/\gamma)\big], \\ 640D\ln(m+1) + \varepsilon_*^{-1}\ln(32B/\gamma) \end{matrix} \right\} \right\rceil. \tag{E.18}$$

The 2^{m^2+1} hypotheses of error at least $1/800$, including equality, survive with total probability at most $\gamma/2$. Partition the remaining interval $[\varepsilon_*, 1/800)$ into at most B dyadic bins with lower endpoints e_j, truncating the final bin. Equation (E.11) bounds each bin's size by $16\exp[320 e_j D \ln(m+1)]$; survival is at most $\exp(-Q_{\mathrm{rank}} e_j)$. Equation (E.18) makes the contribution at most $\gamma/(2B)$ per bin. If $\varepsilon_* = 1/800$, the low-error interval is empty. Thus elimination recovers the correct coset with failure at most γ.

Store raw source records before feedback but do not use them in the coordinate/rank procedure. Conditional on a fixed world and the feedback representative θ^0, raw source transitions remain independent of that procedure. For an observed transition from (s, z) under (b, u) to a state with state bit s', set

$$R = s' + s + b + \theta^0_{u,z}, \qquad X_g = \mathbf{1}\{\ell(u) = g\}(-1)^R, \quad g \in \{0, 1\}. \tag{E.19}$$

On coset recovery, $R = \kappa + d\ell(u)$ in the true-transition component, and it is a fair bit in the reset component, even when the public coordinate is wrong. For a sequentially uniform control, $\mathbb{E}[X_g \mid \text{past}] = (a/2)(-1)^{\kappa+dg}$. Use the first $H-1$ execution transitions in each record. The control U_1 is independent uniform, and $U_2, \ldots, U_{H-1}$ are a strict subset of the tag-conditioned suffix and remain sequentially uniform. Marginalize the unused final control. This is an execution-order proof filtration, not a claim that the raw serialization reveals variables in that order.

For n_{rec} records, bounded martingale differences give a two-sign error probability at most $2\exp[-n_{\mathrm{rec}}(H-1)a^2/8]$. Taking

$$n_{\mathrm{rec}} = \left\lceil \frac{8a^{-2}\ln(6/\delta)}{L} \right\rceil \tag{E.20}$$

makes this at most $\delta/3$. Both gauges, including the endpoint-irrelevant common gauge, are recovered. Correcting θ^0 and the binding bit therefore gives the true first-state target and the full canonical execution on every test prompt. Coordinate, rank and source failure allowances sum to δ.

The raw token cost is

$$N_{\mathrm{pre}} = (2H+2)n_{\mathrm{rec}} \le 32a^{-2}\ln(6/\delta) + 2H + 2. \tag{E.21}$$

The query cost is $Q = Q_{\mathrm{id}} + Q_{\mathrm{rank}}$, generation costs $(H+2)Q$ decisions, and prompts contribute HQ tokens. There are zero gradient updates and no checkpoint or test selection. Dense binary elimination costs conservatively $O(Q(H \log q + m^4) + N_{\mathrm{pre}} \log q)$ bit operations, with $O(m^4 + H \log q + N_{\mathrm{pre}} \log q)$ stored bits when raw source is retained. The verifier separately performs H trusted transitions and H public-coordinate checks per query. Deployment uses H table lookups and $O(H)$ generated symbols; no external shield is inserted.

E.2.7 THE MIXED-OBSERVATION LOWER BOUND

For $m \geq 512$ and $H \geq 5$, restrict to mechanisms with $\theta_{0,v} \neq \theta_{0,0}$ for a fixed $v \neq 0$. There exists a packing of $K_{\mathrm{pack}} = 2^{m^2/128}$ mechanisms with pairwise endpoint disagreement at least $7/16$, even with a fixed known binding. We give the elementary packing calculation to specify the stronger information permissions of the lower bound.

For a random pair of such tables, its difference matrix has independent fair signs except for one tied entry. For the signed matrices A, A_c above, couple each to an iid sign matrix; changing the tied entry costs at most $2/m$ in operator norm. Two $1/4$-nets of size 9^m and scalar Hoeffding bounds yield

$$\Pr\left[\max(\|A\|, \|A_c\|) > \tfrac{1}{2}\right] \leq 4 \exp\left[2m \ln 9 - \frac{(m-4)^2}{32}\right] \leq e^{-m^2/64}. \tag{E.22}$$

A union bound over fewer than K_{pack}^2 pairs proves existence. The fixed initial binding contributes the diagonal sign matrix $D_\eta(z,z) = (-1)^{f(u_\eta, z)}$ and permutation $P_u(z,z') = \mathbf{1}\{z' = z+u\}$, both orthogonal. The signed test agreement is

$$m^{-1} \mathbf{1}^\top D_\eta P_{u_\eta} (A^{H-1} - A_c^{H-1}) \mathbf{1}, \tag{E.23}$$

whose magnitude is at most 2^{2-H}. Thus disagreement is at least $(1 - 2^{2-H})/2 \geq 7/16$.

Let J be a uniformly chosen packing index. Reveal the learner's independent random seed at the start of the transcript and condition on it; subsequent randomized choices are then deterministic functions of that seed and preceding observations. Raw controls and resets convey no mechanism information. Conditional public-coordinate transitions have a mechanism-independent law; each next state bit conveys at most one bit. Hence n_{rec} source records convey at most $n_{\mathrm{rec}} H$ bits. Any adaptive binary verification replies add at most Q bits, even when the learner may choose its training prompts. Conditional on this seed and the preceding transcript, the chosen queries supply no additional mechanism information. Therefore $I(J; \text{transcript}) \leq n_{\mathrm{rec}} H + Q$ bits.

Decode the closest packing endpoint function from the learned predictor. A mistaken index forces endpoint error at least $7/32$ by the triangle inequality. For a randomized predictor use its conditional output probabilities as a randomized endpoint function; malformed outputs only add error. The entropy encoding bound $H(J \mid \text{transcript}) \leq 1 + p_{\mathrm{err}} \log_2 K_{\mathrm{pack}}$ gives

$$\sup_{\theta,\eta} \mathbb{E}[\text{endpoint risk}] \geq \frac{7}{32}\left[1 - \frac{128(n_{\mathrm{rec}} H + Q + 1)}{m^2}\right]_+. \tag{E.24}$$

For worst-case risk at most $1/16$, this implies $Q \geq 5m^2/896 - n_{\mathrm{rec}} H - 1$. At fixed confidence $\delta = 1/16$, fixed $c \in (0, 1/64]$, $a = 1 - c/H$ and $L = \lceil 2^{16} \ln(2m) \rceil$, the upper has $N_{\mathrm{pre}} = O(\log m)$ and $Q = O(m^2)$. Its perfect-recovery event implies the required expected-risk upper. Equation (E.24) gives $Q = \Omega(m^2)$ at that source budget. This matches the query coordinate in this regime, while retaining the whole-record cost in the source coordinate.

E.3 A FINITE-BUDGET BOUNDARY FOR A KNOWN PAIR OF WORLDS

We isolate the gauge ambiguity from the cost of learning an unknown transition table. The following experiment supplies the learner with two candidate worlds and asks which one generated its observations. It does not supply the true member. This restriction permits an exact boundary; it is not an upper bound for the full unknown-table class.

Experiment and initial information. Fix $H \geq 3$ and a sequential world (θ^0, η^0). Let $\theta^1_{u,z} = \theta^0_{u,z} + \ell(u)$, $u_{\eta^1} = u_{\eta^0}$ and $b_{\eta^1} = b_{\eta^0} + \ell(u_{\eta^0})$, with all additions in $\mathbb{F}_2$. Both tables and both

bindings are known candidates. The unknown index is $W \in \{0,1\}$. Let D_0, D_1 be the common prompt laws conditional on even and odd suffix tag parity, respectively. An initial observation O has law P_w in world w and includes all source records, inherited checkpoint weights and other side information jointly. Take its observation space to be standard Borel. After O, at most $Q \in \mathbb{N}_0$ verifier queries receive fresh exogenous iid prompts from $D_\varepsilon = (1-\varepsilon)D_0 + \varepsilon D_1$, $0 \le \varepsilon \le 1$. The prompt sequence is independent of (W, O); every presented prompt costs a query, including a rejection. At a prompt, the learner chooses a candidate trajectory adaptively and receives the deterministic endpoint verifier's reply. There is no uncharged prompt selection. The final test prompt has law D_1 independently of training and W. The loss is endpoint failure, including malformed outputs. Procedures may randomize and have no computational restriction. Write $\mu = P_0 + P_1$, $p_w = dP_w/d\mu$, and $\alpha = \int \min(p_0, p_1)\, d\mu = 1 - \mathrm{TV}(P_0, P_1)$.

Proposition E.3 (Exact pair risk under rare informative prompts)**.** *In the stated experiment, for a prior* $\Pr(W=0) = \pi$*, the optimal Bayes risk is*

$$R^*_\pi(Q) = (1-\varepsilon)^Q \int \min\{\pi p_0, (1-\pi)p_1\}\, d\mu. \tag{E.25}$$

Here $(1-\varepsilon)^0 = 1$*, including at* $\varepsilon = 1$*. If* $r_{\mathrm{mm}}(P_0, P_1)$ *is the infimum, over randomized tests based on* O*, of their maximum error in the two worlds, then*

$$R^*_{\mathrm{mm}}(Q) = (1-\varepsilon)^Q r_{\mathrm{mm}}(P_0, P_1). \tag{E.26}$$

For the equal prior, $R^*_{1/2}(Q) = \frac{1}{2}(1-\varepsilon)^Q \alpha$*. This also equals the minimax risk if a measurable involution of* O *exchanges* P_0 *and* P_1*.*

Proof. If $\varepsilon = 1$ and $Q \ge 1$, every prompt is odd; the separating query in part 1 attains risk zero, which is minimal. Otherwise $(1-\varepsilon)^Q > 0$, so the conditioning below is well defined. **1. Reward equivalence and a separating query.** Public state coordinates are the same in the two worlds. Their private endpoint difference at a fixed prompt is

$$(b_{\eta^1} + b_{\eta^0}) + (\theta^1_{u_\eta, z_0} + \theta^0_{u_\eta, z_0}) + \sum_{i=2}^{H} (\theta^1_{u_i, z_{i-1}} + \theta^0_{u_i, z_{i-1}}) = \sum_{i=2}^{H} \ell(u_i).$$

The first two terms are both $\ell(u_\eta)$ and cancel. On an even prompt, the endpoint targets and all verifier replies coincide: malformed strings fail in both worlds; admissibility checks the common public path, and the final bit is the same. Intermediate private bits are not verified. On an odd prompt, submit the canonical trajectory of candidate world 0. Its public path is legal in either world, while its endpoint is correct precisely when $W = 0$. The reply is therefore $1 - W$. One odd prompt identifies the candidate world exactly.

2. Lower bound with adaptive queries. Presample Q prompts even if a procedure stops early. Let E be the event that all are even; $\Pr(E \mid W, O) = (1-\varepsilon)^Q$. Conditional on E, the prompt law is world independent. At query j, a common history and the procedure's same random seed induce the same submitted string; part 1 gives the same reply. Induction over $j = 1, \dots, Q$ shows that the full training transcript conditional on E is a common Markov kernel applied to O. The independent test prompt supplies only additional world-independent randomness. From any output, define a test choosing world 0 exactly when the output passes its endpoint verifier, and choosing 1 otherwise. Since the two odd-prompt targets are disjoint, this test's error is at most the output's endpoint failure in each world. Thus, conditional on E, no procedure improves on a randomized test based on O. For prior π, pointwise choice between the two weighted densities gives the error integral in (E.25). Dropping nonnegative errors on E^c proves its lower bound. For the minimax claim, the same induced test has two conditional errors e_0, e_1, with $\max_w e_w \ge r_{\mathrm{mm}}$. The unconditional maximum error is at least $(1-\varepsilon)^Q \max_w e_w$, proving the lower bound in (E.26).

3. Attainment and symmetry. Use the separating candidate at the first odd prompt. If there is one, execute the identified world at test time. If none occurs, use a Bayes-optimal test of O, or an arbitrarily near-optimal minimax test, and execute the chosen world. At an odd test prompt the output is correct exactly when this choice is correct. These procedures attain the Bayes bound and approach the minimax infimum. For an involution T exchanging the two laws, transform a test f into $1 - f \circ T$; its two errors are those of f in reverse order. A fair mixture has equal errors, each the equal-prior average error of f. Applying this to a Bayes-optimal test proves that minimax equals equal-prior Bayes risk under symmetry. Without this assumption, the two quantities need not agree. □

An exact source-cost specialization. Suppose O consists of n independent raw source records from the usual even-suffix source law, together with an independent, world-independent initializer. The initial state is uniform, U_1 is uniform, and the suffix is uniform conditional on even tag parity. Each state transition uses $aP_U + (1-a)\Pi$ with $0 \le a \le 1$ and fresh transition noise. The endpoint values $a = 0, 1$ below concern this information experiment; the finite-training results keep their own prescribed $a = 1 - c/H$. Put

$$G_H(z) = \frac{(1+z)\big[(1+z)^{H-1} + (1-z)^{H-1}\big]}{2^H}, \tag{E.27}$$

$$h = a + \frac{2(1-a)}{q}, \qquad r = \frac{a + (1-a)/q}{h}, \qquad c_k = [z^k]G_H(z)^n. \tag{E.28}$$

For $j \ge 0$ define the majority error with a fair tie break by

$$b_j(r) = \sum_{t=0}^{\lfloor (j-1)/2 \rfloor} \binom{j}{t} r^t (1-r)^{j-t} + \frac{\mathbf{1}_{\{j \text{ even}\}}}{2} \binom{j}{j/2} [r(1-r)]^{j/2}. \tag{E.29}$$

For odd j, the tie term is absent; $b_0(r) = 1/2$ and an empty sum is zero.

Corollary E.4 (Source records and reward queries for the pair)**.** *Under this specialization, the equal-prior Bayes and minimax risks both equal*

$$(1-\varepsilon)^Q \sum_{k=0}^{nH} c_k \sum_{j=0}^{k} \binom{k}{j} h^j (1-h)^{k-j} b_j(r). \tag{E.30}$$

In the noiseless case $a = 1$, this reduces to

$$R^*(n, Q) = \tfrac{1}{2}\, 2^{-n(H-1)} (1-\varepsilon)^Q. \tag{E.31}$$

Consequently, in the noiseless case $a = 1$, for $0 < \delta < 1/2$ and $0 \le \varepsilon < 1$, error at most δ is possible exactly when

$$n(H-1)\log 2 + Q[-\log(1-\varepsilon)] \ge \log \frac{1}{2\delta}. \tag{E.32}$$

Proof. **1. Source symmetry.** For a record with states (s_j, z_j) and controls (b_j, u_j), leave controls and public states fixed and replace each private state by $s_j + \sum_{i=1}^{j} \ell(u_i)$, leaving s_0, PAD and EOS fixed. Applying this map twice is the identity. A transition following θ^0 becomes a transition following θ^1; a uniform reset remains uniform. The control law is unchanged, so the map exchanges the two complete-record laws. Apply it separately to all records and leave the independent initializer fixed. The resulting involution exchanges P_0, P_1; Proposition E.3 therefore equates Bayes and minimax risk.

2. Conditional likelihood of source innovations. Given the controls and initial state, define innovations relative to candidate 0:

$$e_j^z = z_j - z_{j-1} - u_j, \qquad e_j^s = s_j - s_{j-1} - b_j - \theta^0_{u_j, z_{j-1}}.$$

This is an invertible sequential transformation of the observed states. When $\ell(u_j) = 0$, the two conditional innovation laws coincide. When $\ell(u_j) = 1$, the innovation $(e_j^z, e_j^s) = (0, w)$ has probability $a + (1-a)/q$ in world w; each of the other $q - 1$ innovations has probability $(1-a)/q$. These probabilities do not depend on previous states. Iterating the conditional factorization shows that, given the controls, the innovation likelihood factors over positions. An innovation with $e_j^z \ne 0$ has world-independent likelihood. The event $e_j^z = 0$ has probability h in either world; conditional on it, the private innovation equals W with probability r. Thus each tag-one position gives either an erasure or an independent binary observation with correctness $r \ge 1/2$. Tag-zero positions give no information. Because the erasure and control laws do not depend on W, they supply no additional likelihood factor.

3. Counts and Bayes error. The balanced tag $\ell(u) = t + x \cdot y$ is a fair bit under uniform u. For one record, its first tag is an independent fair bit, while its $H - 1$ suffix tags are uniform over even-parity bit strings. The generating function of the latter count is $[(1+z)^{H-1} + (1-z)^{H-1}]/2^{H-1}$. Multiplication by $(1+z)/2$ gives (E.27). Independent records give $\Pr(K = k) = c_k$ for the total

tag-one count. Given $K = k$, the number J of nonerased observations is $\mathrm{Bin}(k, h)$. Given $J = j$, the number of correct binary observations is $\mathrm{Bin}(j, r)$. For $1/2 < r < 1$, the likelihood ratio increases with the number voting for world 1, so majority with a fair tie break is Bayes-optimal. At $r = 1/2$ all choices have average error $1/2$, and at $r = 1$ any nonempty sample identifies W; the same formula includes both endpoints. Its error is exactly (E.29). Averaging first over J, then K, and multiplying by the all-even reward probability from Proposition E.3 proves (E.30).

4. Noiseless specialization and the budget identity. At $a = 1$, $h = r = 1$. Any tag-one transition identifies the world, so only $K = 0$ contributes error $1/2$. Since $G_H(0) = 2^{-(H-1)}$, its probability is $2^{-n(H-1)}$. This proves (E.31). Taking logarithms of its comparison to δ gives (E.32), including $n = 0$ or $Q = 0$. □

Scope and limiting cases. For $n = 0$ or $a = 0$, the risk is $(1 - \varepsilon)^Q/2$. For $\varepsilon = 0$, reward cannot improve the source-only risk; for $\varepsilon = 1$ and $Q \geq 1$ it is zero. A checkpoint that already distinguishes the worlds has disjoint initial laws and zero risk even with $Q = 0$. Proposition E.3 accounts for checkpoint information through its joint law with source, including their dependence. This is an observation-based identification result; the optimization results use their specified random initializations. The factor $(1 - \varepsilon)^Q$ relies on charged, exogenous prompts: freely choosing an odd prompt reduces the identification cost to one query. The source channel and likelihood test are available because both candidate tables are known. Learning an arbitrary unknown table still incurs the separate identification costs in Theorem E.2; the pair formula does not remove them.

E.4 Information supplied by contextual source observations

Use the contextual observation model of Section 5.1, with $n_{\mathrm{ctx}} \geq 1$, odd cue dimension $k = p - 2$, independent uniform contexts and exact numerical source observations. A world fixes the slot orientations ξ, contextual signs s and cue vector τ. The error below averages the four held-out amplitude pairs, as well as contexts, queries and cues.

Theorem E.5 (Information supplied by contextual source observations)**.** *Two worlds are reward-equivalent on every training input if and only if*

$$\tau' = \varepsilon\tau, \qquad s'_\ell \xi'_\ell = \varepsilon s_\ell \xi_\ell \quad (\ell = 1, \dots, n_{\mathrm{ctx}}), \qquad \varepsilon \in \{\pm 1\}. \tag{E.33}$$

The n_{ctx} orientations may vary independently; the global sign leaves all predictions unchanged. Flipping an orientation reverses exactly the equal-sign test-amplitude labels in that context.

For the exact numerical source, let N_{obs} count contextual source records, and let $\mathcal{R}_{N_{\mathrm{obs}}}$ be the infimum, over all world-independent learners and finite fixed iid feedback-query budgets, of worst-world expected test error after N_{obs} iid source records. Then

$$\mathcal{R}_{N_{\mathrm{obs}}} = \frac{1}{4}(1 - 1/n_{\mathrm{ctx}})^{N_{\mathrm{obs}}}. \tag{E.34}$$

For $n_{\mathrm{ctx}} = 1$, the factor is one at $N_{\mathrm{obs}} = 0$ and zero at $N_{\mathrm{obs}} \geq 1$. For approximation tolerance $0 < \epsilon < 1$ and $K = n_{\mathrm{ctx}} 2^k$ context–cue types, $\lceil K \log(K/\epsilon) \rceil$ iid reward queries attain error at most $\mathcal{R}_{N_{\mathrm{obs}}} + \epsilon/2$ for every fixed world. With source observations alone, the minimax error is $1/2$ for every N_{obs}.

The proof is given in Section E.5.

E.5 The information supplied by source observations

We prove Theorem 5.1 for the contextual task in Section 5.1. The proof also applies to any fixed $0 < \rho_{\mathrm{mem}} < 1$ and odd $k \geq 1$. The count N_{obs} in this appendix is the number of observed source records.

E.5.1 Complete observation model and feedback equivalence

A source record reveals its context ℓ, the two columns $\kappa_j = Re_j$ of an independent Haar frame in $O(2)$, and $x = U_\ell \kappa_J$ for independent uniform $J \in \{1, 2\}$. In the full input these keys and query are embedded in the observed context block. The numerical fields G_1, G_2 are independent $N(0, I_{k+2})$

vectors, independent of the world, frame, context and index. The predicted bit is $\operatorname{sign}(G_{J,1}G_{J,2})$. In particular, the joint source law in one context contains no information about another context's orientation or about s, τ.

Set $a_\ell = s_\ell \xi_\ell$ and $M_\tau(z) = \operatorname{sign}(\tau^\top z)$. For a uniform sign vector Z,

$$\mathbb{E}[M_\tau(Z)Z_i] = \mu_k \tau_i, \qquad \mu_k = 2^{-(k-1)} \binom{k-1}{(k-1)/2} > 0. \tag{E.35}$$

To derive this identity, absorb τ_j into each Z_j and condition on the other $k-1$ signs. The conditional average of $\operatorname{sign}(Z_i + \sum_{j \neq i} Z_j)Z_i$ is one when the even sum is zero and zero otherwise.

Equality of rewards for every action implies equality of the training labels. Taking their first Fourier coefficients at $q = 1$ gives $a_\ell \tau_i = a'_\ell \tau'_i$ for every ℓ, i. Since all entries are signs, fixing one coordinate gives a common ε with $\tau' = \varepsilon\tau$ and $a' = \varepsilon a$. These relations also suffice, proving (5.1).

For equal-sign test amplitudes with common sign d, the label is $s_\ell d M_\tau(z)$. Changing ξ_ℓ while preserving the training equivalence changes s_ℓ by $-\varepsilon$, reversing this label. For opposite-sign amplitudes with first sign d, the label is $a_\ell v_q d M_\tau(z)$ and is invariant. The global transformation $(s, \tau) \mapsto (\varepsilon s, \varepsilon\tau)$ preserves both training and test predictions. Any changed orientation has equal-sign test witnesses, so this exhausts the prediction-inert transformations after test labels are included.

E.5.2 The exact source-data error floor

Precisely, define

$$\mathcal{R}_{N_{\mathrm{obs}}} = \inf_{Q \in \mathbb{N}_0} \inf_{\mathcal{A}\text{: at most } Q \text{ iid feedback queries}} \sup_{\xi, s, \tau} \mathbb{E}[\mathrm{err}(\mathcal{A})].$$

The algorithm and initialization are chosen independently of the world; observations may determine all subsequent computation. Source count N_{obs} and the feedback cap are fixed. Risk averages test inputs, training observations and algorithmic randomness.

For the lower bound, fix a, τ, choose independent uniform orientations ξ_ℓ, and set $s_\ell = a_\ell \xi_\ell$. All training labels are independent of these orientations. If a context has no source record, its orientation remains uniform conditional on the entire transcript: every observed source record belongs to another context, and every training reward has the same value under its flip. This also holds for adaptive actions or chosen training prompts, since each conditional reply distribution is unchanged. Equal-sign tests have probability $1/2$ and a fair hidden label, giving conditional error at least $1/4$ on a missing context. Its missing probability is $(1 - 1/n_{\mathrm{ctx}})^{N_{\mathrm{obs}}}$. Averaging over the uniform test context, then comparing the worst world to this prior, proves

$$\mathcal{R}_{N_{\mathrm{obs}}} \geq \tfrac{1}{4}(1 - 1/n_{\mathrm{ctx}})^{N_{\mathrm{obs}}}.$$

A visited orientation is identifiable almost surely from exact numerical observations. Its query belongs to $\{\kappa_1, \kappa_2\}$ for the identity and to $\{P\kappa_1, P\kappa_2\}$ for the swap. These sets intersect only at finitely many angles in each component of $O(2)$: a column must be fixed by P or mapped to the other column. These angles have Haar probability zero. Finite N_{obs} therefore permits exact recovery of every visited orientation, even though J is hidden.

For the upper bound, reduce a feedback prompt (ℓ, q, z) to $(\ell, v_q z)$. These reduced cells are uniform over $K = n_{\mathrm{ctx}} 2^k$ possibilities, and the label is $a_\ell M_\tau(v_q z)$. Choosing action $+1$ turns its binary correctness reply into the exact label. After Q iid queries,

$$\Pr[\text{some cell missing}] \leq K e^{-Q/K}.$$

For $0 < \epsilon < 1$, take $Q = \lceil K \log(K/\epsilon) \rceil$. On full coverage, the learner knows $g_\ell(z) = a_\ell M_\tau(z)$. It predicts $v_q d g_\ell(z)$ for an opposite-sign test pair with first sign d. For an equal-sign pair it predicts $\xi_\ell d g_\ell(z)$ when the source context was observed, and an independent fair sign otherwise. Its conditional expected error is exactly $F_{N_{\mathrm{obs}}} = \frac{1}{4}(1 - 1/n_{\mathrm{ctx}})^{N_{\mathrm{obs}}}$. Full feedback coverage is independent of source coverage. If feedback coverage fails, fair prediction gives error $1/2$, so the unconditional error is

$$F_{N_{\mathrm{obs}}} + (1/2 - F_{N_{\mathrm{obs}}}) \Pr[\text{some cell missing}] \leq F_{N_{\mathrm{obs}}} + \epsilon/2.$$

This bound holds for every fixed world. Taking $\epsilon \downarrow 0$ proves the risk infimum; no fixed finite budget is asserted to attain it exactly.

Table 32: What source learning leaves for reward queries. The first two routes compare information resources at fixed confidence and the stated noise regimes; the last gives sufficient observation and query budgets for a finite Adam learning path. Here q is the number of states and controls and H is the execution horizon.

Route and remaining reward problem	Observed source tokens	Training verifier queries
Information resources: identify a mechanism and its task binding		
Few source observations; reward queries recover the mechanism class	$O(\log q)$	$\Theta(q^2)$
Logarithmic H; $a = 1 - c/H$ with fixed c; Theorem E.2.		
Learn a transition table first; rewards identify only the initial operation	$O(q^2 \log q + H)$	At most $q - 1$
Neural training: acquire execution, then adapt the same parameters		
Source Adam and two sampled reward updates	At most NB_sT_s	$2B_R$

For the Adam route, N is tokens per source record, B_s records per source batch, T_s the maximum source batches inspected, and B_R completions per reward batch. There are at most $T_s - 1$ source updates. Each of the two reward updates uses B_R trajectories, totaling $2B_R$ verifier queries. Its horizon and rates are those of Theorem 4.2. Each verifier query additionally processes H prompt tokens and $H + 2$ generated decisions; verifier computation is separate.

Each query generates one binary action and makes one verification. There are no optimizer updates, test queries or selection queries. The construction stores K labels, K occupancy bits, $O(n_{\mathrm{ctx}})$ orientation and context flags, and $O(\log(Q + 1))$ counting bits, in addition to its numerical input and working memory. This finite table establishes information attainability; its exponential dependence on k is separate from the neural learning budgets.

Finally, fix ξ, τ and randomize the independent context signs s_ℓ. Source observations have the same distribution for all signs, while every test label changes with its context's sign. The prior-average source-only error is $1/2$, and fair prediction attains that value. This proves the last assertion of Theorem 5.1.

F Learning routes and resource comparisons

F.1 What additional source data can replace

The sequential family admits two resource endpoints. With little source, feedback learns the observable transition class and source resolves its remaining ambiguity. With abundant source, the mechanism can be learned first, leaving binding search. At failure probability δ_{tab} and source correlation $a > 0$, a classical table learner uses $O(q^2 a^{-2} \log(q/\delta_{\mathrm{tab}}) + H)$ raw source tokens and at most $q - 1$ fresh-prompt verifications, storing q^2 successor indices. The additive H pays for the final whole source record.

The query optimum in Theorem E.2 applies to its fixed-confidence, logarithmic-source and logarithmic-horizon regime. The neural theorem provides another finite training route with its own horizon and block rates. In the contextual task, Theorem 5.1 instead gives an exact source-dependent error floor, approached by a finite label-table learner. Together these results distinguish the cost of identifying the task from the cost of learning a neural computation.

The low-source row takes $\delta = 1/16$, $m = q/2 \geq 512$, fixed $c \in (0, 1/64]$, $a = 1 - c/H$ and $H - 1 = \lceil 2^{16} \log(2m) \rceil$; its upper bound is fresh-iid and its matching lower bound even permits chosen prompts. It solves binary equations for the observable class and then estimates two source sign statistics. The table-first route uses the retained independent triples in Section H.5.7: $M_0 = \lceil 8q^2 a^{-2} \log(2q^3/\delta_{\mathrm{tab}}) \rceil$ and at most $6M_0 + 2H + 2$ raw tokens. It stores a transition table and enumerates bindings. Its guarantee is exact execution on the table-recovery event, whose probability is at least $1 - \delta_{\mathrm{tab}}$. The Adam row instead gives the stated high-probability sampled-success bound under its prescribed horizon, model widths and real-arithmetic updates.

F.2 What remains for feedback and its alternatives

Source-only prediction and source replay contain no task-binding information in the sequential family. Under a uniform independent binding, any such decoder averages endpoint success at most $1/q$. In the contextual family, the corresponding source-only minimax error is $1/2$. More source prediction cannot identify those task parameters.

Task labels or expert supervised fine-tuning (SFT) can supply the missing information. Binary correctness in the contextual model already identifies a binary label once the action is known, making supervised adaptation a legitimate comparison. For batch item b, let R_b be its terminal reward, $Y_{b,j}$ its jth sampled token, and $\pi_w(Y_{b,j} \mid \mathrm{prefix}_{b,j})$ the next-token probability given the preceding tokens. On the exact baseline-zero REINFORCE batch, the reward-weighted likelihood loss

$$-\frac{1}{B_R}\sum_{b=1}^{B_R} R_b \sum_{j=1}^{n} \log \pi_w(Y_{b,j} \mid \mathrm{prefix}_{b,j})$$

has the same update when sampled tokens and rewards are held fixed and the optimizer settings and state agree. Additional off-policy passes or normalization by accepted samples define different updates. Verification-based selection at deployment introduces test-verifier access and its generation cost.

There is also a concrete source-informed controller in the sequential task. On the source event, local state and EOS argmaxes are correct. Freeze this model, choose a candidate first control, and greedily generate the rest. Enumerating at most $q-1$ candidates on fresh training prompts identifies the binding and gives exact deployment execution. In addition to the source budget, this costs at most $q-1$ verifications, $n(q-1)$ generated symbols and $(H+1)(q-1)$ forward evaluations, with zero additional gradient updates. Deployment uses $H+1$ forwards and n symbols without verification; retained context and workspace are charged. Section H.5.6 proves the margins and identification property.

The controller explicitly searches bindings; the Adam construction adapts the source model's own parameters using sampled trajectories. Both use the same acquired transition information, with different query and computation costs.

G Shared Gaussian network foundations

The following initialization and derivative statements are used by both sequential optimizers. Their distinct objectives, rates, stopping rules and probability allocations are verified in the separate applications.

G.1 Finite Gaussian initialization

Let N be the full padded context length, V the vocabulary size, R the head count, d the residual width, k_{head} the head width, and M_{ff} the feed-forward width. The local head width k_{head} equals $d_h = d/R$ in the source budget. The public-state cardinality remains $m = q/2$; write $k_\mu = \|\mu\|^2$ for the kernel-mean quantity in the feedback reference, to distinguish it from the head width.

Prescribed Gaussian member. For the complete block in (G.22), initialize all the following entries independently:

$$\begin{aligned}
E_{ij} &\sim \mathcal{N}(0, d^{-1}), & (Q_h)_{ij}, (K_h)_{ij}, (V_h)_{ij} &\sim \mathcal{N}(0, k_{\mathrm{head}}^{-1}),\\
(O_h)_{ij} &\sim \mathcal{N}(0, d^{-1}), & b_h(\ell) &\sim \mathcal{N}(0,1),\\
B_{ij}, C_{ij}, b_j, c_j &\sim \mathcal{N}(0,1), & U_{ij} &\sim \mathcal{N}(0, (M_{\mathrm{ff}} d)^{-1}),\\
(b_{\mathrm{out}})_i &\sim \mathcal{N}(0, d^{-1}), & (W_{\mathrm{out}})_{yi} &\sim \mathcal{N}(0, d^{-1}).
\end{aligned} \tag{G.1}$$

The attention-head sum has scale $R^{-1/2}$. Its fixed temperature is chosen from these dimensions and the confidence level, without using the unknown mechanism. Every displayed parameter trains from these prescribed width-dependent Gaussian scales.

Let ℓ_h be the lag with maximum relative bias in head h. The winners depend only on the independent relative biases. Conditional on coverage $n_\ell := \#\{h : \ell_h = \ell\} \geq R/(2N)$, put

$$\omega_\ell = \frac{n_\ell}{R}, \qquad w_\ell = \omega_\ell + \mathbf{1}_{\{\ell=0\}}, \qquad p = NV, \qquad D_w = \mathrm{diag}_\ell(w_\ell I_V). \tag{G.2}$$

Lemma G.1 (Hard-lag encoder frame)**.** *The hard-lag reference has*

$$z_{\mathrm{hard}}(x) = \sum_{\ell=0}^{N-1} A_\ell e_{x_\ell}, \qquad A_\ell = \mathbf{1}_{\{\ell=0\}} E + R^{-1/2} \sum_{h:\ell_h=\ell} O_h V_h E. \tag{G.3}$$

Here e_{x_ℓ} is a categorical basis vector, not a supplied semantic state representation. Set $\mathcal{A} = [A_0 \ \cdots \ A_{N-1}]$ and $T = \mathcal{A} D_w^{-1/2}$. For $0 < \xi \leq 1$, the sufficient widths

$$\begin{aligned} d &\geq 2048\xi^{-2}\left[p\log 9 + \log\frac{8}{\delta_{\mathrm{enc}}}\right], \\ k_{\mathrm{head}} &\geq \max\left\{1, \left\lceil \frac{4096N}{R\xi^2}\left[V\log 9 + \log\frac{8N}{\delta_{\mathrm{enc}}}\right]\right\rceil\right\} \end{aligned} \tag{G.4}$$

give $\|T^\top T - I_p\|_{\mathrm{op}} \leq \xi$ with conditional probability at least $1 - \delta_{\mathrm{enc}}$, uniformly over winner allocations satisfying coverage.

Proof. Write $\mathcal{A} D_w^{-1/2} = C_{\mathrm{res}} + OZ$, where O concatenates the O_h and is independent of E and all V_h. Explicitly, $C_{\mathrm{res}} = [E/\sqrt{w_0}, 0, \ldots, 0]$ and the head/lag block of Z is $Z_{h,\ell} = \mathbf{1}_{\{\ell_h=\ell\}} V_h E/\sqrt{R w_\ell}$.

With $\varepsilon = \xi/8$, a Gaussian $1/4$-net bound gives $\|E^\top E - I_V\|_{\mathrm{op}} \leq \varepsilon$. Conditional on E, the stacked value maps in lag group ℓ have $n_\ell k_{\mathrm{head}}$ independent Gaussian rows. Set $H_E = E^\top E$ and stack the lag-group maps $V_h E$ as Y_ℓ. On the encoder event H_E is positive definite. Conditional on E, $G_\ell = \sqrt{k_{\mathrm{head}}} Y_\ell H_E^{-1/2}$ has $n_\ell k_{\mathrm{head}}$ independent standard Gaussian rows. Their normalized Gram matrices are Wishart; the net bound and a union over all N groups give

$$\left\| \frac{G_\ell^\top G_\ell}{n_\ell k_{\mathrm{head}}} - I \right\| \leq \varepsilon, \qquad \left\| \frac{Y_\ell^\top Y_\ell}{n_\ell} - H_E \right\| \leq \|H_E\|\varepsilon \leq (1+\varepsilon)\varepsilon.$$

The second inequality follows by multiplying the first error on both sides by $H_E^{1/2}$. Thus

$$\begin{aligned} \|C_{\mathrm{res}}^\top C_{\mathrm{res}} + Z^\top Z - I\|_{\mathrm{op}} &\leq \varepsilon + (1+\varepsilon)\varepsilon, \\ \|C_{\mathrm{res}}\|_{\mathrm{op}}^2 &\leq 1 + \varepsilon, \qquad\qquad \|Z\|_{\mathrm{op}}^2 \leq (1+\varepsilon)^2. \end{aligned} \tag{G.5}$$

Conditional on E and the value maps, Gaussian subspace isometry for O on $\mathrm{range}(Z)$ bounds its Gram error by $\varepsilon\|Z\|_{\mathrm{op}}^2$. The residual cross term is retained. With orthonormal bases U_C, U_Z of the two fixed column spaces, $U_C^\top O U_Z$ has independent Gaussian entries of variance $1/d$. Two $1/4$-nets under (G.4) bound its norm by ε, giving $\|C_{\mathrm{res}}^\top O Z\|_{\mathrm{op}} \leq \varepsilon(1+\varepsilon)^{3/2}$. The four failure allocations are $\delta_{\mathrm{enc}}/4$ each. The total error is at most

$$\varepsilon + (1+\varepsilon)\varepsilon + \varepsilon(1+\varepsilon)^2 + 2\varepsilon(1+\varepsilon)^{3/2} \leq 6\varepsilon < \xi. \tag{G.6}$$

The elementary Gaussian net estimate used here states that an $M \times s$ standard Gaussian matrix G satisfies $\|G^\top G/M - I\|_{\mathrm{op}} \leq \varepsilon$, except with probability η, whenever

$$M \geq 32\varepsilon^{-2}\left[s\log 9 + \log\frac{2}{\eta}\right].$$

Fixed-direction chi-square tails followed by a $1/4$-net prove it. Gaussian scalar tails and two nets give the residual cross bound. Content-score events are combined afterwards by a union bound; conditioning on them first would invalidate the Gaussian independence. □

Bilinear tensor representation. Define

$$r_0(x) = D_w^{1/2}\,\mathrm{concat}_\ell(e_{x_\ell}), \qquad \|r_0(x)\|^2 = \sum_\ell w_\ell = 2,$$
$$k_{\mathrm{aug}} = p+1, \qquad s_{\mathrm{sym}} = \frac{k_{\mathrm{aug}}(k_{\mathrm{aug}}+1)}{2},$$
$$D_{\mathrm{feat}} = p + s_{\mathrm{sym}} + 1, \qquad \Phi(x) = \begin{bmatrix} r_0(x) \\ \mathrm{symvec}\left(\begin{bmatrix} r_0(x) \\ 1 \end{bmatrix}\begin{bmatrix} r_0(x) \\ 1 \end{bmatrix}^\top\right) \\ 1 \end{bmatrix}. \tag{G.7}$$

Symmetric vectorization is orthonormal for the Frobenius inner product, so off-diagonal coordinates carry a factor $\sqrt{2}$. Then $z_{\mathrm{hard}} = Tr_0$, and

$$\|\Phi(x)\|^2 = 2+9+1 = 12, \qquad h_{\mathrm{hard}}(x) = C_{\mathrm{all}}\Phi(x). \tag{G.8}$$

The latter identity is exact, obtained by expanding the actual bilinear feed-forward block and its biases; it assumes no output-value matching.

Lemma G.2 (Realized coefficient-operator concentration)**.** *Suppose* $0 < \gamma \le 1$ *and* $\|T^\top T - I\|_{\mathrm{op}} \le \gamma/16$. *The additional conditions*

$$d \ge 4096\gamma^{-2}\left[D_{\mathrm{feat}}\log 9 + \log\frac{8}{\delta_{\mathrm{ffn}}}\right], \tag{G.9}$$
$$J = s_{\mathrm{sym}}\log 9 + \log\frac{8}{\delta_{\mathrm{ffn}}}, \qquad A_0 = 1+\gamma^{-2}J,$$
$$L_0 = \log\frac{\mathrm{e}A_0 k_{\mathrm{aug}}}{\delta_{\mathrm{ffn}}\gamma}, \qquad M_{\mathrm{ff}} = \left\lceil 2^{50}A_0^2(k_{\mathrm{aug}}+L_0)^8\right\rceil \tag{G.10}$$

give $\|C_{\mathrm{all}}^\top C_{\mathrm{all}} - I_{D_{\mathrm{feat}}}\|_{\mathrm{op}} \le \gamma$ *with conditional probability at least* $1-\delta_{\mathrm{ffn}}$. *In particular* $d \ge D_{\mathrm{feat}}$.

Proof. The residual Gram error is a principal block of the full Gram error, and must therefore meet the desired total accuracy. Let $H_{\mathrm{aug}} = \mathrm{diag}(T^\top T, 1)$. A neuron's two augmented coefficient vectors are independent $\mathcal{N}(0, H_{\mathrm{aug}})$. Whitening gives independent $a, b \sim \mathcal{N}(0, I_{k_{\mathrm{aug}}})$ and

$$v = \mathrm{symvec}\left(\frac{ab^\top + ba^\top}{2}\right).$$

For each symmetric S,

$$\mathbb{E}(a^\top S b)^2 = \|S\|_F^2, \qquad \mathbb{E}vv^\top = I_{\mathrm{sym}}. \tag{G.11}$$

Furthermore,

$$\mathbb{P}\{\|v\|^2 > (2k_{\mathrm{aug}}+3t)^2\} \le 2\exp(-t),$$
$$\mathbb{E}\|v\|^4 \le [k_{\mathrm{aug}}(k_{\mathrm{aug}}+2)]^2. \tag{G.12}$$

Truncation at squared norm $B_0 = (2k_{\mathrm{aug}}+3t)^2$ has covariance bias at most $5k_{\mathrm{aug}}^2\exp(-t/2)$. Set $\varepsilon = \gamma/16$. It suffices that

$$t \ge \max\left\{1, \log\frac{8M_{\mathrm{ff}}}{\delta_{\mathrm{ffn}}}, 2\log\frac{10k_{\mathrm{aug}}^2}{\varepsilon}\right\},$$
$$M_{\mathrm{ff}} \ge 8B_0^2\varepsilon^{-2}\left[s_{\mathrm{sym}}\log 9 + \log\frac{8}{\delta_{\mathrm{ffn}}}\right]. \tag{G.13}$$

Scalar Hoeffding on a $1/4$-net, together with the truncation bias and the event of any realized truncation, bounds the actual whitened sample-covariance error by ε, with failure probability $\delta_{\mathrm{ffn}}/2$. The explicit choice (G.10) satisfies (G.13) with $t = 128(k_{\mathrm{aug}}+L_0)$; no implicit $\log M_{\mathrm{ff}}$ condition remains. Transforming back through $H_{\mathrm{aug}}^{1/2}$, the actual product-feature matrix F satisfies

$$\|F^\top F/M_{\mathrm{ff}} - I_{\mathrm{sym}}\|_{\mathrm{op}} \le (1+\gamma/16)^2\varepsilon + 2\gamma/16 + (\gamma/16)^2 \le \gamma/4. \tag{G.14}$$

Conditional on T, F, the realized $Z = [UF, b_{\text{out}}]$ has distribution $XQ^{1/2}/\sqrt{d}$, where

$$Q = \operatorname{diag}(F^\top F/M_{\text{ff}}, 1)$$

and X is standard Gaussian. This representation is independent of the now-fixed T. Gaussian nets and (G.9) imply, simultaneously,

$$\|X^\top X/d - I\|_{\text{op}} \le \gamma/8, \qquad \|T^\top X/\sqrt{d}\|_{\text{op}} \le \gamma/4 \tag{G.15}$$

except with probability $\delta_{\text{ffn}}/2$. Hence

$$\|Z^\top Z - I\|_{\text{op}} \le 13\gamma/32, \qquad \|T^\top Z\|_{\text{op}} \le \sqrt{5}\,\gamma/8.$$

Since $C_{\text{all}} = [T, Z]$, the total block error is at most

$$\max\{\gamma/16, 13\gamma/32\} + \sqrt{5}\,\gamma/8 < \gamma.$$

This proves concentration of the realized coefficient operator. □

The ideal kernel is

$$\langle \Phi(x), \Phi(x') \rangle = k_{\text{base}}(x, x')^2 + 3k_{\text{base}}(x, x') + 2, \qquad k_{\text{base}}(x, x') = \sum_\ell w_\ell \mathbf{1}_{\{x_\ell = x'_\ell\}}. \tag{G.16}$$

Thus $K_{\max} = 12$, and on coverage

$$3w_\ell + w_\ell^2 \ge \frac{3}{2N}, \qquad k_{0H} = 2w_0 w_H \ge \frac{1}{N}. \tag{G.17}$$

These singleton weights and pair coefficients are those used in the source reference.

Finite soft attention and nonzero initial readout. For $N \ge 2$ and $0 < \ell < N-1$, the independent bias event has coverage and every pairwise gap at least g, with probability at least $1 - \delta_{\text{bias}}$, when

$$R \ge 8N \log \frac{2N}{\delta_{\text{bias}}}, \qquad g = \frac{\delta_{\text{bias}}}{RN^2}. \tag{G.18}$$

For coverage, uniformly distributed winning lags give failure at most $N \exp(-R/(8N)) \le \delta_{\text{bias}}/2$. A pairwise difference of independent unit Gaussians has density at most $1/(2\sqrt{\pi})$; the union over at most $RN^2/2$ pairs gives gap failure at most $RN^2 g/(2\sqrt{\pi}) < \delta_{\text{bias}}/2$. A separate Gaussian norm event supplies a deterministic M_{init} bounding every initialized block. Content scores then have magnitude at most M_{init}^4. Taking $t_{\text{att}} \le g/(4M_{\text{init}}^4)$ keeps the winner gap, in unscaled score units, at least $g/2$. If also

$$t_{\text{att}} \le \frac{g}{2\log((N-1)/\ell)}, \tag{G.19}$$

each head's off-winner mass is at most ℓ. With $Z_0 = M_{\text{init}} + \sqrt{R}\,M_{\text{init}}^3$,

$$\sup_x \|h_{\text{soft},0}(x) - h_{\text{hard},0}(x)\| \le 2\sqrt{R}\,M_{\text{init}}^3 \ell[1 + 2M_{\text{init}}^3(Z_0 + 1)]. \tag{G.20}$$

This holds uniformly over all strings, including repeated padding and wrong generated tokens. The derivatives are bounded in Section G.2; subsequent motion is controlled separately in Sections H.5 and J.3.1. No infinite-width or zero-temperature limit is interchanged.

For an explicit coarse norm event, let P be the number of initialized scalar parameters, all with variance at most one in (G.1). With failure probability at most δ_{norm}, every block Frobenius, spectral or vector norm is bounded by

$$M_{\text{init}} = \max\left\{1, \sqrt{2P \log \frac{2P}{\delta_{\text{norm}}}}\right\}. \tag{G.21}$$

Bound every scalar Gaussian and then each block's Frobenius norm. This conservative bound is polynomial in the prescribed dimensions. It permits the temperature and all positive rates to be selected before drawing initialization or source data.

G.2 Uniform derivatives of the complete block

The complete one-block member and its derivatives. Use R heads, N retained positions, relative biases $b_h(\ell)$ and fixed temperature $t_{\mathrm{att}} > 0$. Define

$$\begin{aligned} e_\ell &= E(x_\ell), \\ \alpha_h(\ell) &= \mathrm{softmax}_\ell\left[\frac{b_h(\ell)}{t_{\mathrm{att}}} + \frac{\langle Q_h e_0, K_h e_\ell\rangle}{\sqrt{k_{\mathrm{head}}}}\right], \\ z &= e_0 + R^{-1/2}\sum_{h=1}^{R} O_h \sum_{\ell=0}^{N-1} \alpha_h(\ell) V_h e_\ell, \\ h_\psi &= z + U[(Bz+b)\odot(Cz+c)] + b_{\mathrm{out}}. \end{aligned} \tag{G.22}$$

Every displayed matrix, embedding and bias is included in ψ with positive rates specified separately in the Adam and SGD applications. Causal attention sees every real token in the padded context; the displayed equations specify the complete raw architecture.

If every block matrix spectral norm, bias-vector norm and embedding-column norm is at most $M \geq 1$, put

$$\begin{aligned} Z &= M + \sqrt{R}\, M^3, \\ A_{\mathrm{raw}} &= Z + M^3(Z+1)^2 + M, \\ D_z &= 1 + \sqrt{R}[3M^2 + M^3 N(4M^3 + t_{\mathrm{att}}^{-1})], \\ D &= D_z[1 + 2M^3(Z+1)] + 3M^2(Z+1)^2 + 1. \end{aligned} \tag{G.23}$$

Then $\|h_\psi\| \leq A_{\mathrm{raw}}$ and $\|D_\psi h_\psi\|_{\mathrm{op}} \leq D$ in the concatenated Euclidean/Frobenius metric. For a global unit parameter direction, each individual matrix or bias direction has norm at most one. The attention-score derivative is at most $4M^3 + t_{\mathrm{att}}^{-1}$, while

$$\|\dot{\alpha}\|_1 \leq N\|\dot{\mathrm{score}}\|_\infty.$$

Differentiating the O, V, e paths gives D_z. Both feed-forward factors have norm at most $M(Z+1)$, yielding the displayed direct-factor and through-z terms. No block is omitted.

G.3 The nonlinear normalized member

Definition and initialization. Retain every raw attention and bilinear block, replacing only the final output feature by

$$h_{\mathrm{norm}} = g \odot \mathcal{R}(h_{\mathrm{raw}}), \qquad \mathcal{R}(x) = \frac{\sqrt{2}\, x}{\sqrt{1 + \|x\|^2/K}}, \qquad K = 12. \tag{G.24}$$

The gain $g \in \mathbb{R}^d$ is trainable and starts at the task-independent vector of ones. Every original random parameter retains its prescribed Gaussian law. Each original backbone block and the gain update at the same positive backbone rate; the readout uses the specified faster rate. The gain initializer is thus deterministic.

This nonlinear final RMS-type normalization has a fixed positive denominator offset and trainable coordinate gain. It changes the feature and its derivatives away from the constant-norm ideal feature manifold.

The realized normalized initial feature. The normalization has global derivative norm at most $\sqrt{2}$ and global output norm at most $\sqrt{2K}$. On the actual hard-feature event, $\|C_{\mathrm{all}}^\top C_{\mathrm{all}} - I\|_{\mathrm{op}} \leq \gamma$ and $\|\Phi\|^2 = K$. For $\gamma \leq 1/4$,

$$\|\mathcal{R}(C_{\mathrm{all}}\Phi) - C_{\mathrm{all}}\Phi\| \leq 4\gamma. \tag{G.25}$$

To see this, write $\|C_{\mathrm{all}}\Phi\|^2/K = 1 + t$, where $|t| \leq \gamma$. The factor $\sqrt{2}/\sqrt{2+t}$ differs from one by at most γ, and $\|C_{\mathrm{all}}\Phi\| < 4$. Thus the estimate concerns the realized random operator. If the raw soft/hard features differ by at most ζ_{raw}, then

$$\|h_{\mathrm{norm},0} - C_{\mathrm{all}}\Phi\| \leq \sqrt{2}\,\zeta_{\mathrm{raw}} + 4\gamma. \tag{G.26}$$

Derivatives and positive updates. On a radius-one ball in all backbone parameters, $\|g\|_\infty \le 2$. Therefore

$$\|h_{\mathrm{norm}}\| \le 2\sqrt{2K} < 12 = A. \tag{G.27}$$

Let D_{raw} be the raw derivative bound (G.23), evaluated at $M_{\mathrm{init}} + 1$. The concatenated derivative norm is bounded by

$$D_{\mathrm{norm}} = 2\sqrt{2}\, D_{\mathrm{raw}} + \sqrt{2K}. \tag{G.28}$$

The first term differentiates every original attention, embedding and feed-forward parameter through $\mathcal{R}$. The second differentiates the coordinate gain, using $\|\mathcal{R}(h_{\mathrm{raw}})\|_\infty \le \sqrt{2K}$. Both the gain and all original blocks are included in the motion bound.

H Sequential acquisition and sampled Adam adaptation

H.1 Finite source learning and sampled Adam improvement

Use the two architectures and source-to-feedback update rule in Section 4: all parameters train, source cross-entropy has the coupled positive ridge penalty, and bias-corrected Adam uses $(\beta_1, \beta_2) = (0.9, 0.999)$. At the source stopping state, retain every parameter, remove the penalty and reset both moments. Feedback uses fresh on-policy trajectories, baseline zero, and retains every trajectory and every output score. The complete choices of horizon, width, Gaussian scales, positive block rates, stopping threshold and batch sizes are those in Section H.5. Write $\varepsilon_{\mathrm{logit}} := h$ for the tolerance in (H.97). The metrics E, M, S, V are defined in Section 3.

Theorem H.1 (Finite source learning and sampled Adam improvement)**.** *Fix $q = 2^{2r+2}$ with $r \ge 1$, $c = 1/64$ and $\delta \in (0,1)$. For either network member in (4.2)–(4.4), with the explicit horizon, dimensions and training choices of Section H.5, the source algorithm uses at most T_s batches of B_s records and makes at most $T_s - 1$ updates; it is followed by exactly two feedback Adam updates, each on B_R sampled trajectories. For every fixed world, with probability at least $1 - \delta$, source stopping succeeds and, simultaneously for every held-out prompt X,*

$$\min\{E(w_f, X), M(w_f, X), S(w_f, X)\} \ge 7/8, \tag{H.1}$$

$$\max\{E(w_s, X), M(w_s, X), S(w_s, X)\} \le 1/q + 2n\varepsilon_{\mathrm{logit}}, \tag{H.2}$$

$$\min_{J \in \{E,M,S\}} \big[J(w_f, X) - J(w_s, X)\big] \ge 3/4, \tag{H.3}$$

$$V(w_f, X) \le 9c/8 + 1/128, \tag{H.4}$$

where w_s is the source stopping state and w_f the second feedback iterate. With $N = 2H + 2$ source tokens per record and $n = H + 2$ decisions per trajectory,

$$N_{\mathrm{pre}} \le NB_sT_s, \qquad Q = 2B_R, \tag{H.5}$$

$$N_{\mathrm{generated}} = 2nB_R, \qquad N_{\mathrm{updates,source}} \le T_s - 1, \qquad N_{\mathrm{updates,feedback}} = 2. \tag{H.6}$$

The proof is given in Section H.5.

H.2 Applying the shared bounds to Adam

The substitutions into Theorems G.1 and G.2 are $N = 2H + 2$, $V = 2q + 2$, $R = R_{\mathrm{head}}$, $k_{\mathrm{head}} = d/R_{\mathrm{head}}$, $D_{\mathrm{feat}} = k$, and $\xi_{\mathrm{enc}} = \gamma/16$ from Section H.5. In particular, $2048\xi_{\mathrm{enc}}^{-2} = 2^{19}\gamma^{-2}$ and $4096\xi_{\mathrm{enc}}^{-2} = 2^{20}\gamma^{-2}$. Thus the encoder and head premises are exactly the first and third terms in $D_\Gamma\gamma^{-2}$; the second term supplies the output Gram premise. Since $\gamma = d^{-1/8}$, $d \ge D_\Gamma^{4/3}$ implies all three. Rounding d to a multiple of R_{head} preserves them and makes the head width integral. The FFN width is the shared lemma's width with $(J, A_0, L_0) = (J_{\mathrm{ff}}, A_{\mathrm{ff}}, L_{\mathrm{ff}})$. The bias, encoder, FFN and global-norm events each use $\delta/16$; the encoder and FFN conditional bounds are combined by the tower property, then with the other events by a union bound, without conditioning the Gaussian laws on content scores. The temperature satisfies both inequalities in Equation (G.19) and the preceding gap condition; its choice of ℓ makes the raw feature error at most z_{raw}. For the normalized member, $\gamma \le \varepsilon_{\mathrm{src}}/16$ and $z_{\mathrm{raw}} \le \varepsilon_{\mathrm{src}}/(8\sqrt{2})$ give $\sqrt{2}z_{\mathrm{raw}} + 4\gamma \le 3\varepsilon_{\mathrm{src}}/8$. The separate row, readout, geometry, batch and reward events keep their allocations in (H.95); none is replaced by the SGD allocation.

Evaluate the displayed derivative bound at $M = M_{\text{init}} + 1$ for the whole radius-one parameter ball. The source and feedback displacement bounds in Section H.5 keep every supplied backbone parameter in that ball. The actual initialization has bounded features, and its smaller source tube gives the feature envelope $A_{\text{src}} = 12$. The amplitude-free coordinate Adam bound pays motion even when coupled regularizer gradients are large. The gain is included in the backbone parameter count and in every source and feedback update. The source feature tolerance and subsequent motion are selected in Section H.5; their errors include the initial 4γ normalization discrepancy. The proof uses the same raw coefficient operator as an analysis reference and makes no rotation-equivariance claim for the trained normalized network.

H.3 Finite source acquisition by coupled-regularized Adam

The source stage learns a transition function and output format before any terminal reward is observed. We first identify an invariant population ridge optimum, transfer it to the realized hard-feature head, and then show that the actual sampled Adam algorithm reaches a source-observable stopping condition. The finite initializer and full-model derivative bounds are in Section G; Section H.5 chooses the joint horizon, widths, source precision, and failure allocations. Throughout this section, $N = 2H + 2$, $H \geq 5$, $V = 2q + 2$, and $K = 12$. We write T_s, B_s, η_s for the source inspection limit, complete-record batch size, and effective head step. The physical head step is $\eta_s/\sqrt{d}$.

H.3.1 Canonical all-target source structure

Let $F_0(\Theta)$ and $F_+(\Theta)$ denote the unconditioned and tag-plus population source cross-entropies, averaged over all N raw next-token targets. In the unconditioned law, S_0 and all controls are independent uniform. Conditional on the controls, the state transitions use $aP_U + (1-a)\Pi$. The canonical feature map has kernel

$$\langle \Phi(x), \Phi(x') \rangle = k_{\text{base}}(x, x')^2 + 3k_{\text{base}}(x, x') + 2, \qquad k_{\text{base}}(x, x') = \sum_{l=0}^{N-1} w_l \mathbf{1}\{x_l = x'_l\}. \tag{H.7}$$

Its squared feature norm is K, its singleton coefficients are at least $k_{\min} = 3/(2N)$, and its lag-pair coefficient satisfies $k_{0H} \geq 1/N$. The categorical singleton/pair representation and the tensor representation in Section G have the same kernel. Mapping finite linear combinations of one representation to the other preserves their Gram quadratic form, and hence gives an isometry of their spans. Projecting any comparator onto that span cannot increase its norm.

For a state-output row y, define

$$\begin{aligned} h_{1,y}(x) &= \mathbf{1}\{x_0 \in \mathcal{U}, x_H \in \mathcal{S}\} \left(\mathbf{1}\{y = \phi_{x_0}(x_H)\} - q^{-1}\right), \\ h_{2,y}(x) &= \mathbf{1}\{x_0 \in \mathcal{S}, x_H \in \mathcal{U}\} \left(\mathbf{1}\{y = \phi_{x_H}(x_0)\} - q^{-1}\right), \end{aligned} \tag{H.8}$$

and set both functions to zero on nonstate rows. Their pair supports are disjoint. The active lag pair is (U_1, S_0) at the first-state prediction, (S_{i-1}, U_i) at subsequent state predictions, and (S_H, U_1) at EOS. Both functions vanish at the initial control-choice prediction because lag H is PAD.

The gradient of F_0 preserves the sum of mask/class logits and $\alpha_1 h_1 + \alpha_2 h_2$. Here α_j are semantic logit amplitudes, distinct from Adam's moment parameters. Put

$$b_j = \frac{\exp(\alpha_j) - 1}{\exp(\alpha_j) + q - 1}. \tag{H.9}$$

If v_i is the correct state-class mass at state target i and v_E is the incorrect state-class mass at EOS, the semantic coordinates of an auxiliary Euclidean descent step are

$$\begin{aligned} \alpha_1^+ - \alpha_1 &= \frac{\eta k_{0H}}{Nq^2}(a - v_1 b_1), \\ \alpha_2^+ - \alpha_2 &= \frac{\eta k_{0H}}{Nq^2}\left[\sum_{i=2}^{H}(a - v_i b_2) - v_E b_2\right]. \end{aligned} \tag{H.10}$$

The EOS contribution is essential because h_2 is active on the incorrect state class there.

To verify closure, center in both interaction arguments. At the first-state query, independence leaves only the pair (U_1, S_0). At a repeated-state query, averaging the current control eliminates coordinates omitting it; a surviving coordinate pairs that control with the immediately preceding state. At EOS, integrating uniform S_0 makes $(U_2, \dots, U_H, S_1, \dots, S_H)$ independent of U_1. Coordinates omitting both U_1 and S_0 vanish by averaging U_1. Among coordinates retaining U_1, an earlier state leaves a later control to average, and a control leaves uniform S_0; only (U_1, S_H) survives. Among coordinates retaining S_0 but not U_1, an earlier state leaves a later control, S_H leaves a suffix average followed by U_1, and a suffix control leaves another suffix control when $H \geq 3$. The surviving semantic coefficient has factor q^{-2}. Distinct individual record variables are pairwise independent uniform, which also preserves the mask/class space. These categorical identities hold on every prefix with the legal class/PAD pattern, including incorrect generated states.

Let $h_q(a)$ be the entropy of a point mass mixed with the uniform q-state law at weights $a, 1-a$. The unconditioned entropy per optimized target is

$$H_0 = \frac{(H+1)\log q + H h_q(a)}{N}. \tag{H.11}$$

For $\varepsilon > 0$, set

$$\alpha^* = \log \frac{1+(q-1)a}{1-a}, \qquad M = \alpha^* + \log(V/\varepsilon), \qquad P_l = \mathbf{1}\{x_l = \mathrm{PAD}\}. \tag{H.12}$$

Use the class logits

$$\begin{aligned} c_S &= M(P_0 - P_H + P_{2H}), & c_U &= M(P_H - P_0), \\ c_E &= M(1 - P_{2H}), & c_{\mathrm{PAD}} &= 0, \end{aligned} \tag{H.13}$$

and add $\alpha^*(h_1 + h_2)$. They select the correct class at every raw target; the extra α^* in M pays for the active incorrect state partition at EOS. The constant term of c_E can be represented by all lag-zero indicators. This comparator has excess CE at most ε and squared norm bounded by

$$D^*(\varepsilon) = \frac{(5q+V+1)M^2}{k_{\min}} + \frac{2q(q-1)(\alpha^*)^2}{k_{0H}}. \tag{H.14}$$

For deterministic algorithmic choices, we use the upper envelope obtained by substituting $k_{\min} = 3/(2N)$ and $k_{0H} = 1/N$ in this expression.

Let KL denote Kullback–Leibler divergence. For any invariant reference, let $D_j = \mathrm{KL}(T_a \| T_{b_j})$, where T_b denotes the relevant true transition permutation mixed with uniform at weights $b, 1-b$. The exact excess decomposition is

$$D_1 + (H-1)D_2 + \sum_{\text{all stages } s} [-\log v_s] = N\big[F_0(\Theta) - H_0\big]. \tag{H.15}$$

Consequently, if the excess is E, then

$$|b_1 - a| \leq \frac{q}{q-1}\sqrt{NE/2}, \qquad |b_2 - a| \leq \frac{q}{q-1}\sqrt{\frac{NE}{2(H-1)}}. \tag{H.16}$$

For $a = 1 - c/H$, the condition $E \leq c^2(q-1)^2/(32q^2H^2N)$ gives $|b_j - a| \leq c/(8H)$ for both transitions. The first transition is observed at one position per record, which determines the stronger precision requirement.

H.3.2 One-time transfer of the ridge optimum

Write $C_{\mathrm{tag}} = \prod_{i=2}^{H}(-1)^{\ell(U_i)}$ and $\rho_c = 2^{-r}$. The signed contraction holds on the full state space for every unknown mechanism. Indeed, in the Koopman convention $P_A f = f \circ \phi_A$, decompose $f(s,z) = f_+(z) + (-1)^s f_-(z)$. In $M_c = q^{-1}\sum_{b,u}(-1)^{\ell(u)}P_{(b,u)}$, averaging the control bit b annihilates the odd component. The even component reduces to normalized convolution by $(-1)^{t+x\cdot y}$ on the public Z coordinates, independently of θ. Its Walsh eigenvalues are zero at dual t-frequency zero and $2^{-r}(-1)^{\alpha\cdot\beta}$ at dual t-frequency one. Thus $\|M_c\|_{\mathrm{op}} = 2^{-r}$. For the noisy kernels $T_A = aP_A + (1-a)\Pi$, balanced tags annihilate the Π term, giving signed operator norm

$a\rho_c$. For a bounded observable G retaining r_S states and r_{tag} suffix controls, the signed-coordinate estimate is

$$|\mathbb{E}_0[C_{\mathrm{tag}}G]| \le q^{r_S}(a\rho_c)^{H-1-r_{\mathrm{tag}}}\|G\|_\infty. \tag{H.17}$$

Insert diagonal projectors for retained states into the Markov product. Each unretained tagged control contributes an operator of norm $a\rho_c$, while a retained control has norm at most one. Summing retained state values costs at most q^{r_S}. The control U_1 supplies no signed contraction. At an invariant reference, the probability and label terms of each gradient coordinate retain at most three tagged controls and three states. The sum of squared categorical feature weights is at most V^2K. Summing over V output rows and taking a square root therefore gives

$$\|\nabla F_+(\Theta) - \nabla F_0(\Theta)\|_F \le \mu_F, \qquad \mu_F = 2q^3(a\rho_c)^{H-4}V^{3/2}\sqrt{K}. \tag{H.18}$$

This bound will be used only at the invariant ridge optimum.

Define $J_0(\Theta) = F_0(\Theta) + \lambda\|\Theta\|_F^2/2$. Its unique minimizer Θ_λ belongs to the invariant space above: auxiliary zero-start Euclidean GD with step at most $1/(K+\lambda)$ stays in that space by (H.10), ridge scales its coefficients, and strong convexity gives convergence to the unique optimum. This auxiliary sequence only characterizes the optimum; the source algorithm below is Adam.

The comparator yields

$$\begin{aligned} F_0(\Theta_\lambda) - H_0 &\le \varepsilon + \lambda D^*(\varepsilon)/2, \\ \|\Theta_\lambda\|_F^2 &\le D^*(\varepsilon) + 2\varepsilon/\lambda. \end{aligned} \tag{H.19}$$

Choose

$$\varepsilon = e/2, \qquad \overline{D} = \max\{D^*(e/2), e\}, \qquad \lambda = e/\overline{D} \le 1, \qquad R_\lambda = \sqrt{2\overline{D}}. \tag{H.20}$$

Then the excess is at most e and $\|\Theta_\lambda\|_F \le R_\lambda$. Class masses depend only on the legal class/PAD pattern. Equation (H.15) therefore gives $v_s \ge \exp(-Ne)$ on every legal-pattern prefix, including unseen words and incorrect state histories.

Let C be the realized hard-feature operator, $G_C = C^\top C$, and $\|G_C - I\|_{\mathrm{op}} \le \gamma < 1$. The fixed-feature physical head objective

$$J_C(W) = F_+(WC) + \lambda\|W\|_F^2/2 \tag{H.21}$$

is λ-strongly convex in the complete physical head, including its nullspace components. At $\widetilde{W} = \Theta_\lambda C^\top$,

$$\begin{aligned} \nabla J_C(\widetilde{W}) &= [\nabla F_+(\Theta_\lambda G_C) + \lambda\Theta_\lambda]C^\top, \\ r := \|\nabla J_C(\widetilde{W})\|_F &\le \|C\|_{\mathrm{op}}(\mu_F + K\gamma R_\lambda), \qquad \|W^* - \widetilde{W}\|_F \le r/\lambda, \end{aligned} \tag{H.22}$$

where W^* minimizes J_C. Thus, uniformly over all input strings,

$$\|(W^*C - \Theta_\lambda)\Phi(x)\|_\infty \le \sqrt{K}\left[\gamma R_\lambda + \frac{(1+\gamma)(\mu_F + K\gamma R_\lambda)}{\lambda}\right]. \tag{H.23}$$

The tag-conditioned source perturbation is paid once through strong convexity. A further uniform logit error h gives class mass at least $\exp(-Ne - 2h)$ and leakage at most $Ne + 2h$.

H.3.3 Descent with the actual adaptive denominator

We use an inexact first-moment estimate that allows the preconditioner to depend on all current samples. Let F be globally L-smooth and bounded below by F^*. Write $g_t = \nabla F(x_t)$, $z_t = g_t + \xi_t$, $\|\xi_t\| \le b$, and let $\widehat{m}_t(z)$ be the bias-corrected first moment with parameter $\beta \in [0,1)$ and zero initial moment. Suppose

$$x_{t+1} = x_t - \eta P_t \widehat{m}_t(z) + r_t, \qquad \|r_t\| \le \eta b_r, \qquad \lambda_p I \preceq P_t \preceq \Lambda_p I. \tag{H.24}$$

Set $e_{\mathrm{err}} = \Lambda_p b + b_r$ and $u = 1 - \beta$. If

$$\eta \le \frac{\lambda_p u^{3/2}}{8L\Lambda_p^2}, \tag{H.25}$$

then every finite prefix satisfies

$$F(x_{T+1}) + \frac{\eta\lambda_p}{2}\sum_{t=1}^{T}\|g_t\|^2 \le F(x_1) + \frac{3\eta T e_{\mathrm{err}}^2}{2\lambda_p}. \tag{H.26}$$

In particular, with $\Delta = F(x_1) - F^*$, $\sum_t \|g_t\|^2 \le 2\Delta/(\eta\lambda_p) + 3Te_{\mathrm{err}}^2/\lambda_p^2$.

For the proof, write $h_t = \widehat{m}_t(g)$ and absorb the corrected noise into an update error of norm at most e_{err}. Let $S_g = \sum_t \|g_t\|^2$ and $S_h = \sum_t \|h_t\|^2$. Corrected-weight Jensen gives $S_h \le S_g/u$, and the first-moment lag identity gives

$$\sqrt{\sum_t \|h_t - g_t\|^2} \le \frac{L\eta\Lambda_p\beta}{u^{3/2}}\sqrt{S_g} + \frac{L\eta e_{\mathrm{err}}\beta}{u}\sqrt{T}. \tag{H.27}$$

Smoothness and the squared-step bound imply

$$\begin{aligned} F(x_{T+1}) &\le F(x_1) - \frac{3\eta\lambda_p}{4}S_g + \eta D_m e_{\mathrm{err}}\sqrt{TS_g} + L\eta^2 T e_{\mathrm{err}}^2, \\ D_m &= 1 + L\eta\Lambda_p\beta/u \le 9/8. \end{aligned} \tag{H.28}$$

The lag and squared-step losses cost at most $\lambda_p(\beta + \sqrt{u})/8 \le \lambda_p/4$ in the descent coefficient. Young's inequality absorbs the mixed term using $D_m^2/\lambda_p + L\eta \le 89/(64\lambda_p) < 3/(2\lambda_p)$, proving (H.26).

For a target $r > 0$, the conditions

$$e_{\mathrm{err}} \le \lambda_p r/4, \qquad T \ge \max\left\{1, \left\lceil \frac{4\Delta}{\eta\lambda_p r^2} \right\rceil\right\} \tag{H.29}$$

ensure $\|g_t\| \le r$ at some inspected iterate. If $F \ge \mu\|x\|^2/2$ and $F(x_1) \le F_{\mathrm{init}}$, the squared radius of every prefix is at most $2F_{\mathrm{init}}/\mu + 3\eta T e_{\mathrm{err}}^2/(\mu\lambda_p)$. To close a stopped tube without a circular radius, choose a headroom $B_0 > 0$, radius squared $2(2F_{\mathrm{init}} + B_0)/\mu$, impose also $\eta \le B_0/(\lambda_p r^2)$, and use $T = \max\{1, \lceil 4F_{\mathrm{init}}/(\eta\lambda_p r^2)\rceil\}$. Then

$$\eta T\lambda_p r^2 \le 4F_{\mathrm{init}} + B_0, \qquad F(x_t) \le F_{\mathrm{init}} + 3F_{\mathrm{init}}/8 + 3B_0/32 < 2F_{\mathrm{init}} + B_0. \tag{H.30}$$

Bounds established inside this predetermined tube therefore close by first exit.

H.3.4 Ridge confinement of actual Adam coordinates

Suppose each head-gradient coordinate is $g_{t,j} = c_{t,j} + \lambda w_{t,j}$ with $|c_{t,j}| \le C_h/\sqrt{d}$. Use physical step $\eta_s/\sqrt{d}$, $\eta_s \le 1$, and physical denominator offset $\epsilon_0/\sqrt{d}$. Let the moments start at zero and put

$$K_{\mathrm{mom}} = \left[(1-\beta_2)(1-\beta_1^2/\beta_2)\right]^{-1/2}, \qquad M_c = \max\left\{M_0, \frac{C_h}{\lambda} + \frac{\eta_s K_{\mathrm{mom}}}{1-\beta_1}\right\}. \tag{H.31}$$

If $|w_{1,j}| \le M_0/\sqrt{d}$, then $|w_{t,j}| \le M_c/\sqrt{d}$ at every finite step.

Indeed, weighted Cauchy–Schwarz for the complete corrected histories gives $|\widehat{m}_{t,j}| \le K_{\mathrm{mom}}\sqrt{\widehat{v}_{t,j}}$, so each coordinate step is at most $\delta_w = \eta_s K_{\mathrm{mom}}/\sqrt{d}$. The corrected first-moment weights have mean age

$$\sum_s a_{ts}(t-s) = \frac{\beta_1}{1-\beta_1} - \frac{t\beta_1^t}{1-\beta_1^t} \le \frac{\beta_1}{1-\beta_1}. \tag{H.32}$$

Consequently,

$$|\widehat{m}_{t,j} - \lambda w_{t,j}| \le C_h/\sqrt{d} + \lambda\delta_w\beta_1/(1-\beta_1). \tag{H.33}$$

Outside the interval of radius $[C_h/\lambda + \eta_s K_{\mathrm{mom}}\beta_1/(1-\beta_1)]/\sqrt{d}$, momentum has the coordinate's sign, so the step points inward. Inside it, one step stays within $M_c/\sqrt{d}$; an overshoot across zero is at most the step bound. This proves confinement by induction.

Define

$$G_c = C_h + \lambda\max\left\{M_0, C_h/\lambda + K_{\mathrm{mom}}/(1-\beta_1)\right\}. \tag{H.34}$$

Then $|g_{t,j}| \le G_c/\sqrt{d}$, and the actual update is exactly

$$W_{t+1} = W_t - \eta_s P_t \widehat{m}_t, \quad P_t = \operatorname{diag}_j \frac{1}{\sqrt{d\widehat{v}_{t,j}} + \epsilon_0}, \quad \frac{I}{G_c + \epsilon_0} \preceq P_t \preceq \frac{I}{\epsilon_0}. \tag{H.35}$$

The ridge coefficient is constant throughout the moment history. These are spectral bounds on the realized denominator for coupled regularization.

H.3.5 Finite minibatches and the observable stop

We now use $\beta_1 = 0.9$, $\beta_2 = 0.999$, $u = 1 - \beta_1$, and $\epsilon_0 = 1$. Every supplied parameter, including the backbone and any normalization gain, is optimized with the complete source loss plus $\lambda/2$ times the squared norm of all supplied parameters. In the fixed-head comparison (H.21), the constant backbone penalty is omitted; it remains in the actual backbone gradients. All moments start at zero and remain carried throughout source training.

Choose $h \in (0, 1)$ and

$$0 < e \le \min\left\{h/N, \frac{c^2(q-1)^2}{32q^2H^2N}\right\}, \qquad \gamma = d^{-1/8} \le 1/8, \tag{H.36}$$

and use (H.20). Require the realized hard Gram event and

$$\|C^\top C - I\|_{\mathrm{op}} \le \gamma, \qquad \mu_F \le \frac{\lambda h}{64\sqrt{K}}, \qquad \gamma \le \frac{\lambda h}{64\sqrt{K}K(R_\lambda + 1)}. \tag{H.37}$$

Then (H.23) is less than $h/8$.

Here the source feature envelope is $A_s = 12$, distinct from the canonical norm $\sqrt{K}$. For an initialization failure allocation δ_0, put

$$\begin{aligned} M_0 &= \sqrt{2\log(2Vd/\delta_0)}, & W_{\max} &= \sqrt{V}M_0, \\ C_{\mathrm{row}} &= 2\sqrt{2k + 4\log(2d/\delta_0)}, & C_h &= 2\sqrt{K}C_{\mathrm{row}} + 3, \end{aligned} \tag{H.38}$$

where k is the canonical feature dimension. The row and readout events are $\max_i \sqrt{d}\|C_i\| \le C_{\mathrm{row}}$ and $\max_{y,i} |W_{0,yi}| \le M_0/\sqrt{d}$. Use (H.34) and define

$$\lambda_p = (G_c + 1)^{-1}, \qquad \Lambda_p = 1, \qquad L = 2K + 1,$$

$$\begin{aligned} F_{\mathrm{init}} &= \log V + 2A_sW_{\max} + \lambda W_{\max}^2/2, & R &= \sqrt{2(2F_{\mathrm{init}} + 1)/\lambda}, \\ \rho &= \lambda h/(64\sqrt{K}), & r_{\mathrm{goal}} &= \rho/4, \\ \eta_s &= \min\{1, \lambda_p u^{3/2}/(8L), 1/(\lambda_p r_{\mathrm{goal}}^2)\}, & b &= \lambda_p\rho/32, \\ T_s &= \max\left\{1, \left\lceil \frac{4F_{\mathrm{init}}}{\eta_s\lambda_p r_{\mathrm{goal}}^2} \right\rceil\right\}, \\ \varepsilon_{\mathrm{feat}} &= \min\{1, b/[2(A_sR + \sqrt{2})], h/(8R)\}. \end{aligned} \tag{H.39}$$

In addition require

$$\gamma \le \varepsilon_{\mathrm{feat}}/16. \tag{H.40}$$

Choose the initial raw soft/hard error using the finite temperature construction in Section G to be at most

$$z_{\mathrm{raw}} = \min\{\varepsilon_{\mathrm{feat}}/(8\sqrt{2}), 1/(4\sqrt{d})\}. \tag{H.41}$$

The normalized member then has initial discrepancy at most $\sqrt{2}z_{\mathrm{raw}} + 4\gamma \le 3\varepsilon_{\mathrm{feat}}/8$; the base member has a smaller bound. Both initial coordinate envelopes are at most $(C_h - 1)/\sqrt{d}$.

Let D be the complete feature-derivative bound on the radius-one backbone ball and let P_ψ count all supplied backbone coordinates. Set

$$r_\psi = \min\{1, 1/D, \varepsilon_{\mathrm{feat}}/(2D), 1/(D\sqrt{d})\}, \qquad \rho_{\mathrm{back}} = \min\left\{1, \frac{r_\psi\sqrt{d}}{2\eta_s T_s K_{\mathrm{mom}}\sqrt{P_\psi}}\right\} > 0. \tag{H.42}$$

Every backbone coordinate uses physical step $\eta_s\rho_{\mathrm{back}}/\sqrt{d}$. The Adam coordinate-step bound gives total displacement at most $r_\psi/2$, independently of the batches and including the coupled regularizer gradients. Thus all feature coordinates remain bounded by $C_h/\sqrt{d}$ and the full-feature discrepancy stays at most $\varepsilon_{\mathrm{feat}}$. This also includes the normalization gain when present.

With $P_{\mathrm{head}} = Vd$, choose

$$B_s = \left\lceil 8P_{\mathrm{head}}A_s^2 b^{-2}\log\frac{2P_{\mathrm{head}}T_s}{\delta_{\mathrm{batch}}}\right\rceil. \tag{H.43}$$

At each current source iterate, draw B_s fresh complete records and compute the averaged all-target gradient, including the coupled regularizer. If the head gradient's Frobenius norm is at most $\rho/2$, stop before any update. Otherwise update all blocks using actual bias-corrected Adam, unless this was the T_sth check; a failed final check returns failure without another update. A stopping batch is not inserted into a later moment. This algorithm uses at most T_s batches and $T_s - 1$ updates on every outcome. Its selection statistic uses only source observations.

Explicitly, for each performed update and every supplied coordinate, its actual minibatch gradient g_t enters

$$m_t = \beta_1 m_{t-1} + (1-\beta_1)g_t, \quad v_t = \beta_2 v_{t-1} + (1-\beta_2)g_t^2, \quad \widehat{m}_t = \frac{m_t}{1-\beta_1^t}, \quad \widehat{v}_t = \frac{v_t}{1-\beta_2^t}. \tag{H.44}$$

The coordinate subtracts its prescribed physical step times $\widehat{m}_t/(\sqrt{\widehat{v}_t} + 1/\sqrt{d})$. All blocks update simultaneously, with their strictly positive head or backbone rates.

Within $\|W\|_F \le R$, the difference between actual-feature and hard-reference population head gradients is at most $(A_sR + \sqrt{2})\varepsilon_{\mathrm{feat}} \le b/2$. Each coordinate of a complete-record CE gradient lies in $[-A_s, A_s]$. Conditional scalar Hoeffding and a union bound over the P_{head} coordinates and T_s predictable inspections give a batch/mean Frobenius error at most $b/2$, with failure at most δ_{batch}. Thus every inspected actual gradient differs from $\nabla J_C(W)$ by at most b inside the stopped tube.

Coordinate confinement makes the effective preconditioner in (H.35) lie between $\lambda_p I$ and I. Apply (H.26) with $\eta = \eta_s$, noise bound b, no additive update error, and target r_{goal}. The tube construction uses $\mu = \lambda$ and headroom one. If the final check fails, extend its already computed gradient by one counterfactual update only in this argument. The bound then gives some inspected gradient norm at most r_{goal}, while every prefix loss remains below $2F_{\mathrm{init}} + 1$. First exit closes $\|W\|_F < R$. Since $b \le \rho/32$, such a reference gradient necessarily triggers the observable threshold $\rho/2$. Hence failure cannot occur on the good event. Conversely, a triggering gradient gives $\|\nabla J_C(W)\|_F \le \rho/2 + b < \rho$.

Strong convexity now yields

$$\|W - W^*\|_F \le \rho/\lambda, \qquad \sqrt{K}\|C\|_{\mathrm{op}}\rho/\lambda < h/32, \qquad R\varepsilon_{\mathrm{feat}} \le h/8. \tag{H.45}$$

Together with (H.37), these bounds give uniform source-logit discrepancy less than $h/2$ from $\Theta_\lambda\Phi$, including unseen legal-pattern words and incorrect generated state histories. The canonical class masses are at least $\exp(-Ne)$ and $|b_j - a| \le c/(8H)$. The actual source policy inherits the corresponding explicit leakage and transition perturbation bounds, and its initial control choice is close to uniform because the canonical control rows agree. Its worst-case consumption is B_sT_s fresh records, NB_sT_s raw next-token targets, at most $T_s - 1$ updates, and T_s gradient-norm checks, with zero verifier calls.

H.3.6 Explicit source width and row envelopes

Write $t_d = 1 + \log d$ and define quantities independent of d:

$$\begin{aligned}
&\overline{m} = \sqrt{2\max\{1, \log(2V/\delta_0)\}}, \\
&\overline{C} = 2\sqrt{2k + 4\log(2/\delta_0) + 4}, \qquad \overline{c}_h = 2\sqrt{K}\,\overline{C} + 3, \\
&\overline{g} = 1 + \overline{c}_h + \max\{\lambda\overline{m}, \overline{c}_h + \lambda K_{\mathrm{mom}}/u\}, \\
&\overline{f} = \log V + 2A_s\sqrt{V}\,\overline{m} + (\lambda V/2)\overline{m}^2, \qquad \overline{r} = \sqrt{2(2\overline{f} + 1)/\lambda}.
\end{aligned} \tag{H.46}$$

Then $M_0 \le \overline{m}\sqrt{t_d}$, $C_h \le \overline{c}_h\sqrt{t_d}$, $G_c + 1 \le \overline{g}\sqrt{t_d}$, $F_{\mathrm{init}} \le \overline{f}t_d$, and $R \le \overline{r}\sqrt{t_d}$. Put

$$B_{\mathrm{dim}} = \max\{16, 1024\overline{g}(A_s\overline{r} + \sqrt{2})/\rho, 128\overline{r}/h\}. \tag{H.47}$$

The prescribed feature tolerance satisfies $\varepsilon_{\mathrm{feat}} \geq 16/[B_{\dim}(1+\log d)]$. Since $\log d \leq 16d^{1/16}/\exp(1)$, the explicit requirements

$$d \geq [B_{\dim}(1+16/\exp(1))]^{16}, \qquad d \geq \left[\frac{64\sqrt{K}K(R_\lambda+1)}{\lambda h}\right]^8 \tag{H.48}$$

imply (H.40) and the Gram-accuracy inequality in (H.37). An initializer requirement $d \geq D_\Gamma \gamma^{-2}$ becomes $d \geq D_\Gamma^{4/3}$ under $\gamma = d^{-1/8}$. The feature and head widths can then be instantiated without self-reference. The separate signed-source inequality in (H.37), and any feedback-driven tightening of $\varepsilon_{\mathrm{feat}}$, are handled in Section H.5.

For completeness, the row event used above follows from left invariance conditional on coverage and the actual Gram event. In distribution,

$$C = \frac{Z}{\sqrt{d}} H_Z^{-1/2} R_C, \qquad H_Z = Z^\top Z/d, \qquad R_C = (C^\top C)^{1/2}, \tag{H.49}$$

where Z has independent standard Gaussian entries and is independent of R_C. If

$$d \geq 128[k\log 9 + \log(4/\delta_0)], \tag{H.50}$$

then $\|H_Z - I\|_{\mathrm{op}} \leq 1/2$ except with probability $\delta_0/2$. On this event, $\|H_Z^{-1/2} R_C\|_{\mathrm{op}} \leq 3/2 < 2$ because $\gamma \leq 1/8$. The Gaussian moment-generating function at $1/4$ gives

$$\Pr\{\|z_i\|^2 > 2k + 4\log(2d/\delta_0)\} \leq \delta_0/(2d). \tag{H.51}$$

A row union bound therefore supplies $\max_i \sqrt{d}\|C_i\| \leq C_{\mathrm{row}}$ with conditional failure at most δ_0. This does not assert independence of the realized feature rows. The readout-coordinate event has a separate failure allocation δ_0; coverage, full-Gram, parameter-envelope, and source-batch failures are added in Section H.5.

H.4 Two sampled Adam updates after source training

The source model constructed in Appendix H.3 supplies a uniform approximation to a canonical executor. We show that two complete, on-policy feedback batches identify the downstream binding while every supplied parameter continues to train. The argument first controls the actual Adam histories, then compares the correlated finite feature operator with a Gaussian operator uniformly over adaptive histories, and finally couples complete generated trajectories. Appendix H.5 selects the dimensions and precisions simultaneously and combines the probability events.

Throughout this appendix, k denotes the canonical feature dimension, $K = 12$ is its squared feature norm, and $A = \sqrt{K}$. The hard feature map from Appendix G is $h_{\mathrm{hard}}(x) = C\Phi(x)$, with $C \in \mathbb{R}^{d\times k}$ and $\|\Phi(x)\|^2 = K$. We write $\|\cdot\|_F$ for the Frobenius norm. The source and feedback stages use $\beta_1 = 0.9$ and $\beta_2 = 0.999$. At feedback entry the parameters are retained, the source regularizer is removed, and all first and second moments are reset. Feedback ascent uses the complete terminal-reward score, baseline zero, and the original batch denominator.

H.4.1 Complete bias-corrected histories

We record a bound that remains valid when the gradients and denominators are dependent. For a stage-local history $g_1, \ldots, g_t$, initialized with zero moments, define

$$a_{t,s} = \frac{(1-\beta_1)\beta_1^{t-s}}{1-\beta_1^t}, \qquad b_{t,s} = \frac{(1-\beta_2)\beta_2^{t-s}}{1-\beta_2^t}, \tag{H.52}$$

$$\widehat{m}_t(g) = \sum_{s=1}^t a_{t,s} g_s, \qquad \widehat{v}_t(g) = \sum_{s=1}^t b_{t,s} g_s^{\odot 2}. \tag{H.53}$$

Both weight sequences sum to one. For $0 \leq \beta_1 < 1$, $0 < \beta_2 < 1$, and $\beta_1^2 < \beta_2$, put

$$K_{\mathrm{mom}} = \left[(1-\beta_2)(1-\beta_1^2/\beta_2)\right]^{-1/2}. \tag{H.54}$$

Lemma H.2 (Magnitude and history perturbation)**.** *For every coordinate and every finite history, $|\widehat{m}_t(g)| \le K_{\mathrm{mom}}\sqrt{\widehat{v}_t(g)}$. For $\epsilon > 0$, let $\mathcal{D}_t(g) = \widehat{m}_t(g)/(\sqrt{\widehat{v}_t(g)} + \epsilon)$, coordinatewise. Two histories with the same zero initialization satisfy*

$$\|\mathcal{D}_t(g) - \mathcal{D}_t(h)\|_\infty \le \frac{1 + K_{\mathrm{mom}}}{\epsilon} \max_{s \le t} \|g_s - h_s\|_\infty. \tag{H.55}$$

In particular, every Adam coordinate direction has magnitude at most K_{mom}, without an upper gradient bound or a positive lower second moment.

Proof. For one coordinate, weighted Cauchy–Schwarz gives

$$|\widehat{m}_t(g)|^2 \le K_t^2 \widehat{v}_t(g), \tag{H.56}$$

$$K_t^2 = \sum_{s=1}^{t} \frac{a_{t,s}^2}{b_{t,s}} = \frac{(1-\beta_1)^2}{(1-\beta_1^t)^2} \frac{1-\beta_2^t}{1-\beta_2} \sum_{j=0}^{t-1} (\beta_1^2/\beta_2)^j \le K_{\mathrm{mom}}^2. \tag{H.57}$$

Here $1-\beta_1^t \ge 1-\beta_1$; when $\beta_1 = 0$, the first-moment weight is concentrated at the newest gradient and the zeroth power is one. Write $r_g = (\sum_s b_{t,s} g_s^2)^{1/2}$ and define r_h similarly. If $d_t = \max_{s \le t} |g_s - h_s|$, then

$$|\widehat{m}_t(g) - \widehat{m}_t(h)| \le d_t, \qquad |r_g - r_h| \le d_t$$

by the reverse triangle inequality in the weighted Euclidean norm. The identity

$$\frac{m_g}{r_g + \epsilon} - \frac{m_h}{r_h + \epsilon} = \frac{m_g - m_h}{r_g + \epsilon} + \frac{m_h}{r_h + \epsilon} \frac{r_h - r_g}{r_g + \epsilon}$$

now proves (H.55), including coordinates with zero second moment. □

The bound uses the complete history since the common zero initialization. Source moments are retained during source training and reset once at the stage boundary; both feedback histories enter the second update.

H.4.2 The finite operator with adaptive histories

The finite coefficient rows are correlated. Their relevant symmetry follows from the complete initialization in Appendix G.

Lemma H.3 (Initial left invariance)**.** *Conditional on any relative-bias realization satisfying lag coverage, the base member's actual hard coefficient operator obeys $C \overset{\text{law}}{=} QC$ for every deterministic $Q \in O(d)$. Its canonical feature coordinates and lag weights are unchanged by this transformation.*

Proof. Distinguish the FFN matrices $B_{\mathrm{ff}}, C_{\mathrm{ff}}, U_{\mathrm{ff}}$ from the coefficient operator. Apply the deterministic transformation

$$\begin{aligned}
&E \mapsto QE, && O_h \mapsto QO_h, && \\
&Q_h \mapsto Q_h Q^\top, && K_h \mapsto K_h Q^\top, && V_h \mapsto V_h Q^\top, \\
&B_{\mathrm{ff}} \mapsto B_{\mathrm{ff}} Q^\top, && C_{\mathrm{ff}} \mapsto C_{\mathrm{ff}} Q^\top, && U_{\mathrm{ff}} \mapsto QU_{\mathrm{ff}}, \\
&b_{\mathrm{out}} \mapsto Qb_{\mathrm{out}}, && W_{\mathrm{out}} \mapsto W_{\mathrm{out}} Q^\top. &&
\end{aligned}$$

Leave the relative biases and FFN input biases unchanged. Each Gaussian block has its original law under its indicated left or right orthogonal transform; the transformations also preserve joint independence. This remains true conditional on the unchanged relative biases.

The query, key, and value vectors are unchanged, so all soft attention scores and weights are unchanged. The residual feature transforms as $z \mapsto Qz$; the FFN preactivations are unchanged and its output transforms by Q. For the hard reference, the lag blocks satisfy $A'_\ell = QA_\ell$ and $T' = QT$, since the lag-weight matrix is unchanged and invertible on coverage. The augmented FFN coefficient rows $[B_{\mathrm{ff}}T, b_{\mathrm{ff}}]$ and $[C_{\mathrm{ff}}T, c_{\mathrm{ff}}]$ are unchanged, hence so is their symmetric-product coefficient matrix F. The actual expansion therefore gives

$$C' = [QT, QU_{\mathrm{ff}}F, Qb_{\mathrm{out}}] = QC, \qquad (C')^\top C' = C^\top C.$$

This proves the identity for the coefficient columns themselves, including their constant and linear terms, rather than only on sampled inputs. □

This lemma concerns the raw initial operator. For the normalized member the same operator is used as a reference, and the normalization and trainable gain are paid through the feature approximation. No trajectory equivariance of coordinatewise Adam is required.

For rows $c_i^\top$ of C, define the finite and Gaussian operators

$$\mathcal{A}_C(S) = \sum_{i=1}^{d} \frac{c_i c_i^\top}{\sqrt{d\, c_i^\top S c_i} + \epsilon}, \qquad \mathcal{A}(S) = \mathbb{E}_{z\sim N(0,I_k)} \frac{zz^\top}{\sqrt{z^\top S z} + \epsilon}. \tag{H.58}$$

The next estimate is uniform over the second-moment matrices, which permits their dependence on the realized features and training batches.

Lemma H.4 (Uniform correlated adaptive geometry)**.** *Let $k \geq 1$, $d \geq k$, $G \geq 0$, $\epsilon > 0$, $0 < \tau \leq 1$, and $0 < \delta_{\mathrm{geo}} < 1$. Define*

$$K_3 = \frac{8[k(k+2)]^{3/4}}{\delta_{\mathrm{geo}}}, \qquad L_{\mathrm{pert}} = \frac{6k}{\epsilon} + \frac{GK_3}{\epsilon^2}, \tag{H.59}$$

$$\gamma_{\mathrm{geo}} = \min\left\{\frac{1}{8}, \frac{\tau}{16 L_{\mathrm{pert}}}\right\}, \qquad a_{\mathrm{geo}} = \frac{\tau}{2}, \tag{H.60}$$

$$T_{\mathrm{tr}}^2 = \max\left\{1, \frac{64k(k+2)}{\epsilon a_{\mathrm{geo}} \delta_{\mathrm{geo}}}\right\}, \qquad h_{\mathrm{net}} = \left(\frac{a_{\mathrm{geo}}\epsilon^2}{16 T_{\mathrm{tr}}^3}\right)^2, \tag{H.61}$$

$$D_{\mathrm{sym}} = \frac{k(k+1)}{2}, \qquad N_{\mathrm{net}} = \left(1 + \frac{2\sqrt{k}G^2}{h_{\mathrm{net}}}\right)^{D_{\mathrm{sym}}}. \tag{H.62}$$

Suppose C has a left-orthogonally invariant law and the event $\mathcal{E}_C = \{\|C^\top C - I\|_{\mathrm{op}} \leq \gamma_{\mathrm{geo}}\}$ has positive probability. Set

$$d_{\mathrm{geo}} = \left\lceil \max\left\{ 32\gamma_{\mathrm{geo}}^{-2}\left[k\log 9 + \log\frac{16}{\delta_{\mathrm{geo}}}\right], \frac{32k^2 T_{\mathrm{tr}}^4}{\epsilon^2 a_{\mathrm{geo}}^2} \log\frac{16k^2 N_{\mathrm{net}}}{\delta_{\mathrm{geo}}} \right\} \right\rceil. \tag{H.63}$$

If $d \geq d_{\mathrm{geo}}$, then, conditional on $\mathcal{E}_C$, with probability at least $1 - \delta_{\mathrm{geo}}$,

$$\sup_{S\succeq 0,\, \|S\|_{\mathrm{op}} \leq G^2} \|\mathcal{A}_C(S) - \mathcal{A}(S)\|_{\mathrm{op}} \leq \tau. \tag{H.64}$$

The unconditional failure probability is at most $\Pr(\mathcal{E}_C^c) + \delta_{\mathrm{geo}}$.

Proof. On $\mathcal{E}_C$, let $R = (C^\top C)^{1/2}$, which is positive definite and satisfies $\|R - I\|_{\mathrm{op}} \leq \gamma_{\mathrm{geo}}$. Conditional on R, left invariance makes CR^{-1} a uniform Stiefel matrix. Averaging an independent uniform left rotation identifies this conditional law and its independence from R. Thus, conditional on $\mathcal{E}_C$, we may represent

$$C = d^{-1/2} Z M, \qquad M = H_Z^{-1/2} R, \qquad H_Z = Z^\top Z/d, \tag{H.65}$$

where Z has independent standard Gaussian entries and is independent of R. This representation does not assert independence of the rows of C.

For a fixed unit vector v, the chi-square moment generating function gives

$$\Pr\big(|v^\top (H_Z - I) v| > b\big) \leq 2\exp(-db^2/8), \qquad 0 < b < 1.$$

A $1/4$-net of size at most 9^k and the symmetric-matrix net inequality give

$$\Pr\big(\|H_Z - I\|_{\mathrm{op}} > \gamma_{\mathrm{geo}}\big) \leq 2\exp\{k\log 9 - d\gamma_{\mathrm{geo}}^2/32\} \leq \delta_{\mathrm{geo}}/8.$$

On the complementary event,

$$\begin{aligned} \|H_Z^{-1/2} - I\|_{\mathrm{op}} &\leq 2\gamma_{\mathrm{geo}}, & \|R\|_{\mathrm{op}} &\leq 1 + \gamma_{\mathrm{geo}}, \\ \|M - I\|_{\mathrm{op}} &\leq 4\gamma_{\mathrm{geo}} \leq \tfrac{1}{2}, & d^{-1}\sum_i \|z_i\|^2 &\leq 2k. \end{aligned} \tag{H.66}$$

Since $\mathbb{E}\|z\|^3 \le [k(k+2)]^{3/4}$, Markov's inequality also gives $d^{-1}\sum_i \|z_i\|^3 \le K_3$ except with probability $\delta_{\mathrm{geo}}/8$.

Write $\mathcal{A}_Z(S) = d^{-1}\sum_i z_i z_i^\top/(\sqrt{z_i^\top S z_i} + \epsilon)$. Superscript tr denotes truncation to $\|z\| \le T_{\mathrm{tr}}$. Uniformly over S,

$$\|\mathcal{A}(S) - \mathcal{A}^{\mathrm{tr}}(S)\|_{\mathrm{op}} \le \frac{\mathbb{E}[\|z\|^2 \mathbf{1}_{\{\|z\|>T_{\mathrm{tr}}\}}]}{\epsilon} \le \frac{k(k+2)}{\epsilon T_{\mathrm{tr}}^2} \le \frac{a_{\mathrm{geo}}\delta_{\mathrm{geo}}}{64}. \tag{H.67}$$

The corresponding empirical tail is at most $a_{\mathrm{geo}}/8$ except with probability $\delta_{\mathrm{geo}}/8$, by Markov's inequality.

The positive semidefinite (PSD) matrices of operator norm at most G^2 lie in a Frobenius ball of radius $\sqrt{k}G^2$ in dimension D_{sym}. A packing argument supplies an h_{net}-net, with centers in that set, of size at most N_{net}. Each truncated matrix entry has magnitude at most $T_{\mathrm{tr}}^2/\epsilon$. Scalar Hoeffding at entry tolerance $a_{\mathrm{geo}}/(4k)$, unioned over the net and k^2 entries, has failure probability at most

$$2k^2 N_{\mathrm{net}} \exp\left\{-\frac{d a_{\mathrm{geo}}^2 \epsilon^2}{32 k^2 T_{\mathrm{tr}}^4}\right\} \le \delta_{\mathrm{geo}}/8. \tag{H.68}$$

At every net point the operator error is at most $a_{\mathrm{geo}}/4$, since an entrywise bound b gives operator norm at most kb. For two PSD matrices,

$$|\sqrt{z^\top S z} - \sqrt{z^\top S' z}| \le \|z\|\sqrt{\|S - S'\|_{\mathrm{op}}}.$$

Consequently each truncated kernel changes by at most $T_{\mathrm{tr}}^3\sqrt{h_{\mathrm{net}}}/\epsilon^2$ within a net ball. The empirical and population extension costs together are at most $2T_{\mathrm{tr}}^3\sqrt{h_{\mathrm{net}}}/\epsilon^2 = a_{\mathrm{geo}}/8$. Combining both tails, the net bound, and its extension gives

$$\sup_S \|\mathcal{A}_Z(S) - \mathcal{A}(S)\|_{\mathrm{op}} \le \frac{a_{\mathrm{geo}}}{8} + \frac{a_{\mathrm{geo}}}{64} + \frac{a_{\mathrm{geo}}}{4} + \frac{a_{\mathrm{geo}}}{8} = \frac{33 a_{\mathrm{geo}}}{64} < a_{\mathrm{geo}}. \tag{H.69}$$

It remains to transfer through M. For $\mathcal{K}_S(x) = xx^\top/(\sqrt{x^\top S x} + \epsilon)$ and $\kappa = \|M - I\|_{\mathrm{op}} \le 1/2$, direct subtraction gives

$$\|\mathcal{K}_S(M^\top y) - \mathcal{K}_S(y)\|_{\mathrm{op}} \le \kappa\left[\frac{(2+\kappa)\|y\|^2}{\epsilon} + \frac{G\|y\|^3}{\epsilon^2}\right]. \tag{H.70}$$

Here the numerator difference is bounded by $(\|M^\top y\| + \|y\|)\|(M^\top - I)y\|$; both denominators are at least ϵ, and their difference is at most $G\|(M^\top - I)y\|$. Averaging (H.70) and using the second- and third-moment bounds gives

$$\sup_S \|\mathcal{A}_C(S) - \mathcal{A}_Z(S)\|_{\mathrm{op}} \le 4\gamma_{\mathrm{geo}} L_{\mathrm{pert}} \le \tau/4.$$

Together with (H.69), this is at most $3\tau/4 \le \tau$. The Gaussian auxiliary failures total at most $\delta_{\mathrm{geo}}/2$, uniformly over the admissible R; integrating its conditional law proves the lemma. □

For complete coefficient histories u_s, set

$$m_t = \sum_s a_{t,s} u_s, \qquad S_t = \sum_s b_{t,s} u_s u_s^\top. \tag{H.71}$$

If $\|u_s\| \le G$, then $\|m_t\| \le G$ and $\|S_t\|_{\mathrm{op}} \le G^2$. Equation (H.64) therefore applies to histories chosen using C and any training information. For histories in a subspace H_0, Gaussian reflection gives $P_{H_0^\perp}\mathcal{A}(S_t)m_t = 0$ and hence $\|P_{H_0^\perp}\mathcal{A}_C(S_t)m_t\| \le \tau G$. The same Gaussian operator satisfies

$$\mathcal{A}(S) \succeq \frac{I}{2(G\sqrt{2(k+2)} + \epsilon)}.$$

Indeed, retain $\|z\| \le \sqrt{2(k+2)}$ and use $\mathbb{E}[(v^\top z)^2\|z\|^2] = k+2$ for each unit v. This lower bound concerns action on the momentum vector; current gradients need not align with previous momentum.

For fixed features, a physical gradient Cu_s and coefficient coordinate $\vartheta = C^\top w$ yield the exact ascent identity

$$\epsilon_{\mathrm{physical}} = \epsilon/\sqrt{d}, \qquad \eta_{\mathrm{physical}} = \eta/\sqrt{d} \quad \Longrightarrow \quad \Delta\vartheta = \eta\mathcal{A}_C(S_t)m_t. \tag{H.72}$$

The current denominator remains inside $\mathcal{A}_C$.

H.4.3 THE LEARNED EXECUTOR AND THE TWO-UPDATE REFERENCE

Let $q \geq 16$, $c = 1/64$, and let the source conditional transition parameters satisfy $|b_j - (1 - c/H)| \leq c/(8H)$ for $j = 1, 2$, as proved in Appendix H.3. Define

$$B_{\text{exec}} = b_1 b_2^{H-1}, \qquad d_j = b_j + 2(1 - b_j)/q, \qquad D_{\text{exec}} = d_1 d_2^{H-1}. \tag{H.73}$$

At a valid transition of type j, the correct and hidden-bit-flipped states have probabilities $\alpha_j = b_j + (1 - b_j)/q$ and $\gamma_j = (1 - b_j)/q$. Their sum is d_j and difference is b_j. Summing all even and odd hidden-error histories gives, conditional on the choice and correct output classes,

$$r_* = \frac{D_{\text{exec}} + B_{\text{exec}}}{2}, \qquad \bar{r} = \frac{D_{\text{exec}} - B_{\text{exec}}}{2}, \qquad r_a = 0 \quad \text{for the other public arms.} \tag{H.74}$$

Here r_* is the reward of the bound operation and $\bar{r}$ that of its hidden-bit opposite. Other public operations cannot reach the required public endpoint while satisfying the grammar and public-coordinate checks. Repaired hidden-bit histories are included in (H.74). Conditional on the correct output classes, the public-coordinate checks pass with probability D_{exec}, independently of the selected control, and

$$r_* \geq B_{\text{exec}} \geq 1 - 9c/8, \qquad 0 \leq \bar{r} \leq \sum_j \gamma_j \leq \frac{9c}{8q}, \qquad 1 - D_{\text{exec}} \leq 9c/8. \tag{H.75}$$

The sum over j counts one first transition and $H - 1$ repeated transitions. Thus $\Delta = r_* - \bar{r} \geq 1/2$ and $\bar{r} < r_*/q$.

Let μ be the mean canonical feature of a training choice prompt. A degree-two coordinate retains at most two controls, so $H \geq 5$ leaves an unused balanced suffix tag. Its expectation is therefore the same in either tag class. Against each fixed legal choice prompt, averaging the kernel's matching indicators also gives the same constant. Consequently

$$\mu^\top \Phi(x) = \|\mu\|^2 =: \kappa \in [2, K] \quad \text{for every training and held-out choice prompt.} \tag{H.76}$$

This identity keeps the reference choice probabilities prompt-independent.

The Gaussian reference starts with zero coefficient increments $Y_0 = 0$ and uniform choice probabilities. Its row gradient is

$$u_{t,a} = c_{t,a}\mu, \qquad c_{t,a} = p_{t-1,a}(r_a - J_{t-1}), \qquad J_{t-1} = \sum_a p_{t-1,a} r_a. \tag{H.77}$$

Each row uses its own corrected moments and update $\eta_t \mathcal{A}(S_{t,a}) m_{t,a}$. Set

$$\epsilon_f = \frac{1}{64q}, \qquad \rho_0 = \frac{2e^{-2}}{\sqrt{2\pi}}, \qquad C_0 = \frac{2(1 - \beta_1)\rho_0}{64\sqrt{2} + 1}, \tag{H.78}$$

$$\eta_1 = \frac{1}{100 K_{\text{mom}} \sqrt{K}}, \qquad \eta_2 = \frac{\log(15(q-1))}{C_0}, \qquad S_{\text{step}} = \eta_1 + \eta_2. \tag{H.79}$$

Lemma H.5 (Two Gaussian reference updates)**.** *With (H.79), the reference chooses the bound operation with probability at least* $15/16$ *after the second update, on every training and held-out choice prompt. Its total Frobenius path length is at most*

$$R_{\text{ref}} = S_{\text{step}} C_{\text{move}}, \qquad C_{\text{move}} = K_{\text{mom}} + 32\sqrt{K}. \tag{H.80}$$

Proof. For a row history, write $m_{t,a} = \bar{c}_{t,a}\mu$ and $S_{t,a} = v_{t,a}\mu\mu^\top$. Gaussian symmetry gives

$$\mathcal{A}(v\mu\mu^\top)\mu = \mu\mathcal{B}(v), \qquad \mathcal{B}(v) = \mathbb{E}_{Z \sim N(0,1)} \frac{Z^2}{\sqrt{\kappa v}|Z| + \epsilon_f} > 0. \tag{H.81}$$

Every prompt logit increment is $\eta_t \kappa \bar{c}_{t,a} \mathcal{B}(v_{t,a})$. Whenever $p_{t-1,*} \geq 1/q$, $J_{t-1} \geq r_*/q > \bar{r}$: the correct row has positive gradient and momentum, and every wrong row has negative gradient and momentum. Thus the correct probability is nondecreasing, and prompt independence follows from (H.76).

Before $p_{t-1,*}$ reaches $15/16$,

$$c_{t,*} \geq \Delta p_{t-1,*}(1 - p_{t-1,*}) \geq \Delta/(16q), \tag{H.82}$$

$$c_{s,*} \leq r_* p_{s-1,*} \leq r_* p_{t-1,*} \leq (16 r_*/\Delta) c_{t,*} \leq 32 c_{t,*} \qquad (s \leq t). \tag{H.83}$$

The newest first-moment weight is at least $1 - \beta_1$, so $\bar{c}_{t,*} \geq (1 - \beta_1) c_{t,*}$ and $\sqrt{v_{t,*}} \leq 32 c_{t,*}$. The event $1 \leq |Z| \leq 2$ has probability at least ρ_0. Therefore

$$\mathcal{B}(v_{t,*}) \geq \frac{\rho_0}{2\sqrt{\kappa v_{t,*}} + \epsilon_f}.$$

Writing $\chi = 16 q \epsilon_f / \Delta \leq 1$, the correct logit increment is at least

$$\eta_t (1 - \beta_1) \rho_0 \frac{\kappa}{64\sqrt{\kappa} + \chi} \geq \eta_t C_0. \tag{H.84}$$

The fraction increases for $\kappa \geq 2$. Wrong logits decrease. If the target has not already been reached, the second step alone therefore raises every correct-versus-wrong logit gap by at least $\log(15(q-1))$. Since the gaps started at zero and increased on the first step, $p_{2,*} \geq 15/16$.

For the path bound, Lemma H.2 and Gaussian dual Cauchy–Schwarz bound each row field by K_{mom}. The first step thus changes each logit by at most $1/100$. At both gradient evaluations,

$$p_{t-1,a} \leq e^{0.02}/q \leq 50/(49q), \qquad t = 1, 2.$$

Using $J_{t-1} \leq (3/2)(50/49) r_*/q$ and negativity of wrong coefficients gives $|c_{t,a}| \leq 2 r_*/q^2$ for wrong rows. Their corrected first moments have norm at most $2 r_* \sqrt{K}/q^2$. Since $\|\mathcal{A}(S)\|_{\mathrm{op}} \leq 1/\epsilon_f$, each wrong row field has norm at most $128 r_* \sqrt{K}/q$. Their aggregate norm is at most $128\sqrt{K}/\sqrt{q} \leq 32\sqrt{K}$, while the correct row contributes at most K_{mom}. Summing the two step lengths proves (H.80). □

H.4.4 Complete sampled trajectories and moving parameters

Generate exactly $n = H + 2$ full-vocabulary decisions in the designated choice, state, and EOS slots. The verifier checks grammar, public-coordinate consistency, and the terminal endpoint. Every sampled trajectory, including a zero-reward one, contributes through the original full batch denominator. Define

$$X_t = (W_{\mathrm{fast},t} - W_{\mathrm{fast},s}) C, \qquad X_0 = 0,$$

where $W_{\mathrm{fast},s}$ is the source control-row readout. The fixed-executor choice-only population gradient, with the actual training-prompt distribution, is denoted $U(X)$. Bounded rewards and $\|\Phi\| \leq A$ give the global bound

$$\|U(X) - U(Y)\|_F \leq L_c \|X - Y\|_F, \qquad L_c = 6K. \tag{H.85}$$

At the reference iterates this gradient is (H.77).

For two histories, the Gaussian row fields together are Lipschitz in the maximum Frobenius history error, with constant

$$D_A = \frac{\sqrt{2}(1 + K_{\mathrm{mom}})}{\epsilon_f}. \tag{H.86}$$

To see this, apply (H.55) to each scalar history $z^\top u_s$. The squared scalar maximum is bounded by the sum of its two squared errors. Summing over output rows and using dual Cauchy–Schwarz against a unit Frobenius test matrix gives (H.86); Gaussian second moments replace each scalar square by its coefficient norm squared. There is no factor depending on the number of canonical coordinates or rows. The physical-coordinate version uses epsilon $\epsilon_f/\sqrt{d}$.

Set $G = n\sqrt{2}A$. Apply Lemma H.4 with this G, $\epsilon = \epsilon_f$, and tolerance τ. A complete coefficient score has norm at most G, so the same event covers all actual row histories.

Compare the iterates until their coefficient error exceeds

$$E_* = (1024A)^{-1}, \qquad B_{\exp} = A(R_{\mathrm{ref}} + 1). \tag{H.87}$$

Within this region $\|X_t\|_F \leq R_{\mathrm{ref}} + 1$. Let $h \leq 1$ be the source logit tolerance. The source entry and slow-rate choices in Appendices H.3 and H.5 supply three uniform bounds: the source hard-logit

error plus slow-head and feature-motion errors is at most $2h$; the actual feature error is at most e_f; and the canonical source class leakage is at most h, using $Ne \le h$.

On each matching legal-class history, the actual incorrect-class mass is at most $\ell = h\exp(2B_{\exp}+4)$. The factor pays the change in class odds caused by arbitrary fast-row coefficients in the stopped region. Once an incorrect-class token is generated the coupling has already failed, so no class-mass bound is required on its malformed continuation. Conditional state laws only change through the $2h$ perturbation of state rows; fast control rows do not change the within-state normalization. Conditional choice and state laws differ from the fixed-executor laws by at most $8h$ in total variation.

Let $g_{\mathrm{coeff},t}$ denote the complete reward-weighted score expressed using Φ, with the probabilities and trajectories of the actual policy. Maximal conditional couplings on matching histories give

$$\|\mathbb{E}[g_{\mathrm{coeff},t} \mid \mathcal{F}_{t-1}] - U(X_{t-1})\|_F \le C_{\mathrm{pop}}h, \qquad C_{\mathrm{pop}} = 256n^2A\exp(2B_{\exp}+4). \tag{H.88}$$

Here the update indexed t is evaluated at iterate $t-1$. Indeed the path total variation is at most $n(8h+\ell)$, and both complete coefficient score vectors have norm at most $n\sqrt{2K}$. On a matched valid path, each non-choice fast-control score has norm at most $\sqrt{2K}\ell$, whereas the reference fast score is zero there. The choice fast-score difference is at most $8\sqrt{2K}(h+\ell)$. The same deterministic verifier returns the same reward on matched paths, including histories whose hidden errors repair. Hence the expectation error is at most

$$2n\sqrt{2K}\,n(8h+\ell) + 8n\sqrt{2K}(h+\ell) \le C_{\mathrm{pop}}h.$$

This accounts for non-choice prefixes, EOS, distribution change, and malformed trajectories without a teacher-prefix substitution.

For a desired sampling error ε_g, take

$$B_R = \left\lceil 8qkn^2A^2\varepsilon_g^{-2}\log\frac{4qk}{\delta_R}\right\rceil. \tag{H.89}$$

Each complete-trajectory coefficient entry lies in $[-nA, nA]$. Conditional Hoeffding and a union over qk coordinates and the two predictable iterates give both Frobenius errors at most ε_g with probability at least $1-\delta_R$; the smaller threshold $\varepsilon_g/2$ already meets this allocation. Only trajectories within each fresh batch are independent.

The actual physical fast gradient differs from $g_{\mathrm{coeff},t}C^\top$ by at most $n\sqrt{2}e_f$. Apply the physical history bound before projection: for two histories the squared sum of coordinatewise maximal errors is at most the sum of the two squared Frobenius errors. Multiplying the physical rate $\eta_t/\sqrt{d}$ by the reciprocal physical epsilon $\sqrt{d}/\epsilon_f$ cancels d. Postmultiplication by C, with $\|C\|_{\mathrm{op}} \le 2$, therefore costs at most $2\eta_t D_A n\sqrt{2}e_f$. The fixed-feature update is exactly (H.72); its uniform operator error costs at most $\eta_t\tau G$ across all rows. Writing $E_t = \max_{0\le s\le t}\|X_s - Y_s\|_F$, Gaussian history comparison and (H.85) give, for $t = 1, 2$,

$$E_t \le (1+\eta_t D_A L_c)E_{t-1} + \eta_t\left[D_A(C_{\mathrm{pop}}h+\varepsilon_g) + \tau G + 2D_A n\sqrt{2}e_f\right]. \tag{H.90}$$

Both the moment numerator and the current random denominator are retained. The equality $X_0 = Y_0 = 0$ concerns increments from the learned source; its nonzero source function error enters (H.88).

Define

$$Q_{\mathrm{amp}} = (1+\eta_1 D_A L_c)(1+\eta_2 D_A L_c), \qquad \xi = \frac{E_*}{8S_{\mathrm{step}}Q_{\mathrm{amp}}}. \tag{H.91}$$

Choose

$$h \le \min\left\{\frac{1}{2^{20}n\exp(2B_{\exp}+4)}, \frac{\xi}{D_A C_{\mathrm{pop}}}\right\}, \qquad \varepsilon_g \le \xi/D_A, \tag{H.92}$$

$$\tau \le \xi/G, \qquad e_f \le \min\left\{\frac{\xi}{2D_A n\sqrt{2}}, \frac{h}{32R_W}\right\}. \tag{H.93}$$

Here a deterministic bound for the complete physical readout is $R_W = R + S_{\mathrm{step}}K_{\mathrm{mom}}\sqrt{q}+1$, where R is the source head bound. The forcing in (H.90) is at most 4ξ, so $E_2 \le 4S_{\mathrm{step}}Q_{\mathrm{amp}}\xi = E_*/2$ and

the stopped region cannot be exited. This argument controls the large second update from its gradient evaluation inside the region.

The source feature error is included in the total e_f allowance, rather than reset at feedback entry. It suffices to cap it at $e_f/2$, to require slow readout movement at most $h/(32A)$, and to bound additional backbone movement by

$$\min\left\{\frac{1}{2}, \frac{e_f}{2D}, \frac{h}{32DR_W}\right\} \tag{H.94}$$

inside the remaining derivative ball. Appendix H.5 uses the stronger source allowance $e_f/4$, reserves the radius remaining after source training, and specifies a positive common slow multiplier. The bound on each Adam coordinate from Lemma H.2 makes these movement bounds deterministic. They include every backbone parameter and the normalized member's gain. The full R_W also includes physical readout nullspace components and pays the head/backbone cross term through the feature error.

At the final iterate the choice logit error from $X_2 - Y_2$ is at most AE_*, with an additional base error at most $2h$. The remaining complete-path error is bounded by $n(8h + h\exp(2B_{\text{exp}} + 4))$. The reference task utility and first-state retention are at least $(15/16)(1 - 9c/8)$, and its grammar-or-public-coordinate violation probability is at most $9c/8$. Appendix H.5 combines these margins with the source baseline, the finite-width events, and (H.93) to obtain the stated capability and gain bounds.

The two feedback batches use exactly $2B_R$ fresh training prompts, $2B_R$ binary verifier calls, $2nB_R$ unrestricted generated decisions, and two optimizer updates. The $2HB_R$ state, $2B_R$ control, and $2B_R$ EOS counts refer to designated positions; the actual emitted token classes may differ. There is no feedback checkpoint selection or test-time verifier use.

H.5 An explicit joint source and feedback budget

We give an explicit order for choosing parameters for the Gaussian architecture of Appendix G, the sampled ridge-Adam source procedure of Appendix H.3, and the two sampled feedback updates of Appendix H.4. All logarithms are natural unless their base is displayed. Fix

$$q = 2^{2r+2} \geq 16, \qquad V = 2q + 2, \qquad c = \frac{1}{64}, \qquad K = 12, \qquad \beta_1 = 0.9, \quad \beta_2 = 0.999.$$

The source estimates use $A_{\text{src}} = 12$, whereas feedback uses $A_{\text{fb}} = \sqrt{K}$. These two constants have different roles. Here S is S_{step} in Appendix H.4, and ε_f below is its feature tolerance e_f. For a requested failure probability $0 < \delta < 1$, allocate

$$\begin{aligned} \delta_{\text{bias}} = \delta_{\text{enc}} = \delta_{\text{ffn}} = \delta_{\text{norm}} = \delta/16, \qquad & \delta_0 = \delta/16, \\ \delta_{\text{geo}} = \delta_{\text{batch}} = \delta_R = \delta/8. & \end{aligned} \tag{H.95}$$

The allocation δ_0 is used separately for the row envelope and the readout-coordinate event.

H.5.1 Task, horizon, and source tolerances

First define constants that depend only on q and the fixed optimizer parameters:

$$\begin{aligned} u &= 1 - \beta_1, & K_{\text{mom}} &= \left[(1-\beta_2)(1-\beta_1^2/\beta_2)\right]^{-1/2}, \\ \epsilon_f &= \frac{1}{64q}, & \rho_0 &= \frac{2\exp(-2)}{\sqrt{2\pi}}, \\ C_0 &= \frac{2(1-\beta_1)\rho_0}{64\sqrt{2}+1}, & \eta_1 &= \frac{1}{100K_{\text{mom}}\sqrt{K}}, \\ \eta_2 &= \frac{\log(15(q-1))}{C_0}, & S &= \eta_1 + \eta_2, \\ C_{\text{move}} &= K_{\text{mom}} + 32\sqrt{K}, & R_{\text{ref}} &= SC_{\text{move}}, \\ B_{\text{exp}} &= A_{\text{fb}}(R_{\text{ref}} + 1), & E_{\text{exp}} &= \exp(2B_{\text{exp}} + 4), \end{aligned}$$

$$D_A = \frac{\sqrt{2}(1+K_{\mathrm{mom}})}{\epsilon_f}, \qquad L_c = 6K,$$
$$E_* = \frac{1}{1024 A_{\mathrm{fb}}}, \qquad Q_{\mathrm{amp}} = (1+\eta_1 D_A L_c)(1+\eta_2 D_A L_c),$$
$$\xi = \frac{E_*}{8SQ_{\mathrm{amp}}}.$$

Set

$$h_0 = \min\left\{\frac{1}{2^{20}E_{\mathrm{exp}}}, \frac{\xi}{256 A_{\mathrm{fb}} D_A E_{\mathrm{exp}}}\right\},$$
$$C_q = \frac{2}{3}(5q+V+1)+2q(q-1),$$
$$L_0 = \max\left\{1, \log\frac{24qV}{ch_0}\right\},$$
$$B_H = \frac{7372800 K C_q q^3 V^{3/2} L_0^2}{h_0^2}.$$

Choose any integer horizon satisfying

$$H \geq \lceil \max\{128,\, 2\log_2 B_H\} \rceil, \qquad n = H+2, \qquad N = 2H+2. \tag{H.96}$$

For the minimum prescribed horizon, the right side determines H directly from q; larger prescribed horizons are also allowed. Set the tolerances exactly as follows:

$$h = \frac{h_0}{(H+2)^2}, \qquad e = \min\left\{\frac{h}{N}, \frac{c^2(q-1)^2}{32q^2H^2N}\right\},$$
$$\overline{D} = \max\{D_*(e/2), e\}, \qquad \lambda = \frac{e}{\overline{D}}, \qquad R_\lambda = \sqrt{2\overline{D}}. \tag{H.97}$$

Here D_* is the deterministic comparator envelope from Appendix H.3: its kernel weights are replaced by $k_{\min} = 3/(2N)$ and $k_{0H} = 1/N$. Explicitly, with

$$a = 1 - c/H, \qquad \beta_* = \log\frac{1+(q-1)a}{1-a}, \qquad M = \beta_* + \log\frac{2V}{e},$$

the value used in (H.97) is

$$D_*(e/2) = \frac{2N}{3}(5q+V+1)M^2 + 2Nq(q-1)\beta_*^2.$$

Thus these choices use neither realized lag frequencies nor the hidden mechanism nor any test score.

We check the horizon substitution explicitly. For $H \geq 5$, the above choices give

$$h \geq \frac{h_0}{4H^2}, \qquad e \geq \frac{h_0}{12H^3}, \qquad \beta_* \leq \log(qH/c),$$
$$\overline{D} \leq 3C_q H(L_0 + 4\log H)^2 \leq 75 C_q L_0^2 H^3, \qquad \lambda h \geq \frac{h_0^2}{3600 C_q L_0^2 H^8}.$$

In the source-tag estimate, $\rho_c \leq 1/2$ and

$$\mu_F = 2q^3(a\rho_c)^{H-4}V^{3/2}\sqrt{K}, \qquad 64\sqrt{K}\,\mu_F \leq 2048 K q^3 V^{3/2} 2^{-H}.$$

For $H \geq 128$, $H^8 \leq 2^{H/2}$; moreover, (H.96) gives $B_H \leq 2^{H/2}$. Hence $2^H \geq B_H H^8$, which proves

$$\mu_F \leq \frac{\lambda h}{64\sqrt{K}}.$$

The constant B_H contains no H. In particular, this step does not require a source tolerance exponentially small in the horizon. The equality choice of h also gives both feedback requirements

$$h \leq \frac{1}{2^{20} n E_{\mathrm{exp}}}, \qquad h \leq \frac{\xi}{D_A C_{\mathrm{pop}}}, \qquad C_{\mathrm{pop}} = 256 n^2 A_{\mathrm{fb}} E_{\mathrm{exp}}.$$

H.5.2 Geometry and residual width

Define the feature dimensions and feedback precisions before selecting the residual width:

$$p = NV, \qquad k_{\mathrm{aug}} = p + 1, \qquad s_{\mathrm{sym}} = k_{\mathrm{aug}}(k_{\mathrm{aug}} + 1)/2, \qquad k = p + s_{\mathrm{sym}} + 1,$$
$$G = n\sqrt{2}A_{\mathrm{fb}}, \qquad \varepsilon_g = \xi/D_A, \qquad \tau = \xi/G.$$

For completeness, instantiate Appendix H.4.2 with k, G, ϵ_f, τ, and δ_{geo}. Its explicit constants are

$$K_3 = \frac{8[k(k+2)]^{3/4}}{\delta_{\mathrm{geo}}}, \qquad L_{\mathrm{pert}} = \frac{6k}{\epsilon_f} + \frac{GK_3}{\epsilon_f^2},$$
$$\gamma_{\mathrm{geo}} = \min\{1/8, \tau/(16L_{\mathrm{pert}})\}, \qquad a_{\mathrm{geo}} = \tau/2,$$
$$T_{\mathrm{geo}}^2 = \max\left\{1, \frac{64k(k+2)}{\epsilon_f a_{\mathrm{geo}} \delta_{\mathrm{geo}}}\right\}, \qquad h_{\mathrm{geo}} = \left(\frac{a_{\mathrm{geo}}\epsilon_f^2}{16T_{\mathrm{geo}}^3}\right)^2,$$
$$D_{\mathrm{geo}} = k(k+1)/2, \qquad \mathcal{N}_{\mathrm{geo}} = \left(1 + \frac{2\sqrt{k}G^2}{h_{\mathrm{geo}}}\right)^{D_{\mathrm{geo}}}.$$

The sufficient width is

$$d_{\mathrm{geo}} = \left\lceil \max\left\{32\gamma_{\mathrm{geo}}^{-2}\left[k\log 9 + \log\frac{16}{\delta_{\mathrm{geo}}}\right], \frac{32k^2T_{\mathrm{geo}}^4}{\epsilon_f^2 a_{\mathrm{geo}}^2}\log\frac{16k^2\mathcal{N}_{\mathrm{geo}}}{\delta_{\mathrm{geo}}}\right\}\right\rceil. \tag{H.98}$$

All these quantities are fixed before d.

We next recall the dimension-independent source envelopes, retaining $A_{\mathrm{src}} = 12$:

$$\overline{m} = \sqrt{2\max\{1, \log(2V/\delta_0)\}},$$
$$\overline{C} = 2\sqrt{2k + 4\log(2/\delta_0) + 4}, \qquad \bar{c}_h = 2\sqrt{K}\,\overline{C} + 3,$$
$$\bar{g} = 1 + \bar{c}_h + \max\{\lambda\overline{m}, \bar{c}_h + \lambda K_{\mathrm{mom}}/u\},$$
$$\overline{f} = \log V + 2A_{\mathrm{src}}\sqrt{V}\,\overline{m} + \frac{\lambda V}{2}\overline{m}^2, \qquad \overline{r} = \sqrt{\frac{2(2\overline{f}+1)}{\lambda}},$$
$$\rho = \frac{\lambda h}{64\sqrt{K}}, \qquad B_{\mathrm{dim}} = \max\left\{16, \frac{1024\bar{g}(A_{\mathrm{src}}\overline{r} + \sqrt{2})}{\rho}, \frac{128\overline{r}}{h}\right\}.$$

The source formulas below satisfy, for $t = 1 + \log d$,

$$R(d) \le \overline{r}\sqrt{t}, \qquad \varepsilon_{\mathrm{feat,src}}(d) \ge \frac{16}{B_{\mathrm{dim}}t}.$$

Put

$$R_W(d) = R(d) + SK_{\mathrm{mom}}\sqrt{q} + 1, \qquad \overline{r}_W = \overline{r} + SK_{\mathrm{mom}}\sqrt{q} + 1,$$
$$\varepsilon_{f,0} = \frac{\xi}{2D_A n\sqrt{2}}, \qquad \overline{\varepsilon}_f = \min\{\varepsilon_{f,0}, h/(32\overline{r}_W)\},$$
$$\varepsilon_f(d) = \min\{\varepsilon_{f,0}, h/(32R_W(d))\}, \qquad \varepsilon_{\mathrm{src}}(d) = \min\{\varepsilon_{\mathrm{feat,src}}(d), \varepsilon_f(d)/4\},$$
$$B_{\mathrm{joint}} = \max\{B_{\mathrm{dim}}, 64/\overline{\varepsilon}_f\}.$$

Here ϵ_f denotes Adam's feedback denominator constant, whereas ε_f denotes feature approximation error. Since $R_W(d) \le \overline{r}_W\sqrt{t}$, we have

$$\varepsilon_f(d) \ge \overline{\varepsilon}_f/\sqrt{t}, \qquad \varepsilon_{\mathrm{src}}(d) \ge \frac{16}{B_{\mathrm{joint}}t}.$$

Consequently, using $\log d \le 16d^{1/16}/\mathrm{e}$, the bound $d \ge [B_{\mathrm{joint}}(1 + 16/\mathrm{e})]^{16}$ ensures $d^{-1/8} \le \varepsilon_{\mathrm{src}}(d)/16$.

Choose the head count and concentration constants by

$$R_{\mathrm{head}} = \left\lceil 8N \log \frac{2N}{\delta_{\mathrm{bias}}} \right\rceil,$$
$$D_{\mathrm{head}} = 2^{20} N \left\lceil V \log 9 + \log \frac{8N}{\delta_{\mathrm{enc}}} \right\rceil,$$
$$D_\Gamma = \max\left\{ 2^{19} \left\lceil p \log 9 + \log \frac{8}{\delta_{\mathrm{enc}}} \right\rceil, 4096 \left\lceil k \log 9 + \log \frac{8}{\delta_{\mathrm{ffn}}} \right\rceil, D_{\mathrm{head}} \right\}.$$

Define

$$d_{\mathrm{req}} = \max\Bigg\{ [B_{\mathrm{joint}}(1+16/\mathrm{e})]^{16}, \left\lceil \frac{64\sqrt{K}\, K(R_\lambda+1)}{\lambda h} \right\rceil^8, \gamma_{\mathrm{geo}}^{-8}, d_{\mathrm{geo}}, 128 \left\lceil k \log 9 + \log \frac{4}{\delta_0} \right\rceil, D_\Gamma^{4/3} \Bigg\}, \qquad d = R_{\mathrm{head}} \left\lceil \frac{d_{\mathrm{req}}}{R_{\mathrm{head}}} \right\rceil, \qquad \gamma = d^{-1/8}. \tag{H.99}$$

Every quantity on the right of d_{req} is independent of d. The condition $d \geq D_\Gamma^{4/3}$ is precisely the substitution of $\gamma = d^{-1/8}$ into $d \geq D_\Gamma \gamma^{-2}$. It supplies the actual encoder and full-feature Gram bounds with encoder tolerance $\gamma/16$. In particular the D_{head} term ensures the head-width bound while preserving the architecture's integer relation

$$k_{\mathrm{head}} = d/R_{\mathrm{head}}.$$

After fixing d, choose the FFN width by

$$J_{\mathrm{ff}} = s_{\mathrm{sym}} \log 9 + \log(8/\delta_{\mathrm{ffn}}), \qquad A_{\mathrm{ff}} = 1 + \gamma^{-2} J_{\mathrm{ff}},$$
$$L_{\mathrm{ff}} = \log \frac{\mathrm{e} A_{\mathrm{ff}} k_{\mathrm{aug}}}{\delta_{\mathrm{ffn}} \gamma}, \qquad m_{\mathrm{ffn}} = \left\lceil 2^{50} A_{\mathrm{ff}}^2 (k_{\mathrm{aug}} + L_{\mathrm{ff}})^8 \right\rceil.$$

H.5.3 Source rates, finite temperature, and positive slow rates

At the now chosen width, the source constants of Appendix H.3 are explicitly

$$M_0 = \sqrt{2\log(2Vd/\delta_0)}, \qquad W_{\max} = \sqrt{V} M_0,$$
$$C_{\mathrm{row}} = 2\sqrt{2k + 4\log(2d/\delta_0)}, \qquad C_h = 2\sqrt{K} C_{\mathrm{row}} + 3,$$
$$G_c = C_h + \lambda \max\{M_0, C_h/\lambda + K_{\mathrm{mom}}/u\}, \qquad \lambda_p = (G_c+1)^{-1},$$
$$L_s = 2K+1, \qquad F_{\mathrm{init}} = \log V + 2A_{\mathrm{src}} W_{\max} + \lambda W_{\max}^2/2,$$
$$R = \sqrt{2(2F_{\mathrm{init}}+1)/\lambda}, \qquad r_{\mathrm{goal}} = \rho/4,$$
$$\eta_s = \min\left\{1, \frac{\lambda_p u^{3/2}}{8L_s}, \frac{1}{\lambda_p r_{\mathrm{goal}}^2}\right\}, \qquad b = \lambda_p \rho/32,$$
$$T_s = \max\left\{1, \left\lceil \frac{4F_{\mathrm{init}}}{\eta_s \lambda_p r_{\mathrm{goal}}^2} \right\rceil\right\},$$
$$\varepsilon_{\mathrm{feat,src}} = \min\left\{1, \frac{b}{2(A_{\mathrm{src}} R + \sqrt{2})}, \frac{h}{8R}\right\},$$
$$B_s = \left\lceil \frac{8VdA_{\mathrm{src}}^2}{b^2} \log \frac{2VdT_s}{\delta_{\mathrm{batch}}} \right\rceil.$$

These formulas also define the width-dependent functions used above; their dimension-independent envelopes were fixed first. Source Adam starts with fresh zero moments, uses physical denominator offset $1/\sqrt{d}$, and uses head rate $\eta_s/\sqrt{d}$. At every inspected iterate it computes the complete-record minibatch head gradient, including the coupled term λW. It stops before an update when that gradient's Frobenius norm is at most $\rho/2$. A failed T_sth check declares failure without another update.

On the joint good event, the source result proves that this failure cannot occur and that the output has the required uniform source logit accuracy. The norm criterion uses only source observations.

All dimensions are fixed before drawing initialization. Let P_{total} be the number of initialized scalars and P_ψ the number of supplied backbone scalars. Define

$$\begin{aligned}
M_{\mathrm{init}} &= \max\left\{1, \sqrt{2P_{\mathrm{total}}\log(2P_{\mathrm{total}}/\delta_{\mathrm{norm}})}\right\},\\
Z_0 &= M_{\mathrm{init}} + \sqrt{R_{\mathrm{head}}}M_{\mathrm{init}}^3,\\
z_{\mathrm{raw}} &= \min\left\{\frac{\varepsilon_{\mathrm{src}}}{8\sqrt{2}}, \frac{1}{4\sqrt{d}}\right\},\\
\ell &= \min\left\{\frac{1}{2N}, \frac{z_{\mathrm{raw}}}{2\sqrt{R_{\mathrm{head}}}M_{\mathrm{init}}^3[1+2M_{\mathrm{init}}^3(Z_0+1)]}\right\},\\
g_{\mathrm{bias}} &= \frac{\delta_{\mathrm{bias}}}{R_{\mathrm{head}}N^2},\\
t_{\mathrm{att}} &= \min\left\{\frac{g_{\mathrm{bias}}}{4M_{\mathrm{init}}^4}, \frac{g_{\mathrm{bias}}}{2\log((N-1)/\ell)}\right\} > 0.
\end{aligned}$$

Thus the implemented attention has a finite positive temperature. For the radius-one derivative bound, write $M_1 = M_{\mathrm{init}} + 1$ and set

$$\begin{aligned}
Z_1 &= M_1 + \sqrt{R_{\mathrm{head}}}M_1^3,\\
D_z &= 1 + \sqrt{R_{\mathrm{head}}}[3M_1^2 + M_1^3N(4M_1^3 + t_{\mathrm{att}}^{-1})],\\
D_{\mathrm{raw}} &= D_z[1+2M_1^3(Z_1+1)] + 3M_1^2(Z_1+1)^2 + 1,\\
D &= \begin{cases} D_{\mathrm{raw}}, & \text{base member},\\ 2\sqrt{2}D_{\mathrm{raw}} + \sqrt{2K}, & \text{normalized gain member}. \end{cases}
\end{aligned}$$

The Gaussian parameter envelope and gain initialization are included in this radius-one bound. In the source precision choices use the capped $\varepsilon_{\mathrm{src}}$, and put

$$\begin{aligned}
r_\psi &= \min\left\{1, \frac{1}{D}, \frac{\varepsilon_{\mathrm{src}}}{2D}, \frac{1}{D\sqrt{d}}\right\},\\
\rho_{\mathrm{back}} &= \min\left\{1, \frac{r_\psi\sqrt{d}}{2\eta_s T_s K_{\mathrm{mom}}\sqrt{P_\psi}}\right\} > 0.
\end{aligned}$$

Every source backbone coordinate uses the physical rate $\eta_s\rho_{\mathrm{back}}/\sqrt{d}$. Its total source displacement is at most $r_\psi/2 \le 1/2$, and the source feature error is at most $\varepsilon_{\mathrm{src}} \le \varepsilon_f/4$.

At feedback, remove the source regularizer and reset all first and second moments. Choose the positive common slow multiplier

$$\rho_R = \min\left\{1, \frac{h}{32A_{\mathrm{fb}}SK_{\mathrm{mom}}\sqrt{V}}, \frac{\sqrt{d/P_\psi}}{SK_{\mathrm{mom}}}\min\left\{\frac{1}{2}, \frac{\varepsilon_f}{2D}, \frac{h}{32DR_W}\right\}\right\}. \tag{H.100}$$

Use physical rate $\eta_t/\sqrt{d}$ for fast control-output rows and $\eta_t\rho_R/\sqrt{d}$ for every slow supplied parameter, at $t = 1, 2$. Every feedback block uses physical denominator offset $\epsilon_f/\sqrt{d}$. Indeed, (H.100) gives

$$\begin{aligned}
\rho_R SK_{\mathrm{mom}}\sqrt{V} &\le h/(32A_{\mathrm{fb}}),\\
\rho_R SK_{\mathrm{mom}}\sqrt{P_\psi/d} &\le \min\{1/2, \varepsilon_f/(2D), h/(32DR_W)\}.
\end{aligned}$$

Source and feedback backbone displacements total at most one, so the same D remains valid during both phases. The bound R_W includes physical readout components in the nullspace of the feature operator. These choices supply total feature error at most ε_f and base, slow-row, and feature-motion logit discrepancy at most $2h$; fast coefficients are controlled by the feedback comparison.

Each of the two feedback updates averages complete on-policy trajectories using batch size

$$B_R = \left\lceil \frac{8qkn^2A_{\mathrm{fb}}^2}{\varepsilon_g^2}\log\frac{4qk}{\delta_R}\right\rceil. \tag{H.101}$$

Zero-reward trajectories and all token-score terms are retained. The baseline is zero. The history-sensitive comparison in Appendix H.4 has per-step forcing at most 4ξ and therefore

$$\max_{t\le 2}\|X_t - Y_t\|_F \le 4SQ_{\mathrm{amp}}\xi = E_*/2.$$

Both nonzero Adam histories participate in the second update. The width and precision choices control the current random denominator before coefficient projection, through the uniform geometry of Appendix H.4.2.

H.5.4 Joint probability and capability margins

The bias, encoder, FFN/full-Gram, and parameter-envelope failures total $4\delta/16$. The row and readout-coordinate failures total $2\delta/16$. Uniform geometry, source batches, and the two feedback batches contribute $3\delta/8$. Thus the union bound is

$$4\delta/16 + 2\delta/16 + 3\delta/8 = 3\delta/4 < \delta.$$

The encoder's four internal allocations total δ_{enc} and the FFN's two internal allocations total δ_{ffn}. The row envelope is conditional on coverage and the actual Gram event; the geometry event is uniform over histories. These facts suffice for the union bound without assuming independent realized feature rows. Gaussian initialization is not conditioned on later content-score events when its independence lemmas are applied. For every fixed unknown world, fresh batches obey the stopped conditional concentration estimates. This is not a simultaneous guarantee for one dataset across all unknown worlds.

The canonical source selects a uniform operation. Sharp transitivity makes its first state uniform, and subsequent doubly stochastic noisy-permutation kernels keep its endpoint uniform. Its valid-class endpoint success and first-state retention are consequently $1/q$, and its task utility is at most $1/q$. Full-vocabulary and actual-logit coupling bounds each of the three actual source metrics by

$$1/q + 2nh.$$

Writing r_{opp} for the opposite arm's reward, the separate exact identity is $r_* + r_{\mathrm{opp}} = D_{\mathrm{exec}} \le 1$; the source uniformity conclusion uses the task structure, rather than the individual reward upper bounds.

After the two reference updates, the correct choice has probability at least $15/16$. Reference task utility and first-state retention are at least

$$\frac{15}{16}\left(1 - \frac{9c}{8}\right) = \frac{7545}{8192}.$$

The coefficient comparison contributes choice-logit error at most $A_{\mathrm{fb}}E_* = 1/1024$. The base, format, and state contribution is

$$n(8h + hE_{\mathrm{exp}}) \le 9/2^{20},$$

and the choice-distribution contribution is at most $1/256$. Their total deployment-law discrepancy is strictly less than $1/128$. Hence all three actual metrics E, S, M exceed $7/8$, and each exceeds its corresponding source value by more than $3/4$. The probability of a grammar or public-coordinate violation is at most

$$\frac{9c}{8} + \frac{1}{128}.$$

The bound uses the original verifier, which checks grammar, public-coordinate consistency and the endpoint. In the reference, only the correct arm and its hidden-bit opposite can receive reward, including histories whose private-state errors repair.

H.5.5 Polynomial substitution and resource accounting

We make the resource dependence explicit, including the late choices of temperature and positive rates. With fixed optimizer parameters, η_1 is constant, $\eta_2 = O(\log q)$, and $R_{\mathrm{ref}} = O(\log q)$. In fact

$$E_{\mathrm{exp}} = \exp\big(2A_{\mathrm{fb}}(C_{\mathrm{move}}\eta_1 + 1) + 4\big)[15(q-1)]^{2A_{\mathrm{fb}}C_{\mathrm{move}}/C_0}.$$

This is a fixed power of q. Moreover $D_A = O(q)$, $Q_{\mathrm{amp}} = O(q^2 \log q)$, and $\xi^{-1} = O(q^2(\log q)^2)$, so h_0^{-1} is polynomial in q up to logarithms. Thus B_H is explicit and the minimum horizon in (H.96) is finite. For any admitted H,

$$h^{-1} = \frac{(H+2)^2}{h_0}, \qquad e^{-1} \le \frac{12H^3}{h_0}, \qquad \lambda^{-1} \le \frac{900 C_q L_0^2 H^6}{h_0}.$$

These bounds make the dimension-independent source envelopes polynomial in the expanded variables q, H, δ^{-1}, up to logarithms.

The geometry net does not introduce an exponential width: it enters only through

$$\log \mathcal{N}_{\mathrm{geo}} = D_{\mathrm{geo}} \log \left(1 + \frac{2\sqrt{k} G^2}{h_{\mathrm{geo}}}\right).$$

All other geometry constants are powers, products, or positive reciprocals of already chosen quantities. Thus d_{geo}, B_{joint}, and d_{req} are polynomial in those expanded variables. Rounding adds less than R_{head} to d_{req}. The number of heads is polynomial, and $k_{\mathrm{head}} = d/R_{\mathrm{head}}$ is integral. Since $\gamma^{-2} = d^{1/4}$,

$$A_{\mathrm{ff}} = 1 + d^{1/4} J_{\mathrm{ff}}, \qquad m_{\mathrm{ffn}} = O\Big((1 + d^{1/4} J_{\mathrm{ff}})^2 [k_{\mathrm{aug}} + \log d + \log(\delta_{\mathrm{ffn}}^{-1})]^8\Big),$$

with the explicit multiplier 2^{50} given above. The complete architecture of Appendix G has polynomially many scalars in these dimensions, hence P_{total}, P_ψ, and M_{init} are polynomial up to logarithms.

The source cap gives $\varepsilon_{\mathrm{src}}^{-1} \le B_{\mathrm{joint}}(1 + \log d)/16$. Therefore z_{raw}^{-1} and ℓ^{-1} are polynomial as well. In particular

$$t_{\mathrm{att}}^{-1} = \max\left\{\frac{4M_{\mathrm{init}}^4}{g_{\mathrm{bias}}}, \frac{2\log((N-1)/\ell)}{g_{\mathrm{bias}}}\right\}$$

is polynomial. The displayed derivative formula is a fixed-degree polynomial in M_1, N, $\sqrt{R_{\mathrm{head}}}$, and t_{att}^{-1}; consequently D is polynomial for both members.

Finally, the displayed formulas make η_s^{-1}, T_s, B_s, and B_R polynomial. For the late backbone rate,

$$r_\psi^{-1} = \max\{1, D, 2D/\varepsilon_{\mathrm{src}}, D\sqrt{d}\}, \qquad \rho_{\mathrm{back}}^{-1} = \max\left\{1, \frac{2\eta_s T_s K_{\mathrm{mom}} \sqrt{P_\psi/d}}{r_\psi}\right\},$$

and (H.100) similarly gives a polynomial bound for ρ_R^{-1}. Multiplying by $\sqrt{d}$ and the reciprocals of η_s, η_1, η_2 proves the same statement for every positive physical learning-rate reciprocal. Neither D nor these late rates occurs on the right side of the residual-width choice. This ordered substitution gives a polynomial real-arithmetic bound in expanded q, H, δ^{-1}, up to logarithms, with the absolute exponents specified above.

The source phase uses at most $B_s T_s$ fresh complete raw records and $N B_s T_s$ predicted next-token observations. On every outcome it performs at most $T_s - 1$ optimizer updates and T_s gradient-norm checks, and makes zero task-verifier calls. The feedback phase uses exactly $2B_R$ fresh prompts, containing $2HB_R$ prompt tokens, and exactly $2B_R$ terminal verifier calls. It generates $2nB_R$ unrestricted tokens and performs exactly two optimizer updates. There are $2HB_R$ designated state positions, $2B_R$ designated control positions, and $2B_R$ designated EOS positions. These are position counts: the full-vocabulary sampler can emit an incorrect type at any position, so each emitted-type count is bounded by $2nB_R$ unless measured from the actual sequence. There is no test-time verifier, reward filtering, source-success labeling, or feedback checkpoint selection. Section F compares the corresponding learning routes and resources.

H.5.6 A FROZEN-SOURCE CONTROLLER AND SEPARATE EXECUTION COSTS

On the same source event, the canonical correct-state probability at every legal prefix is at least $(1-h)(1-9c/(8H))$. The actual source logit error is below $h/2$, so its probability is at least $(1-h)(1-9c/(8H)) - h > 1/2$ under (H.97). EOS also has probability above $1/2$. Thus externally forcing any candidate first control and greedily decoding the frozen source model yields its exact state trajectory and EOS, including on unseen words. By sharp transitivity, a training query accepts

exactly when that candidate equals the unknown binding. Enumerate at most $q-1$ candidates on fresh training prompts and infer the last if all reject. This uses at most $q-1$ verifier calls, $n(q-1)$ generated symbols, $H(q-1)$ prompt tokens, and $(H+1)(q-1)$ neural forward evaluations, with zero further gradient updates. Store the binding index and an enumeration counter in addition to the model and its retained context. Deployment requires $H+1$ forwards and n generated symbols with no verifier. This comparator retains the original source cost.

Padding is separate from observed information. Each source record starts with N PAD tokens; separate full-prefix evaluation of its N targets uses N^2 context-token occurrences, including $N(N+1)/2$ PAD occurrences. Each feedback trajectory uses nN context-token occurrences and H prompt symbols. Each verifier performs H private transitions and H public-coordinate checks. Source stopping uses an already-computed head gradient norm, without extra data or verifier access. The model stores P_{total} real parameters and the corresponding optimizer moments. These costs use real arithmetic and real-valued parameter storage.

H.5.7 Opaque-table recovery and fresh-prompt binding selection

Under the raw source law, execution controls $U_1, \dots, U_{H-1}$ are jointly iid uniform: U_1 is independent, and $U_2, \dots, U_{H-1}$ form a strict subset of the tag-conditioned suffix. Retain odd execution transitions through $H-1$ and average the intervening even controls. Their inputs become independent uniform states because $q^{-1}\sum_{U\in\mathcal{U}} T_U = \Pi$. Thus a complete record supplies $\lfloor H/2 \rfloor$ iid triples in this marginal retained experiment. Take

$$M_0 = \left\lceil 8q^2 a^{-2} \log \frac{2q^3}{\delta_{\mathrm{tab}}} \right\rceil, \qquad n_{\mathrm{rec}} = \left\lceil \frac{M_0}{\lfloor H/2 \rfloor} \right\rceil. \tag{H.102}$$

The charged raw-token count satisfies

$$N n_{\mathrm{rec}} = (2H+2) n_{\mathrm{rec}} \le 6M_0 + 2H + 2.$$

The empirical mode for every opaque state/control row recovers all q^2 successors with probability at least $1-\delta_{\mathrm{tab}}$. This estimator requires neither a coordinate chart nor downstream goal labels.

For completeness, let $M = n_{\mathrm{rec}}\lfloor H/2 \rfloor$ be the number of retained iid triples and put $t = \log(2q^3/\delta_{\mathrm{tab}})$. Each row count has mean $\mu = M/q^2 \ge 8a^{-2}t$. Binomial lower-tail concentration bounds the probability of fewer than $\mu/2$ observations by $e^{-\mu/8}$. Conditional on n observations of a row, true-minus-competitor indicator differences lie in $[-1,1]$ and have mean a. Their centered log moment generating function is at most $\lambda^2/2$, so an incorrect empirical mode, including a tie, has probability at most $e^{-na^2/2}$. Union bounding the q^2 row counts and $q^2(q-1)$ comparisons gives failure probability at most δ_{tab}. This proves recovery with full- record rounding. It does not assume independence of all H transitions within a record.

On recovery, enumerate candidate bindings on separately drawn fresh uniform training prompts. Generate each candidate's exact path using the recovered table. The binary terminal reward equals one exactly when that candidate is η: suffix invertibility and sharp transitivity of the first transition prove the equivalence. Test at most $q-1$ candidates, stopping at success or inferring the last binding if all reject. No canonical-prompt query selection is needed.

The additional costs are at most $q-1$ verifier calls, $(H+2)(q-1)$ generated symbols, and H table lookups per candidate or deployment. There are zero gradient updates and no test-time oracle. Final storage is q^2 successor indices and one binding index; training may use q^3 counts. Deployment has perfect utility and public-coordinate compliance on the recovery event. This comparator uses the same observations with a table-based computation, separating information requirements from the neural training route. The $2H+2$ granularity term and every unused raw token remain charged.

H.6 Output format and regularized source learning

The source theorem uses a regularizer to select a structured optimum. The following obstruction explains why accurate conditional execution and a small unregularized source gradient would not suffice. It concerns population risk in the canonical coefficient class, rather than the trajectory selected by neural Adam.

Proposition H.6 (A source near-minimizer with incorrect held-out format)**.** *Use the raw source law with $H \geq 5$ and $0 < a < 1$. Let $f_\Theta(x) = \Theta\Phi(x)$ contain all positively weighted categorical singleton and pair features, with $\|\Phi(x)\| = \sqrt{K}$, $K = 12$, and let F_+ be population source CE. Fix a forbidden suffix $w = (w_2, \dots, w_H)$. For every $\varepsilon > 0$ and $\delta_f \in (0, 1)$, there is a finite Θ such that*

$$F_+(\Theta) - \inf_{\Theta'} F_+(\Theta') \leq \varepsilon, \qquad \|\nabla F_+(\Theta)\|_F \leq \varepsilon,$$

all within-state conditional predictions equal the noisy true transitions, but EOS has probability at least $1 - \delta_f$ at the first-state prediction on w, uniformly over the initial state and preceding chosen control.

Proof. Let $P_l = \mathbf{1}\{x_l = \mathrm{PAD}\}$. The polynomial $D_1 = P_{H+1} - P_H$ is one exactly when the raw prefix length is $H + 1$, the first-state target, and zero at every other source target. At that stage U_j lies at lag $H + 1 - j$. Define

$$d_w(x) = \sum_{j=2}^{H} \big(1 - \mathbf{1}\{x_{H+1-j} = w_j\}\big), \qquad \Delta f_{\mathrm{EOS}}(x) = M\big(\tfrac{1}{2} - d_w(x)\big) D_1(x). \tag{H.103}$$

This is a degree-two categorical function. Every training suffix differs from w, so $d_w \geq 1$ and the change suppresses the incorrect EOS logit by at least $M/2$ wherever it acts. On w it raises that logit by $M/2$.

Choose finite base logits with exact noisy state kernels, uniform initial-state and control conditionals, and a format separator whose leakage is at most ζ. The separator in (H.13), with sufficiently large finite amplitude, provides this property even at the actual EOS, where the state interaction remains active.

The resulting within-class population objective has zero gradient in every coefficient. State predictions equal their true conditional laws; the initial state and early controls are uniform. At the last suffix control U_H, the total tag constrains one bit, but any degree-two prefix feature retains at most two of the $H - 2 \geq 3$ preceding suffix controls. An unused independent balanced tag annihilates the target-feature correlation. Uniform control logits therefore have zero population gradient at this target as well. Convexity implies the within-class minimum

$$H_0 = \frac{(H + 1) \log q + H h_q(a)}{N}.$$

Full CE is that objective plus nonnegative class loss, and finite separators approach zero class loss. Hence $\inf F_+ = H_0$. The unrestricted conditional entropy is lower by $\log 2/N$; the coefficient-class infimum is not being identified with it.

The perturbation changes no correct-class conditional law and only reduces source class leakage. Consequently

$$F_+(\Theta) \leq H_0 - \log(1 - \zeta), \qquad \|\nabla F_+(\Theta)\|_F \leq 2\sqrt{K}\,\zeta,$$

independently of M. The second bound follows because the full probability vector differs from its within-class counterpart by at most 2ζ in ℓ_1; average the residual outer products over all targets. Taking $\zeta = \min\{1/2, 1 - e^{-\varepsilon}, \varepsilon/(2\sqrt{K})\}$ supplies the source bounds.

At a forbidden first-state prefix, the EOS probability is

$$p_{\mathrm{EOS}} = \frac{1}{1 + R(x) e^{-M/2}}, \qquad R(x) = \sum_{y \neq \mathrm{EOS}} e^{f_y(x) - f_{\mathrm{EOS}}(x)}.$$

The base logits and the set of initial-state/control combinations are finite, so $R_{\max} < \infty$. Choose $M \geq \max\{0, 2 \log[R_{\max}(1 - \delta_f)/\delta_f]\}$. This proves the uniform EOS bound. Under the fixed-length sampler an EOS token in a state slot is a grammar failure, regardless of subsequent tokens. □

The perturbation increases the coefficient norm with M and targets one forbidden suffix. It isolates the role of regularization in extending source prediction to the required output format.

I CONTEXTUAL ACQUISITION AND SAMPLED ADAM ADAPTATION

I.1 ACQUISITION, CONTEXTUAL TRANSFER AND INTERVENTION

Use the task and attention–bilinear architecture in Section 5.1, with $n_{\mathrm{ctx}} \geq 1$ and odd $p \geq 17$. The fixed normalizer is the finite positive mixture in (5.4), with weights summing to one, powers in $[0, 1]$ and offsets in $[0, Z_{\max}]$. Initialize every factor coordinate independently from a standard Gaussian. Use $0 \leq \beta_1 < 1$, $0 < \beta_2 < 1$, $\beta_1^2 < \beta_2$, and the physical step sizes and denominator offsets specified below, which implement a common normalized update. Retain all parameters between stages and reset only the two Adam moments.

The physical query/key rows are $Q_i = q_i/\sqrt{d_{\mathrm{width}}}$ and $K_i = k_i/\sqrt{d_{\mathrm{width}}}$; physical head rows are a_i, b_i, o_i. A normalized step η_t and positive denominator offset ϵ use query/key step $\eta_t/\sqrt{d_{\mathrm{width}}}$ and offset $\epsilon/\sqrt{d_{\mathrm{width}}}$, and head step η_t and offset $\epsilon/d_{\mathrm{width}}$. Every normalized coordinate therefore has the same nonzero rate and offset. At an exceptional zero attention pool, define $g = 0$ and its selected derivative as zero; the proof events avoid this point.

Theorem I.1 (Acquisition, contextual transfer and intervention)**.** *Fix the task dimensions, an architecture (5.4), Adam parameters (β_1, β_2) as above, a world (U, s, τ) and $\delta \in (0, 1)$. Choose width, batches, steps, offsets and input precision as in Sections I.7 and I.8; they are explicit and finite. With probability at least $1 - \delta$ over initialization and all training samples, jointly:*

1. Acquisition. *Initial source matching is at most 0.51; after source training it is at least 0.9 uniformly over contexts, frames and queries, and greedy source-bit accuracy exceeds 0.9.*
2. Transfer. *At the final feedback iterate, sampled success is at least $19/20$ on every held-out context, query, cue and amplitude pair, an improvement of at least $2/5$ over the source checkpoint.*
3. Intervention. *The same schedules and budgets with the initial query/key parameters frozen in both stages, and every head factor trained, end with sampled test success at most $1/2 + 1/64$; the paired gap exceeds $2/5$.*

The conclusions hold for the finite coordinate alphabet of Section I.8 with supplied numerical decoding and real-valued optimizer states. At fixed $n_{\mathrm{ctx}}, p, Z_{\max}, m_{\mathrm{norm}}$ the budget formulas may be shared across mixture members, with probability evaluated for each fixed member.

The proof is given in Section I.7.

I.2 SOURCE LEARNING IN THE RADIAL CLASS

We prove Theorem 5.2 with explicit constants. All core norms below are Frobenius norms, combined by the Euclidean direct sum. The row coordinates are the normalized parameters of Section 5.1. Time derivatives in this and the next section describe a deterministic reference; Section I.6 constructs that reference and couples the actual finite Adam runs to this analytical reference.

I.2.1 GAUSSIAN REDUCTION AND STRICT DRIFT

At a source core $M = \beta U$, $\beta \geq 0$, write

$$a = \sigma(\beta), \quad S = a^2 + (1 - a)^2, \quad \varrho = a^2/S, \quad \rho = \sqrt{\varrho}.$$

Then $S_\beta/S = 2(\varrho - a)$ and $\varrho_\beta = 2\varrho(1 - \varrho)$. Gaussian conditioning gives a pooled content vector $h = \sqrt{S}G$, $G \sim N(0, I_p)$, and

$$\mathbb{E}[y \mid G] = m_\rho(G_1)m_\rho(G_2), \qquad m_\rho(z) = 2\Phi\left(\frac{\rho z}{\sqrt{1 - \rho^2}}\right) - 1.$$

Here Φ and ϕ denote the standard Gaussian cumulative distribution function (CDF) and density.

Source sign and permutation symmetries force the reference head core to have only $C_{12} = C_{21} = t/2$ in its p-dimensional content block. The indices in this subsection refer to that block. Flipping an

inactive value coordinate and its two factor coordinates preserves the source law and kills entries with exactly one such index. Negating o together with both first-channel factors preserves source alignment and kills diagonal and inactive–inactive mean entries. Exchanging a, b equates the remaining two entries. Context-factor sign flips apply because source context inputs vanish. These are signed parameter permutations, so uniqueness of the coordinatewise reference preserves the row law. At the resulting core, the population head gradient itself has only its $12, 21$ entries: the same Gaussian sign symmetries kill the other derivatives individually. This stronger fact will preserve the inactive row factors.

The source logit is $f = th_1h_2\psi(\|h\|^2)$, and its loss has the exact decomposition

$$L = \mathbb{E}\log(1 + e^{-|f|}) + \frac{1}{2}\mathbb{E}\left[|f|\{1 - y\,\mathrm{sign}(G_1G_2)\}\right]. \tag{I.1}$$

For one component (α, ζ) with $\alpha > 0$, put $v = \zeta/S$ and define

$$D = \frac{1}{2}\mathbb{E}\left[\frac{|G_1G_2|}{(\|G\|^2 + v)^\alpha}\{1 - m_\rho(G_1)m_\rho(G_2)\,\mathrm{sign}(G_1G_2)\}\right].$$

The Gamma integral and elementary Gaussian integration yield

$$D = \frac{1 - \varrho}{\pi\Gamma(\alpha)}\int_0^\infty \frac{u^{\alpha-1}e^{-vu}(1 + 2u)^{-p/2}}{1 + 2u(1 - \varrho)}\,du. \tag{I.2}$$

Indeed,

$$\mathbb{E}[Ze^{-uZ^2}m_\rho(Z)] = \frac{\sqrt{2/\pi}\rho}{(1 + 2u)\sqrt{1 + 2u(1 - \rho^2)}}.$$

Multiply this identity for the two active Gaussian coordinates and by $(1 + 2u)^{-(p-2)/2}$ for the rest, then subtract from the perfect-sign expression. The Gamma representation follows by substitution in its defining integral; nonnegative sign-adjusted integrands justify Tonelli.

Let $\mathbb{E}_w$ integrate under the normalized positive integrand in (I.2), and write $B_u = 1 + 2u(1 - \varrho)$. Integration of the derivative of u times that integrand gives

$$\mathbb{E}_w B_u^{-1} = 1 - \alpha + p\mathbb{E}_w\frac{u}{1 + 2u} + v\mathbb{E}_w u.$$

The origin boundary vanishes for $\alpha > 0$, and the infinity boundary and derivatives are integrable for $p \geq 17$, including $v = 0$. Differentiating with $v_\beta = -2(\varrho - a)v$ gives

$$-\frac{\partial_\beta(S^{1-\alpha}D)}{S^{1-\alpha}D} = 2\left[(1 - \alpha)a + \varrho p\mathbb{E}_w\frac{u}{1 + 2u} + av\mathbb{E}_w u\right] > 0. \tag{I.3}$$

For $\alpha = 0$ the direct formula is $D = (1 - \varrho)/\pi$, and the ratio is $2a$; no $\Gamma(0)$ is used.

For every component, the derivative with respect to S of its margin factor is

$$\partial_S\frac{S}{(S\|G\|^2 + \zeta)^\alpha} = \frac{\zeta + (1 - \alpha)S\|G\|^2}{(S\|G\|^2 + \zeta)^{\alpha+1}} \geq 0.$$

The first term of (I.1) therefore cannot increase with β. Its second term is t times the positive weighted sum of $S^{1-\alpha}D$. Equation (I.3) proves $L_\beta < 0$ whenever $t > 0$ and β is finite. The argument averages the noise terms componentwise; it does not replace the logistic loss of a mixture by a mixture of logistic losses.

I.2.2 Uniform source constants

Set $a_\star = 19/20$, $\varrho_\star = a_\star^2/[a_\star^2 + (1 - a_\star)^2]$, $\rho_\star = \sqrt{\varrho_\star}$, $s_\star = \rho_\star/\sqrt{1 - \rho_\star^2}$, and

$$P_E = \phi(2)^2\Pr(|Z| \leq 1)^{p-2}, \qquad m_0 = 2\phi(1),$$

$$m_\star = 2\Phi(2s_\star) - 1, \qquad A_{\min} = \frac{P_E m_0^2}{2(p + 6 + 2Z_{\max})},$$

$$D_{\min} = \frac{P_E(1 - m_\star^2)}{2(p + 6 + 2Z_{\max})},$$

$$\gamma_s = \min\left\{\frac{D_{\min}}{2}, \frac{p(1-\varrho_\star)e^{-4Z_{\max}}5^{-p/2-2}}{3\pi}\right\}. \tag{I.4}$$

These are explicit positive real numbers. On $G_1, G_2 \in [1,2]$ with all other coordinates in $[-1,1]$, the probability is at least P_E and $\|G\|^2 + \zeta/S \le p + 6 + 2Z_{\max}$. Its correlation and remaining uncertainty give $A_{\min}, D_{\min}$ uniformly until $a_\star$. For $\alpha \le 1/2$, the first term of (I.3) lower bounds the unscaled decrease by $D_{\min}/2$. For $\alpha \ge 1/2$, restrict the integral to $u \in [1,2]$, and use $\Gamma(\alpha) \le 3$, $v \le 2Z_{\max}$ and $1 + 2u \le 5$ to obtain the second term in γ_s. Since $S^{1-\alpha} \ge 1/2$,

$$-L_\beta \ge t\gamma_s/2 \qquad (a \le a_\star). \tag{I.5}$$

The coefficient of t in f is bounded by $|G_1G_2| + 1/2$, whence $L_{tt} < 1$. At $t = 0$, $-L_t \ge A_{\min}/2$. Thus, with

$$t_{\min} = A_{\min}/8, \qquad d_s = 3A_{\min}/8,$$

we have $-L_t \ge d_s$ on $0 \le t \le t_{\min}$. The correlation-event bound also holds beyond $a_\star$ whenever $\beta \ge 0$, which will make $t_{\min}$ an invariant lower boundary.

I.2.3 Factor dynamics and the retained source law

Let $F_\epsilon(z) = z/(|z| + \epsilon)$ act coordinatewise on each row's scaled population gradient. Align $k' = Uk$. The signed-coordinate source symmetries give $M = \beta U$ and the head form above. Haar averaging in the active block and uniform contexts give the full gradient

$$\nabla_M L = \frac{L_\beta}{2n_{\text{ctx}}} U. \tag{I.6}$$

The head gradient has no extra $1/n_{\text{ctx}}$: each context has the same head loss. Each aligned query/key pair obeys

$$\dot q = F_\epsilon(\chi k'), \quad \dot k' = F_\epsilon(\chi q), \qquad \chi = -L_\beta/(2n_{\text{ctx}}).$$

Its product derivative is nonnegative for $\chi \ge 0$. Initially $q, k' \in [1,2]$ has probability at least $\phi(2)^2$; there both remain at least one, and their product derivative is at least one when $\chi \ge \epsilon$. The negative event gives the same contribution. Consequently

$$\dot\beta \ge p_0 := e^{-4}/\pi = 2\phi(2)^2 \qquad (\chi \ge \epsilon).$$

The active head coordinates (a_1, a_2, b_1, b_2, o) have alignment $t = \mathbb{E}[o(a_1b_2 + a_2b_1)]$. With $d = -L_t$, their equations are

$$\begin{aligned} \dot a_1 &= F_\epsilon(dob_2), & \dot a_2 &= F_\epsilon(dob_1), \\ \dot b_1 &= F_\epsilon(doa_2), & \dot b_2 &= F_\epsilon(doa_1), \\ \dot o &= F_\epsilon\big(d(a_1b_2 + a_2b_1)\big). \end{aligned}$$

Every alignment contribution has the form $uF_\epsilon(du)$. On the event that all five initial entries are in $[1,2]$, with probability at least $p_5 = \phi(2)^5$, all remain there or above during warmup, and

$$\dot t \ge 6p_5 \frac{d}{d+\epsilon}.$$

At $t = 0$ the direction is positive; at $t = t_{\min}$ it is strictly inward. Thus t reaches and then stays above $t_{\min}$, even if its derivative later changes sign. Choose

$$\epsilon_s = \frac{t_{\min}\gamma_s}{4n_{\text{ctx}}}, \quad S_{\text{warm}} = \frac{t_{\min}(d_s + \epsilon_s)}{6p_5 d_s}, \quad S_s = S_{\text{warm}} + \frac{\log 19}{p_0} + 1. \tag{I.7}$$

After warmup, (I.5) implies $\chi \ge \epsilon_s$ until $a_\star$, so matching reaches $a_\star$ by S_s. Thereafter β cannot decrease because $L_\beta < 0$ and $t > 0$. Every block remains trainable, and every normalized coordinate has speed at most one.

Sparse context support supplies the block decomposition. Simultaneous query/key sign flips in each block kill cross-context reference moments already at initialization. Their finite deviations remain in the full-core norm of Section I.6; acquired matching is within the supplied blocks.

At the switch, the n_{ctx} context head coordinates and k inactive cue coordinates retain their independent standard-Gaussian reference law, independent of the active five-coordinate evolution. Their population source gradients vanish; source context sample gradients vanish identically. The transformation $(o, a_1, b_1) \mapsto (-o, -a_1, -b_1)$ preserves source alignment and the coordinatewise flow. Hence o has a symmetric continuous law. Finite-time locally Lipschitz flow preserves absolute continuity, so $\Pr(o \ge 0) = 1/2$. This entire source-produced law is retained in the feedback proof.

I.3 Contextual feedback from the acquired state

Let β_s be the source contrast at S_s. The two source-active content coordinates are zero on all target inputs. Thus the reference source checkpoint has zero target logit without replacing any parameter. The starting law of this stage is exactly the law constructed in Section I.2.3.

I.3.1 The full target core and coordinate equations

The target-active head core has the invariant form

$$C_{\ell i} + C_{i\ell} = Bs_\ell \tau_i, \qquad C_{\text{context,context}} = C_{\text{cue,cue}} = 0, \tag{I.8}$$

where i ranges over cue coordinates. To see this for arbitrary s, conjugate context factors by the diagonal signs s and replace the uniform cue in context ℓ by $s_\ell z$. The quadratic logit is invariant under an overall input sign, reducing to all-positive s. The inactive source factors retain their Gaussian law under these transformations. Context permutations and signed cue permutations equate the aligned cross coefficients. Finally, negating o, both first source-channel factors and every cue factor preserves the source law; on target inputs it maps $f(z)$ to $-f(-z)$. Labels also change sign with the cue, so this symmetry kills even target blocks and preserves the cross block. All these transformations are signed permutations, under which F_ϵ is equivariant.

Let P_- project onto $(1,-1)$ and $P_+ = I - P_-$. The reference attention blocks are

$$M_{\ell\ell} = (\beta_s P_+ + \lambda P_-)U_\ell, \qquad c = \tanh(\lambda/2),$$

with common λ and zero cross-context blocks. A training pool has cue cvz, where $v \in \{\pm 1\}$ is the selected slot sign. Define

$$\begin{aligned} T &= \tau^\top z, & w(c) &= c\psi(1+kc^2), \\ J(B,c) &= \mathbb{E}\sigma(Bw(c)|T|), & E_T &= \mathbb{E}[|T|\sigma'(Bw(c)|T|)], \\ d_f &= w(c)E_T/k, & g_f &= (1-c^2)Bw'(c)E_T/2. \end{aligned} \tag{I.9}$$

Direct differentiation gives the full, unsymmetrized core gradients

$$(\nabla_C J)_{\ell i} = (\nabla_C J)_{i\ell} = \frac{d_f}{n_{\text{ctx}}} s_\ell \tau_i, \qquad (\nabla_M J)_{\ell\ell} = \frac{g_f}{n_{\text{ctx}}} U_\ell P_-. \tag{I.10}$$

Cue symmetry makes its aligned gradient proportional to τ; the scalar product with τ is $w(c)E_T$, giving $1/k$. Uniform contexts give $1/n_{\text{ctx}}$. The attention factor follows from two-slot softmax, query averaging and $dc/d\lambda = (1-c^2)/2$. There is no additional source factor $1/2$ in g_f.

For a head row define

$$A_c = \frac{s^\top a_{\text{context}}}{\sqrt{n_{\text{ctx}}}}, \quad D_c = \frac{s^\top b_{\text{context}}}{\sqrt{n_{\text{ctx}}}}, \quad X_c = \frac{\tau^\top a_{\text{cue}}}{\sqrt{k}}, \quad Y_c = \frac{\tau^\top b_{\text{cue}}}{\sqrt{k}}.$$

For the output factor, contracting the core gradient with its row feature gives the normalized row gradient

$$\begin{aligned} g_o &:= \langle \nabla_C J, ab^\top \rangle_F \\ &= \frac{d_f}{n_{\text{ctx}}} \left[(s^\top a_{\text{context}})(\tau^\top b_{\text{cue}}) + (\tau^\top a_{\text{cue}})(s^\top b_{\text{context}}) \right] \\ &= d_f \sqrt{\frac{k}{n_{\text{ctx}}}} (A_c Y_c + X_c D_c). \end{aligned}$$

Summing the coordinatewise equations with their signs therefore yields

$$\begin{aligned} \dot{A}_c &= \sqrt{n_{\text{ctx}}} F_\epsilon \left(\frac{d_f \sqrt{k}}{n_{\text{ctx}}} o Y_c \right), & \dot{D}_c &= \sqrt{n_{\text{ctx}}} F_\epsilon \left(\frac{d_f \sqrt{k}}{n_{\text{ctx}}} o X_c \right), \\ \dot{X}_c &= \sqrt{k} F_\epsilon \left(\frac{d_f}{\sqrt{n_{\text{ctx}}}} o D_c \right), & \dot{Y}_c &= \sqrt{k} F_\epsilon \left(\frac{d_f}{\sqrt{n_{\text{ctx}}}} o A_c \right), \\ \dot{o} &= F_\epsilon \left(d_f \sqrt{\frac{k}{n_{\text{ctx}}}} (A_c Y_c + X_c D_c) \right), & B &= \frac{\mathbb{E}[o(A_c Y_c + X_c D_c)]}{\sqrt{n_{\text{ctx}} k}}. \end{aligned} \tag{I.11}$$

These equations follow by summing the original coordinatewise updates. All contributions to $\dot{B}$ have sign d_f. At the switch, A_c, D_c, X_c, Y_c are independent standard Gaussians, independent of the symmetric sign of o. The event that these four variables are at least one and $o \geq 0$ has probability at least $p_g = \phi(2)^4/2$. These inequalities persist while $d_f > 0$. The o contribution alone implies

$$\dot{B} \geq \frac{2p_g}{\sqrt{n_{\mathrm{ctx}}k}} \frac{2d_f\sqrt{k/n_{\mathrm{ctx}}}}{2d_f\sqrt{k/n_{\mathrm{ctx}}} + \epsilon_f}. \tag{I.12}$$

The acquired contrast makes the right side positive; with $c = 0$ the training cue disappears.

For attention align $k' = Uk$ and set $u = q_1 - q_2$, $v = k'_1 - k'_2$ within a block. Then

$$\begin{aligned} \dot{u} &= 2F_\epsilon(g_f v/(2n_{\mathrm{ctx}})), & \dot{v} &= 2F_\epsilon(g_f u/(2n_{\mathrm{ctx}})), \\ \lambda &= \mathbb{E}[uv]/2, & \dot{\lambda} &= \mathbb{E}\left[vF_\epsilon(g_f v/(2n_{\mathrm{ctx}})) + uF_\epsilon(g_f u/(2n_{\mathrm{ctx}}))\right]. \end{aligned} \tag{I.13}$$

Thus $\dot{\lambda}$ has the sign of g_f. Sum coordinates are pointwise constant. Simultaneous coordinate exchange preserves vanishing mixed sum/contrast moments of the source law; their independence is unnecessary. Block sign symmetries and context permutations retain the complete attention-core form.

I.3.2 A NORMALIZATION BARRIER AND A HELD-OUT MARGIN

For every component,

$$\frac{d}{dc}\frac{c}{(1+kc^2+\zeta)^\alpha} = \frac{1+\zeta+(1-2\alpha)kc^2}{(1+kc^2+\zeta)^{\alpha+1}}. \tag{I.14}$$

It is nonnegative at $c_{\min} = 1/\sqrt{k}$. Equations (I.12) and (I.13), the source bound $c \geq 0.9$, and uniqueness give joint invariance of $B \geq 0$ and $c \geq c_{\min}$. Some members decrease contrast above that boundary; actual Adam deviations from the reference barrier will be paid separately.

In this region $w_{\min} \leq w(c) \leq 1$, where $w_{\min} = c_{\min}/(1+k+Z_{\max})$. Put

$$\begin{aligned} L_\star &= \log 19 + 1, & B_\star &= \frac{2(1+k+Z_{\max})L_\star}{c_{\min}}, \\ \mu_k &= \frac{\binom{k-1}{(k-1)/2}}{2^{k-1}}, & d_{\min,f} &= w_{\min}\mu_k e^{-B_\star k}/4, \\ \epsilon_f &= d_{\min,f}\sqrt{k/n_{\mathrm{ctx}}}, & S_f &= \frac{3\sqrt{n_{\mathrm{ctx}}k}B_\star}{4p_g} + 1. \end{aligned} \tag{I.15}$$

Conditioning one sign in the odd Boolean sum, or cancelling paired binomial terms, gives $\mathbb{E}|T|/k = \mu_k$. Since $|T| \leq k$ and $\sigma'(z) \geq e^{-z}/4$ for $z \geq 0$, we have $d_f \geq d_{\min,f}$ until B reaches $B_\star$. Then (I.12) gives $\dot{B} \geq 4p_g/(3\sqrt{n_{\mathrm{ctx}}k})$. Hence $B \geq B_\star$ by S_f and remains so.

For any two amplitudes from $\{\pm 1, \pm\rho_{\mathrm{mem}}\}$, let the correct weight be $a = (1+c)/2$ and the pooled cue multiplier be $z_0 = au_{\mathrm{selected}} + (1-a)u_{\mathrm{other}}$. Then

$$\begin{aligned} \mathrm{sign}(u_{\mathrm{selected}})z_0 &\geq a\rho_{\mathrm{mem}} - (1-a) = c - \frac{1+c}{2}2^{-m_{\mathrm{mem}}} \geq c_{\min}/2, \\ |z_0| &\leq 1. \end{aligned}$$

The signed logit is at least $Bc_{\min}/[2(1+k+Z_{\max})]$, since $|T| \geq 1$. The reference therefore has margin at least $L_\star$ on all sixteen such pairs, including the four disjoint test pairs. The bound

$$\left|\frac{d}{dz}\{z\psi(1+kz^2)\}\right| \leq 1 \tag{I.16}$$

holds componentwise and hence for the positive mixture, and will control finite core error.

I.4 THE FROZEN INTERVENTION AND SOURCE-INFORMATION BOUND

The frozen branch shares initialization and schedules with the main run but sets query/key updates to zero in both stages. Its reference $M = 0$ is an expectation identity; the actual finite initialization is

retained in the coupling. Source head symmetry still leaves only the parity-channel coefficients at uniform retrieval. No source-acquisition conclusion is needed for this branch.

On target training records its reference pool is exactly e_ℓ. Conditional labels are balanced, so

$$\mathbb{E}_y \sigma(yC_{\ell\ell}) = 1/2, \qquad \nabla_C J = 0.$$

All target-active reference coefficients vanish at the switch; the target logit is zero for every attention core there, and $\nabla_M J = 0$ as well. After the moment reset, every reference head row is stationary during feedback. Its logit stays zero on every test pair. Section I.7 bounds the finite branch with its random initial attention and trained head.

For the lower-bound direction of Theorem 5.1, pair worlds by independently flipping any subset of

$$(U_\ell, s_\ell) \longmapsto (U_\ell P_{\mathrm{swap}}, -s_\ell).$$

Every permitted antipodal training prompt has the same reward for each action in paired worlds. This pointwise equality survives adaptive prompts, randomized policies, parameter updates, arbitrary computation and any number of permitted feedback queries. Source records from other contexts are also unaffected.

If context ℓ is absent from N_{pre} iid source records, its flip remains uniform under an independent uniform prior on the n_{ctx} flips, conditional on the full transcript and supplied τ. Equal-sign test amplitude pairs have opposite paired labels, whereas opposite-sign pairs agree. Equal-sign mass is $1/2$, so a missing context has conditional test error at least $1/4$. It is missing with probability $(1 - 1/n_{\mathrm{ctx}})^{N_{\mathrm{pre}}}$. Averaging over uniform test contexts and then bounding the supremum by the prior mean proves the lower bound.

Receiving fewer than a worst-case cap of iid observations cannot improve on receiving the full cap. An expected stopping budget is different: a coupon-collector rule can wait until every context appears. Chosen-context source queries also change the experiment. The information result concerns unobserved orientations; the frozen-attention comparison concerns the specified neural training path (Section B).

I.5 A SUFFICIENT CONTRAST REGION AND ITS SIGNAL SCALE

The contextual reference path stays in $c \geq 1/\sqrt{k}$, and the sampled-path error budget controls its deviation. The reciprocal normalization below separates this sufficient invariant region from the presence of a learning signal.

Proposition I.2 (Nonmonotone signal within the reciprocal member)**.** *Let $k > 1$, $0 \leq c \leq 1$, and $\psi(u) = u^{-1}$, corresponding to a single mixture component with weight one, exponent one and shift zero. Then $w(c) = c\psi(1 + kc^2) = c/(1 + kc^2)$ increases strictly on $[0, 1/\sqrt{k}]$, decreases strictly on $[1/\sqrt{k}, 1]$, and has maximum $1/(2\sqrt{k})$. It is positive at every $c > 0$.*

Proof. The denominator is positive throughout the interval. The quotient rule gives

$$w'(c) = \frac{(1 + kc^2) - 2kc^2}{(1 + kc^2)^2} = \frac{1 - kc^2}{(1 + kc^2)^2}.$$

Its sign is positive for $c < 1/\sqrt{k}$, zero at that point and negative above it. Integrating its strict sign between distinct points on either side proves the two monotonicity statements. Substitution gives $w(1/\sqrt{k}) = (1/\sqrt{k})/(1+1) = 1/(2\sqrt{k})$. For $c > 0$, numerator and denominator are positive. Thus this member has a positive signal below $1/\sqrt{k}$, while the convergence argument uses the invariant region above it. □

I.6 ACTUAL SAMPLED ADAM AND THE CUBIC MOMENT MAP

We now justify the reference and control both finite runs. Within each stage, scalar gradients g_t generate

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t, \qquad v_t = \beta_2 v_{t-1} + (1 - \beta_2) g_t^2,$$
$$\widehat{m}_t = m_t/(1 - \beta_1^t), \qquad \widehat{v}_t = v_t/(1 - \beta_2^t), \qquad D_t = \frac{\widehat{m}_t}{\sqrt{\widehat{v}_t} + \epsilon}.$$

Source subtracts $\eta_t D_t$; feedback adds it. If $H_i = D\varphi_i^\top$ times the batch core gradient, physical query/key and head gradients are $H_i/\sqrt{d_{\text{width}}}$ and H_i/d_{width}, respectively. Homogeneity of both moments with the prescribed physical offsets yields exactly $\eta_t D_t$ in each normalized coordinate. This uses the physical first and second moments; it does not replace them by a population gradient.

I.6.1 A UNIFORM HISTORY BOUND AND ROW ENVELOPES

Define

$$K_A = \frac{1-\beta_1}{\sqrt{(1-\beta_2)(1-\beta_1^2/\beta_2)}}, \qquad \Lambda = \frac{\beta_1}{1-\beta_1} + \frac{\sqrt{\beta_2(1+\beta_2)}}{1-\beta_2}. \tag{I.17}$$

Weighted Cauchy–Schwarz gives $|D_t| \le K_A$ for every coordinate history. Before adding ϵ, the squared ratio is bounded by

$$\frac{(1-\beta_1)^2}{1-\beta_2}\frac{1-\beta_2^t}{(1-\beta_1^t)^2}\frac{1-(\beta_1^2/\beta_2)^t}{1-\beta_1^2/\beta_2}.$$

The inequality $(1-x)(1-y) \le (1-\sqrt{xy})^2$ bounds this by K_A^2, and also shows $K_A \ge 1$. A positive offset can only decrease the ratio. Thus all-zero rewards and a zero current gradient with nonzero old momentum are covered.

Put

$$\begin{aligned} &d_1 = 4n_{\text{ctx}}, && d_2 = 2(n_{\text{ctx}}+p)+1, && H = S_s + S_f + 2, \\ &v_0 = K_A\sqrt{\max(d_1,d_2)}, && a_0 = 1 + v_0 H, && \rho_i = a_0 + \|Z_i\|, \end{aligned}$$

where Z_i is a row's initial Gaussian vector. Every actual or reference row lies in its radius-ρ_i ball through time H. For dimension d define

$$\begin{aligned} \mu_3(d) &= 4\left[a_0^3 + 2^{3/2}\frac{\Gamma((d+3)/2)}{\Gamma(d/2)}\right], \\ \mu_6(d) &= 32[a_0^6 + d(d+2)(d+4)]. \end{aligned}$$

The two row moment maps, embedded in the direct-sum core, are

$$\varphi_1(q,k) = qk^\top, \qquad \varphi_2(a,b,o) = oab^\top.$$

Since $\rho_i \ge 1$, both satisfy

$$\|\varphi_i\| \le \rho_i^3, \qquad \|D\varphi_i\| \le 3\rho_i^2, \qquad \|D^2\varphi_i\| \le 6\rho_i. \tag{I.18}$$

For an allocation δ_c set

$$B_3 = \frac{4}{\delta_c}\{\mu_3(d_1) + \mu_3(d_2)\}.$$

Markov's inequality gives, with failure at most $\delta_c/4$,

$$\sum_{\text{groups}} \frac{1}{d_{\text{width}}} \sum_i \rho_i^3 \le B_3. \tag{I.19}$$

It bounds both actual cores, their coupled empirical reference cores and their speeds by $V = 3v_0 B_3$. Deterministic population cores satisfy the same envelope. Freezing query/key velocities only reduces it.

I.6.2 CORE REGULARITY AT SINGULAR NORMALIZERS

For $u > 0$, the mixture obeys

$$\psi(u) \le \max(1, u^{-1}), \qquad |u\psi'(u)| \le \psi(u), \qquad u^2\psi''(u) \le 2\psi(u).$$

For $A(d_{\text{width}0}) = \sqrt{\psi(d_{\text{width}0}^2)}$, differentiation gives $|A'| \le A/d_{\text{width}0}$ and $|A''| \le 6A/d_{\text{width}0}^2$. Hence $\|D_h g\| \le 2A$ and $\|D_h^2 g\| \le 9A/d_{\text{width}0}$; the latter follows by bounding $d_{\text{width}0}|A''| + 3|A'|$ in the derivative of $A(d_{\text{width}0})h$.

On source records let W be the maximum of the $2p$ Gaussian magnitudes, $Y = \sqrt{p}W$, and $X = Y/\|h\|$. Unit keys/query and softmax give

$$\|h\| \le Y, \quad \|D_M h\| \le 2Y, \quad \|D_M^2 h\| \le 8Y.$$

Therefore

$$\|g\| \le 1 + Y, \qquad \|D_M g\| \le 4(Y + X),$$
$$\|D_M^2 g\| \le 16(Y + X) + 36(YX + X^2).$$

On $\|C\| \le B_3$, conservative joint-core bounds are

$$\|Df\| \le C_1(1 + Y^2 + X^2), \quad \|D^2 f\| \le C_2(1 + Y^3 + X^3), \tag{I.20}$$

where $C_1 = 64(B_3 + 1)$ and $C_2 = 4096(B_3 + 1)$.

For $1 \le j \le 16$ define $m_j = 2^{j/2}\Gamma((j + 1)/2)/\sqrt{\pi}$. For $1 \le j \le 8$ put

$$Y_j = p^{j/2}(2p)m_j, \qquad X_j = p^{j/2}\sqrt{\frac{(2p)m_{2j}2^j}{\prod_{\ell=1}^{j}(p - 2\ell)}}. \tag{I.21}$$

These bound $\mathbb{E}Y^j$ and $\mathbb{E}X^j$ uniformly over attention cores. Conditional on the keys/query, h is isotropic Gaussian with variance $S_\pi = \sum_j \pi_j^2 \ge 1/2$. Cauchy–Schwarz bounds the product $W^j\|h\|^{-j}$ using its $2j$ moments; inverse chi-square integration gives the denominator. Independence of W and h is not used. The strongest denominator remains positive at $p = 17$.

Consequently, the population source gradient bound, its Lipschitz constant, and the sample-gradient second moment are bounded respectively by

$$G_s = C_1(1 + Y_2 + X_2), \qquad \sigma_s^2 = 3C_1^2(1 + Y_4 + X_4),$$
$$L_s = \sigma_s^2/4 + C_2(1 + Y_3 + X_3). \tag{I.22}$$

To justify derivatives at zero architectural shifts, first add a common auxiliary positive squared shift. The envelopes above are uniform in that shift. Along deterministic convergent core/shift sequences the limiting pool is nonzero almost surely, so sample derivatives converge. The eighth moments bound squared Hessians and give uniform integrability. Vitali convergence establishes continuity of expected derivatives and the C^2 limit on compact core sets. No sample-adaptive supremum of inverse radii is asserted; the algorithm uses the prescribed shifts.

Target pools have norm between one and $\sqrt{k + 1}$ for arbitrary M. Define

$$A_f = 64(B_3 + 1)(k + 1)^2, \qquad H_f = 4096(B_3 + 1)(k + 1)^3,$$
$$G_f = A_f/4, \qquad L_f = (A_f^2 + H_f)/4, \qquad \sigma_f^2 = A_f^2. \tag{I.23}$$

The actual sample gradient is $R_{\text{rew}}A\sigma(-Af)Df$. Its norm is at most A_f; its conditional action expectation is $y\sigma'(f)Df$, exactly the derivative of sampled success. Thus the stated variance bound applies to actual terminal rewards. Set

$$G = \max(G_s, G_f), \quad L = \max(L_s, L_f), \quad \sigma^2 = \max(\sigma_s^2, \sigma_f^2), \quad \epsilon_{\min} = \min(\epsilon_s, \epsilon_f).$$

I.6.3 Existence and averaged sensitivity of the reference

Write $\Psi = (M, C)$, and let $F = L$ in the source stage and $F = J$ in feedback. For a prescribed continuous core path, each reference row solves the locally Lipschitz ordinary differential equation (ODE) with velocity $\pm F_\epsilon(D\varphi_i^\top \nabla F)$; the sign depends on the stage. A frozen row projects its query/key velocities to zero. Bounded coordinate speed prevents finite-time divergence. For two row/core pairs, (I.18) gives

$$\|H_i(X) - H_i(Y)\| \le 6G\rho_i\|X_i - Y_i\| + 3L\rho_i^2\|\Psi_X - \Psi_Y\|. \tag{I.24}$$

Solving this inequality for each row before averaging gives a finite Volterra Lipschitz constant proportional to $\mathbb{E}[\rho^4 e^{6GH\rho/\epsilon_{\min}}]$. The map from a prescribed core path to the expectation of the evolved row moment is therefore a contraction on sufficiently short time intervals and maps the bounded core region into itself. Iterating these intervals constructs the unique self-consistent reference

on both stages. The same inequality proves uniqueness; the switch concatenates the paths with their row states retained. This justifies the symmetry arguments above.

Put $\lambda_0 = 6GH/\epsilon_{\min}$ and

$$E_4(d) = 8\,2^{d/2}[a_0^4 + 4d(d+2)]e^{\lambda_0 a_0 + \lambda_0^2},$$
$$B_4 = \frac{4}{\delta_c}\{E_4(d_1) + E_4(d_2)\}. \tag{I.25}$$

The inequality $e^{\lambda_0\|Z\|} \le e^{\lambda_0^2 + \|Z\|^2/4}$, a Gaussian tilt and $(a + d_{\text{width}0})^4 \le 8(a^4 + d_{\text{width}0}^4)$ give this moment bound. A second unconditional Markov event, with failure at most $\delta_c/4$, gives

$$\sum_{\text{groups}} \frac{1}{d_{\text{width}}} \sum_i \rho_i^4 e^{\lambda_0 \rho_i} \le B_4. \tag{I.26}$$

The averaged radius moment controls the cubic row maps without a width-dependent maximum-radius bound.

I.6.4 Adam histories and discrete errors

For any current vector h_0 and bias-corrected history, let b_{ts} be the normalized nonnegative second-moment weights. Subtracting fractions and using weighted reverse triangle inequalities gives

$$\|D_t - F_\epsilon(h_0)\| \le \frac{\|\widehat{m}_t - h_0\| + \sqrt{\sum_s b_{ts}\|g_s - h_0\|^2}}{\epsilon}. \tag{I.27}$$

The normalized first-moment lag mean is at most $\beta_1/(1-\beta_1)$, and the second-moment lag second moment is at most $\beta_2(1+\beta_2)/(1-\beta_2)^2$, producing Λ in (I.17).

Let η bound all deterministic steps. On (I.19), population row gradients at two indices in the same stage differ by at most η times their index distance times

$$A_i = 6Gv_0\rho_i + 3LV\rho_i^2.$$

If the fresh core batch error is at most ν, its row error is at most $3\rho_i^2\nu$. Equation (I.27) then bounds the actual direction error relative to the normalized current population gradient by

$$\frac{\eta\Lambda A_i + 6\rho_i^2\nu}{\epsilon_{\min}}. \tag{I.28}$$

The reference Euler local error is at most $\eta_t^2 A_i/(2\epsilon_{\min})$. Both moments reset at the boundary, so history estimates stay within their objective; accumulated parameter error remains. Shortening the final step improves these lag and Euler bounds.

Couple actual and reference rows using the same initial Gaussian row. Let $e_{i,\ell}$ be row discrepancy and Δ_ℓ the actual core error relative to its population reference. Then

$$e_{i,\ell+1} \le \left(1 + \eta_\ell \frac{6G\rho_i}{\epsilon_{\min}}\right) e_{i,\ell} + \eta_\ell \frac{3L\rho_i^2}{\epsilon_{\min}}\Delta_\ell + \frac{\eta_\ell}{\epsilon_{\min}}\left\{\eta(\Lambda + 1/2)A_i + 6\rho_i^2\nu\right\}.$$

Solve rowwise, multiply by $3\rho_i^2$, and average using (I.26). If the reference empirical-core discrepancy at all deterministic grid times is at most w, the result is

$$\Delta_\ell \le w + \sum_{j<\ell} \eta_j(K_H\Delta_j + R_H),$$
$$K_H = 9LB_4/\epsilon_{\min},$$
$$R_H = \frac{B_4}{\epsilon_{\min}}\left\{\eta(\Lambda + 1/2)(18Gv_0 + 9LV) + 18\nu\right\}. \tag{I.29}$$

Discrete Gronwall gives, including at the stage boundary,

$$\Delta_\ell \le (w + HR_H)e^{K_H H}. \tag{I.30}$$

The same argument controls the frozen branch with its own reference. Shared Gaussian radius events suffice. Its finite initial attention core is included in w; actual trained rows are not asserted to remain independent.

I.7 Explicit budgets and functional outcomes

Fix the task, architecture and Adam coefficients first. Equations (I.7) and (I.15) determine $\epsilon_s, S_s, \epsilon_f, S_f$, hence $H, \epsilon_{\min}$. Set $\delta_c = \delta/4$, and form all constants in Section I.6. They do not depend on width, step count, or a sampled maximum row norm.

Define $c_p = \Gamma(p/2)/[\sqrt{\pi}\Gamma((p-1)/2)]$ and

$$\begin{aligned}\kappa &= \min\left\{\frac{1}{256\pi}, \frac{t_{\min}}{160000c_p^2}, \frac{c_{\min}}{8}, \frac{1}{32(k+1+\sqrt{2k}B_3)}\right\},\\ \tau_{\text{core}} &= \kappa/4, \quad E_0 = e^{K_H H}, \quad w = \tau_{\text{core}}/(4E_0),\\ \nu &= \frac{\epsilon_{\min}\tau_{\text{core}}}{144HB_4E_0},\\ \eta &= \min\left\{1, \frac{\epsilon_{\min}\tau_{\text{core}}}{8HB_4E_0(\Lambda+1/2)(18Gv_0+9LV)}\right\}.\end{aligned} \tag{I.31}$$

Put $T_s = \lceil S_s/\eta \rceil$, $T_f = \lceil S_f/\eta \rceil$, and $T = T_s + T_f$. Use step η, shortening the final step of each stage to make its total time exactly S_s or S_f. Every step is positive. Choose integers, rounded up and at least one,

$$d_{\text{width}} \geq \frac{8(T+1)\{\mu_6(d_1)+\mu_6(d_2)\}}{\delta_c w^2}, \qquad B_{\text{mb}} \geq \frac{8T\sigma^2}{\delta_c \nu^2}. \tag{I.32}$$

These choices make the width and sample budget explicit for every fixed member.

At a deterministic reference time the core is an average of iid row moments with total variance at most $\{\mu_6(d_1)+\mu_6(d_2)\}/d_{\text{width}}$. Chebyshev and the union over $T+1$ times give error at most w with failure at most $\delta_c/8$ per branch. We apply this concentration bound before intersecting with the radius events, preserving row independence. Given each actual past, a fresh batch has variance at most σ^2/B_{mb} while (I.19) holds. Chebyshev and the update union bound give batch error at most ν, failing with probability at most $\delta_c/8$ per branch. The branches may share samples; no interbranch independence is needed.

The two moment events cost $\delta_c/2$, the two reference-grid events $\delta_c/4$, and the two batch events $\delta_c/4$. On their intersection, (I.30) and (I.31) give $\Delta \leq \tau_{\text{core}}/2$, in particular $\Delta \leq \kappa/4$, for both branches throughout the grid. Source parameter error is carried into feedback.

Initially $\|M\| \leq w$, so correct source attention is at most $e^{2w}/2 < 0.51$. At the source endpoint the reference odds are at least 19; core error at most κ gives actual odds at least $19e^{-2\kappa} > 9$. Matching is therefore at least 0.9 uniformly over unit source frames and queries.

Conditional on a fixed frame/query and trained parameters, each relevant pooled Gaussian channel has correlation at least $9/\sqrt{82}$ with its selected channel. Gaussian planar angles bound the parity-sign error by $2\arctan(1/9)/\pi$. Coefficient error changes this sign only if

$$|h_1 h_2|/\|h\|^2 \leq \kappa/t_{\min}.$$

The direction $h/\|h\|$ is uniform on the $(p-1)$-sphere; each of its first two coordinate densities is at most c_p. A union bound gives probability at most $4c_p\sqrt{\kappa/t_{\min}} \leq 0.01$. The radial multiplier is positive. The total source error is below 0.1.

At the source checkpoint the reference target-active coefficients vanish, giving actual target logit magnitude at most $(k+1)\kappa$. At the final endpoint the reference signed logit is at least $\log 19 + 1$. A core attention error Δ_M changes a softmax weight by at most $\Delta_M/2$ and the pooled scalar by at most Δ_M. Equation (I.16) bounds the resulting logit error by $kB_{\text{final}}\Delta_M$. Moreover,

$$\|C_{\text{ref}}\|_F^2 \geq bkB_{\text{final}}^2/2, \qquad kB_{\text{final}} \leq \sqrt{2k/n_{\text{ctx}}}B_3 \leq \sqrt{2k}B_3.$$

Coefficient error contributes at most $(k+1)\Delta_C$. The choice of κ makes the total final error and source-entry magnitude at most $1/32$. Hence final success exceeds $19/20$, entry success is at most $1/2 + 1/128$, and their difference exceeds $2/5$.

For the frozen branch, the reference head uses absent source channels and has zero target logit for every M. Its actual test logit, even on unequal-amplitude memories, has magnitude at most $(k+1)\Delta_C \leq 1/16$. The $1/4$-Lipschitz sigmoid gives sampled success at most $1/2 + 1/64$, proving the paired gap. This completes the continuous-input part of Theorem 5.2.

I.8 FINITE COORDINATE ALPHABETS

We fix all continuous budgets first, then specify a deterministic observation encoding. Write $N_s = B_{\mathrm{mb}} T_s$ and $N_f = B_{\mathrm{mb}} T_f$. Both branches can share source records, so a single source-data event suffices. Optimizer states remain real-valued.

I.8.1 A POOL LOWER BOUND UNIFORM OVER ATTENTION

Set $\delta_v = \delta/8$ and

$$s_0 = \sqrt{\frac{p-1}{e}} \left(\frac{\delta_v}{4N_s} \right)^{1/(p-1)}, \quad L_v = 2\sqrt{p + \log(2N_s/\delta_v)}, \quad m_v = s_0^2 / L_v. \tag{I.33}$$

For each p-by-two Gaussian value matrix, a chi-square Chernoff bound allocates probability at most $\delta_v/(4N_s)$ to each of two events: its first-column norm is below s_0, or the distance of its second column from the first-column span is below s_0. The first dimension is $p \geq p-1$, so the latter bound suffices for both. Their union costs $\delta_v/(2N_s)$. The Frobenius upper-tail event above L_v also costs at most $\delta_v/(2N_s)$. The QR determinant identity therefore gives $\sigma_{\min} \geq m_v$. Union over the N_s records yields, with failure at most δ_v,

$$\|V\|_F \leq L_v, \qquad \sigma_{\min}(V) \geq m_v, \qquad \|V\pi\| \geq m_v/\sqrt{2} \quad \text{for every simplex weight } \pi. \tag{I.34}$$

In the first expression, $\|V\|_F$ denotes the Frobenius norm of the two-column value matrix. The event controls adaptive weights without selecting or rejecting any record. A value perturbation of operator norm at most $m_v/2$ leaves every interpolation pool at norm at least $m_v/(2\sqrt{2})$.

Use an odd, sign-preserving saturating quantizer with spacing $\gamma = 2^{-q}$. It rounds each nonzero magnitude upward to the next grid point and leaves zero unchanged, with clipping level $\lceil L_{\mathrm{clip}}/\gamma \rceil \gamma$. Error within the clipping range is at most γ. It encodes source queries, keys and values, commutes with U, and preserves the parity bit. Embedded zeros remain exact. Rounded frames need not be orthonormal; we couple their paths to the continuous ones.

I.8.2 AN EXPLICIT GRAPH BOUND

Let $P = d_{\mathrm{width}}(6n_{\mathrm{ctx}} + 2p + 1)$, $d_h = n_{\mathrm{ctx}} + p$, $\epsilon_{\mathrm{dec}} = 1/100$, and define

$$\begin{aligned}
R_0 &= \sqrt{2\log(16P/\delta)}, \\
R_{\mathrm{bd}} &= 1 + R_0 + K_A H, \\
L_{\mathrm{clip}} &= \max\{L_v, 2, \sqrt{2\log(8p/\epsilon_{\mathrm{dec}})}\}, \\
h_0 &= \min\{1, m_v/(2\sqrt{2})\}, \\
V_{\mathrm{bd}} &= L_{\mathrm{clip}} + 2, \\
F_{\max} &= d_h^2 R_{\mathrm{bd}}^3 V_{\mathrm{bd}}^2 (1 + h_0^{-1})^2, \\
B_{\mathrm{prim}} &= \exp\big(100[1 + d_{\mathrm{width}} F_{\max} + d_{\mathrm{width}} R_{\mathrm{bd}}^2 + d_h V_{\mathrm{bd}}^2 + h_0^{-2} + Z_{\max} + m_{\mathrm{norm}}]\big), \\
N_{\mathrm{op}} &= 100 d_{\mathrm{width}} (3n_{\mathrm{ctx}} + p + m_{\mathrm{norm}} + 1)^2, \\
D &= P + 8n_{\mathrm{ctx}} + 2p + 20, \\
K_{\mathrm{graph}} &= (4B_{\mathrm{prim}})^{4N_{\mathrm{op}}}, \\
L_{\mathrm{graph}} &= d_{\mathrm{width}} D K_{\mathrm{graph}}, \\
L_{\mathrm{state}} &= (7n_{\mathrm{ctx}} + 3p)(R_{\mathrm{bd}} + 1)^2.
\end{aligned} \tag{I.35}$$

A Gaussian maximum event, failing with probability at most $\delta/8$, bounds all initialized coordinates by R_0. The Adam coordinate bound then keeps both continuous and encoded paths within R_{bd}. Scores have magnitude at most $16R_{\mathrm{bd}}^2$, since an embedded query/key has only two nonzero coordinates.

The graph may compute the full $2n_{\mathrm{ctx}}$-by-$2n_{\mathrm{ctx}}$ core, softmax, pool and head. The operation count covers scalar additions, products, reciprocals, exponentials and logarithms. Unaveraged core/head sums cost at most $d_{\mathrm{width}} R_{\mathrm{bd}}^2$ or $d_{\mathrm{width}} F_{\max}$. Evaluate a radial component as $\exp[-\alpha_j \log(u + \zeta_j)]$ and its mixture square root as the exponential of half its logarithm. On the interpolation domain,

$$h_0^2 \leq u \leq d_h V_{\mathrm{bd}}^2, \qquad \frac{1}{1 + d_h V_{\mathrm{bd}}^2 + Z_{\max}} \leq \psi(u) \leq \max(1, h_0^{-2}).$$

Thus no inverse small mixture weight is needed. The bound B_{prim} covers every primitive's value and first two derivatives, including softmax/logistic denominators and the normalizer logarithms. Induction over unary and binary nodes bounds scalar graph derivatives through order two by K_{graph}.

Normalized row gradients equal d_{width} times literal normalized-coordinate derivatives. Hence L_{graph} bounds their state/input Lipschitz constants in maximum norm and also logit differences. Batch averaging preserves this bound. Subtracting quadratic and cubic moment entries and taking Frobenius norms gives the core map's maximum-norm state bound L_{state}.

I.8.3 History coupling and the precision choice

For any two actual Adam histories, subtracting moment fractions and applying weighted reverse triangle inequalities yields

$$\|D_t(g) - D_t(g')\|_\infty \le \frac{1+K_A}{\epsilon} \max_{s\le t} \|g_s - g'_s\|_\infty. \tag{I.36}$$

Let E_t be the running parameter discrepancy while all sampled actions in a paired continuous/encoded history agree. With $A_0 = (1+K_A)/\epsilon_{\min}$,

$$\begin{aligned} E_{t+1} &\le (1+\eta_t A_0 L_{\mathrm{graph}})E_t + \eta_t A_0 L_{\mathrm{graph}}\gamma, \\ E_{\mathrm{final}} &\le \gamma\{e^{A_0 L_{\mathrm{graph}} H} - 1\}. \end{aligned} \tag{I.37}$$

Both paths reset moments at the switch while carrying accumulated parameter error. Define

$$E_\star = \min\left\{1, \frac{\kappa}{8L_{\mathrm{state}}}, \frac{\delta}{32N_f L_{\mathrm{graph}}}\right\}.$$

Choose an integer $q \ge m_{\mathrm{mem}}$ such that $\gamma = 2^{-q}$ is no larger than each of

$$\frac{1}{4}, \quad \frac{m_v}{2\sqrt{2p}}, \quad \frac{\kappa}{100(R_{\mathrm{bd}}+1)^2}, \quad \frac{\pi\epsilon_{\mathrm{dec}}^2}{96p^{3/2}L_{\mathrm{clip}}}, \quad \frac{E_\star}{2}e^{-A_0 L_{\mathrm{graph}} H}. \tag{I.38}$$

This is a finite, explicit choice after width and sample counts. Target amplitudes $0, \pm 1, \pm\rho_{\mathrm{mem}}$ are exact on the grid, source perturbations satisfy the interpolation condition, and parameter discrepancy is at most $E_\star/2$.

Couple actual policy actions by shared uniform random variables and stop at the first differing action. Until that event, target inputs coincide; the sigmoid Lipschitz bound makes its conditional mismatch probability at most $L_{\mathrm{graph}}E_\star/8$. Across both branches' $2N_f$ samples, the union costs at most $\delta/128$. On the complementary event all sampled labels/rewards coincide and (I.37) applies to the actual algorithms. Deterministic path Lipschitzness alone would not justify this on-policy step.

The continuous two-branch proof costs $\delta/4$, the maximum event $\delta/8$, the QR event $\delta/8$, and mismatches at most $\delta/128$. Their sum is $65\delta/128 < \delta$. Combined core discrepancy is at most $\kappa/4 + \kappa/16 = 5\kappa/16$. Rounded queries and keys alter a source score by at most $9R_{\mathrm{bd}}^2\gamma$, within the remaining matching slack. Target inputs are exact, so the endpoint margins in Section I.7 still apply.

I.8.4 Fresh source evaluation and observation costs

Source-bit accuracy on a fresh encoded record requires an additional sign argument. Condition on the trained head and fresh rounded keys/query; its attention weights are independent of fresh Gaussian values. Normalize a nonzero symmetric quadratic coefficient to Frobenius norm one. Some eigenvalue has magnitude at least $1/\sqrt{p}$. Conditioning on the other Gaussian coordinates and using the squared-Gaussian density gives, for every fixed center z_0,

$$\Pr(|h^\top C h - z_0| \le u) \le \frac{2}{\sqrt{\pi}} p^{1/4}\sqrt{u/S_\pi}, \qquad S_\pi \ge 1/2.$$

An interval of fixed length has maximal squared-Gaussian mass at zero. On all $2p$ magnitudes being at most M_{clip}, rounding changes a quadratic score by at most $p\gamma(2M_{\mathrm{clip}} + \gamma)$. Thus its sign-change probability is at most

$$4pe^{-M_{\mathrm{clip}}^2/2} + 2\sqrt{2/\pi}\, p^{3/4}\sqrt{\gamma(2M_{\mathrm{clip}} + \gamma)}.$$

Take $M_{\mathrm{clip}} = \sqrt{2\log(8p/\epsilon_{\mathrm{dec}})}$. Equation (I.38) makes this at most ϵ_{dec}. A positive radial multiplier preserves the greedy sign, including the zero-pool convention. Adding this error to the Gaussian-proxy source error still gives less than 0.1. This uniform argument is specific to quadratic heads; unrestricted finite-grid decoders are not covered.

Each coordinate has at most $2\lceil L_{\mathrm{clip}}/\gamma\rceil + 1$ symbols, requiring

$$b_{\mathrm{coord}} = \lceil \log_2[2\lceil L_{\mathrm{clip}}/\gamma\rceil + 1]\rceil$$

bits in a fixed-length encoding. A literal source record contains $8n_{\mathrm{ctx}} + 2p$ coordinate fields and one predicted bit. Public sparse serialization uses $2p + 6$ coordinate fields, a context index and the next bit; it does not reveal the selected-slot index. Numerical decoding is part of the interface and is not learned by a token embedding. The trained and frozen neural runs use this same alphabet. Deterministic encoding preserves the lower-bound direction of Theorem 5.1, since it cannot reveal an unvisited orientation. The equality and its attainability proof in Section E.5 use exact numerical source observations.

For one algorithm run, the resources are N_s source records and predicted bits, N_f generated binary actions, N_f verifier queries, and $T_s + T_f$ updates, with zero validation or selection queries. Coordinate bits are additionally charged by the chosen serialization. The normalizer evaluates m_{norm} components per pool and the network has P parameters. Optimization uses real-valued states; the resource counts additionally specify observation precision and normalizer evaluation. All conclusions of Theorem 5.2 now follow.

J THE SUPPLEMENTARY SGD PATH

J.1 SUPPLEMENTARY SGD LEARNING PATH

The sequential task and the two architecture members in (4.2) also admit a finite stochastic gradient descent (SGD) learning path. This supplementary result uses an unregularized source objective, a predetermined source endpoint and its own horizon and rate choices. Its probability and resource statements apply to this schedule; the Adam schedule is specified separately in Theorem 4.2. The following subsections provide the complete source, optimizer-specific coupling and feedback proofs, using the common initialization in Section G.

Source SGD uses a readout rate and a smaller strictly positive rate for every other block. Feedback uses the actual complete-trajectory terminal REINFORCE score with baseline zero and fresh minibatches. Control-output rows have predictable rate $\eta_t = \eta_0/p_{\max,t}$, where $p_{\max,t}$ is the maximum conditional control probability on one fixed training prompt. It is inspected before the fresh batch and uses no reward or test information. Here η_0 is the base feedback step and $\rho_R > 0$ the relative step multiplier for non-control blocks, which have rate $\eta_t\rho_R$. These rates are distinct from the task binding η. Batch denominators are unchanged even when all rewards are zero.

Write w_s, w_f for the source-trained and final parameters, B_{src}, B_R for source-record and feedback-trajectory batch sizes, and T_{src}, T for their update counts. In the bounds below, h is the prescribed per-decision approximation tolerance, ν the feedback gradient-noise tolerance, and A the feature-norm envelope (6 for the base architecture and 12 for its normalized variant), distinct from the generated control A in the task definition.

Theorem J.1 (Finite source acquisition and feedback improvement)**.** *Fix $q = 2^{2r+2}$, $r \geq 1$, $0 < c \leq 1/64$ and $0 < \delta < 1$. For either architecture member, choose the explicit nonempty horizon range (J.22), widths, initialization and rates of Sections G, J.2.4 and J.4. After T_{src} raw next-token SGD updates and T sampled REINFORCE updates of the same parameter vector, for every fixed world, with probability at least $1 - \delta$, simultaneously for every test prompt X,*

$$\min\{E(w_f, X), M(w_f, X), S(w_f, X)\} \geq 7/8, \tag{J.1}$$

$$\max\{E(w_s, X), M(w_s, X), S(w_s, X)\} \leq 1/q + nh, \tag{J.2}$$

$$\min_{J\in\{E,M,S\}} \big[J(w_f, X) - J(w_s, X)\big] \geq 3/4, \tag{J.3}$$

$$V(w_f, X) \leq 9c/8 + 4nh. \tag{J.4}$$

Here w_s is the predetermined source checkpoint, and $h = \nu/(128n^2A^2)$ with all constants given in the appendices. The budgets are

$$N_{\text{pre}} = NB_{\text{src}}T_{\text{src}}, \quad Q = B_R T, \quad N_{\text{generated}} = nQ, \quad N_{\text{updates}} = T_{\text{src}} + T. \tag{J.5}$$

There are zero source verifier calls and zero checkpoint/test selection. Rate inspections add T inference-only evaluations and HT prompt-token appearances; padding and full-context costs are listed in Section J.4.10. Under real arithmetic, the finite budgets are polynomial in expanded $q, H, c^{-1}, \delta^{-1}$ up to logarithms; the explicit exponents are given in the appendices.

J.2 Finite neural source acquisition

We give the complete source construction and its finite budgets. The feedback constants in Section J.4.2 and (J.83) are chosen first and depend only on q: $K = 12$, η_0, T, s, Λ, ν, B_0 and $B_{\exp}$. The horizon and source accuracy are chosen next, followed by widths, temperature, derivative envelopes and strictly positive slow rates. This order prevents a width–accuracy circularity. In this and the neural appendices, $N = 2H + 2$ is both the record length and padded context length; $n = H + 2$ is the number of generated decisions.

J.2.1 The finite family and its signed operator

Each $\phi_{(b,u)}$ is a permutation: recover $z = z' + u$ and then $s = s' + b + \theta_{u,z}$. For any fixed initial and desired final state there is exactly one control attaining that final state. Consequently

$$q^{-1} \sum_{b,u} P_{(b,u)} = \Pi, \tag{J.6}$$

where Π is uniform averaging on the state alphabet. The family has 2^{m^2} distinct labeled tables and needs no group-closure assumption. The subfamily $\theta_{0,0} \neq \theta_{0,v}$ for fixed $v \neq 0$ has 2^{m^2-1} elements and contains noncommuting operations: compare $\phi_{(0,0)}$ and $\phi_{(0,v)}$ at public coordinate zero.

For the Koopman convention $P_U f = f \circ \phi_U$ for a full control $U = (b, u)$, decompose $f(s, z) = f_+(z) + (-1)^s f_-(z)$. Averaging over b in $M_c = q^{-1} \sum_{b,u} (-1)^{\ell(u)} P_{(b,u)}$ kills the odd state-bit subspace. On the even subspace it is precisely the signed XOR convolution in Section E.2.4. Hence $\|M_c\| = 2^{-r} = \rho_c$ for every unknown table. With source noise $aP + (1 - a)\Pi$, the signed norm is $a\rho_c$. Transposing to forward distributions preserves the norm.

J.2.2 An exact all-target source reference

Consider first the unconditioned source distribution, with all controls independent uniform. The diagonal-lag quadratic kernel realized by the Gaussian feature construction in Equation (G.16) is

$$K(x, x') = k_{\text{base}}(x, x')^2 + 3k_{\text{base}}(x, x') + 2, \quad k_{\text{base}} = \sum_{l=0}^{N-1} w_l \mathbf{1}\{x_l = x'_l\}. \tag{J.7}$$

It has diagonal $K = 12$, singleton coefficients at least $3/(2N)$ and the lag-pair coefficient $k_{0H} \geq 1/N$. Here x_l denotes the token at lag l from the prediction position, with leading PAD tokens as needed. Define the two centered state-output functions

$$h_{1,y}(x) = \mathbf{1}\{x_0 \in \mathcal{U}, x_H \in \mathcal{S}\} \left(\mathbf{1}\{y = \phi_{x_0}(x_H)\} - \frac{1}{q} \right), \tag{J.8}$$

$$h_{2,y}(x) = \mathbf{1}\{x_0 \in \mathcal{S}, x_H \in \mathcal{U}\} \left(\mathbf{1}\{y = \phi_{x_H}(x_0)\} - \frac{1}{q} \right), \tag{J.9}$$

and set them to zero on nonstate output rows. Their categorical supports are disjoint. Zero-initialized population CE descent stays in the sum of mask/class logits and $\beta_1 h_1 + \beta_2 h_2$. Write $b_j = (e^{\beta_j} - 1)/(e^{\beta_j} + q - 1)$, let v_i be the correct state-class mass at state prediction i, and let v_E be the incorrect state-class mass at EOS. The simultaneous updates are exactly

$$\beta_1^+ - \beta_1 = \frac{\eta k_{0H}}{Nq^2}(a - v_1 b_1), \tag{J.10}$$

$$\beta_2^+ - \beta_2 = \frac{\eta k_{0H}}{Nq^2}\left[\sum_{i=2}^{H}(a - v_i b_2) - v_E b_2\right]. \tag{J.11}$$

In particular EOS contributes a negative term; there is no source grammar mask.

For completeness, the categorical-coordinate argument checks more than pairwise independence. At the first-state query, all input states and controls are jointly independent; centering both arguments retains only the lag pair $(0, H)$. At a repeated-state query, averaging its current control eliminates coordinates omitting that control; a surviving coordinate must pair it with the current state. At EOS, integrating the uniform S_0 makes $(U_2, \ldots, U_H, S_1, \ldots, S_H)$ independent of U_1. Coordinates omitting both U_1, S_0 vanish by averaging U_1. Among coordinates containing U_1, only its pair with S_H survives: an earlier state leaves a later control to average, and a control leaves uniform S_0. Among coordinates containing S_0 but not U_1, an earlier state leaves a later control, S_H leaves a suffix average followed by U_1, and a suffix control leaves another suffix control when $H \geq 3$. Thus the only semantic EOS contribution is (S_H, U_1), with factor q^{-2}. Mask/class residuals remain in their span because distinct individual record variables are pairwise independent uniform. The resulting identities are functions of categorical prefixes and therefore also hold on incorrect generated state histories with the same class/PAD pattern.

Let $h_q(a)$ be the entropy of a point mass mixed with the uniform q-state law at weights $a, 1-a$. The irreducible unconditioned entropy per optimized token is $[(H+1)\log q + H h_q(a)]/N$. It is positive. Define

$$\beta^* = \log\frac{1+(q-1)a}{1-a}, \quad M = \beta^* + \log(V/\epsilon), \quad P_l = \mathbf{1}\{x_l = \mathrm{PAD}\}. \tag{J.12}$$

The class logits

$$c_S = M(P_0 - P_H + P_{2H}), \quad c_U = M(P_H - P_0), \quad c_E = M(1 - P_{2H}), \quad c_{\mathrm{PAD}} = 0 \tag{J.13}$$

select the correct class at every raw target. Add $\beta^*(h_1 + h_2)$. The extra β^* in M pays for the active wrong-class state partition at EOS. A squared comparator-norm bound is

$$D^* \leq \frac{(5q+V+1)M^2}{k_{\min}} + \frac{2q(q-1)(\beta^*)^2}{k_{0H}}, \qquad k_{\min} \geq \frac{3}{2N}. \tag{J.14}$$

The constant part of c_E uses all lag-zero indicators. The comparator's CE excess is at most ϵ. Convex smooth descent at $\eta \leq 1/K$ and $T_{\mathrm{src}} \geq D^*/(2\eta\epsilon)$ gives last-iterate excess $\mathcal{E} \leq 2\epsilon$. For any invariant reference, if $D_j = \mathrm{KL}(T_a \| T_{b_j})$, the exact decomposition is

$$D_1 + (H-1)D_2 + \sum_{\text{stages}} -\log(\text{correct-class mass}) = N\mathcal{E}. \tag{J.15}$$

Pinsker's inequality gives $|b_1 - a| \leq \frac{q}{q-1}\sqrt{N\mathcal{E}/2}$ and $|b_2 - a| \leq \frac{q}{q-1}\sqrt{N\mathcal{E}/[2(H-1)]}$. The first-state transition has only one observation position per record and requires the stronger accuracy.

J.2.3 Conditioning on the training-word split

For a signed low-coordinate observable F, retaining r_S state values and r_{tag} of the tagged controls $U_2, \ldots, U_H$, Markov-product contraction gives

$$|\mathbb{E}_0[C_{\mathrm{suffix}} F]| \leq q^{r_S}(a\rho_c)^{H-1-r_{\mathrm{tag}}}\|F\|_\infty. \tag{J.16}$$

Indeed insert coordinate projectors for retained states. A fixed control has transition norm at most one, each averaged tagged control has norm $a\rho_c$, and U_1 supplies no signed contraction. Summing retained state values gives the factor q^{r_S}. Reference CE-gradient coordinates retain conservatively at most three tagged controls and three states. Their discrepancy is at most

$$\mu_{\mathrm{src}} = 2q^3(a\rho_c)^{H-4}. \tag{J.17}$$

The squared sum of categorical feature weights is at most V^2K, so the gradient discrepancy is at most $\mu_{\mathrm{src}} V^{3/2}\sqrt{K}$. Nonexpansiveness of the conditional population CE map couples it to the unconditioned reference, giving uniform logit error at most $s_{\mathrm{src}}\mu_{\mathrm{src}}V^{3/2}K$. Take $H \geq 5$ to retain an unused balanced tag also in operation-label terms. The reference risk is unchanged by conditioning, but the conditional Bayes entropy is lower by $\log 2/N$. The comparison therefore retains the conditional reference approximation gap of $\log 2/N$.

J.2.4 EXPLICIT NONEMPTY SOURCE BUDGETS

Let $V = 2q + 2$, $\alpha = -\log \rho_c$ and $0 < c \le 1/64$. Use $A = 6$ for the base architecture and $A = 12$ for the normalized member. With the feedback constants chosen first as stated above, put

$$C_\epsilon = \min\{c^2/1024,\ \nu e^{-B_{\exp}}/(2^{14}A^2)\}, \qquad C_e = \frac{\nu}{13824A^2(1+3As)}, \tag{J.18}$$

$$C_M = \log\frac{qV}{cC_\epsilon} + 4, \qquad C_B = \log(q/c) + 1, \tag{J.19}$$

$$C_D = 2(5q+V+1)C_M^2 + 6q(q-1)C_B^2, \qquad C_s = C_D/(2C_\epsilon) + 1, \tag{J.20}$$

$$K_{\mathrm{cond}} = 2C_s q^3 \rho_c^{-4} V^{3/2} K/C_e. \tag{J.21}$$

Choose any integer

$$H \ge \left\lceil \max\left\{5, (18/\alpha)^2, 2\max(0, \log K_{\mathrm{cond}})/\alpha\right\}\right\rceil. \tag{J.22}$$

Set $N = 2H + 2$, $n = H + 2$, $a = 1 - c/H$, and

$$\epsilon_{\mathrm{src}} = C_\epsilon/H^3, \tag{J.23}$$

$$\beta^* = \log(qH/c - q + 1), \tag{J.24}$$

$$M = \beta^* + \log(V/\epsilon_{\mathrm{src}}), \tag{J.25}$$

$$D^* = (2N/3)(5q+V+1)M^2 + 2Nq(q-1)(\beta^*)^2, \tag{J.26}$$

$$\eta_{\mathrm{src}} = 1/A^2, \tag{J.27}$$

$$T_{\mathrm{src}} = \left\lceil D^*/(2\eta_{\mathrm{src}}\epsilon_{\mathrm{src}})\right\rceil, \tag{J.28}$$

$$s_{\mathrm{src}} = \eta_{\mathrm{src}} T_{\mathrm{src}}, \tag{J.29}$$

$$h = \frac{\nu}{128n^2A^2}, \tag{J.30}$$

$$e_0 = \frac{h}{48(1+\sqrt{2}nAs)}. \tag{J.31}$$

All checkpoints are fixed before data and initialization. The reference satisfies $|b_j - a| \le 2\sqrt{3C_\epsilon}/H < c/(8H)$ for both j. Correct-class leakage is at most $N\mathcal{E} \le 6C_\epsilon/H^2 \le he^{-B_{\exp}}/2$. Using $\beta^* \le C_B H$, $M \le C_M H$, $D^* \le C_D H^3$ gives $s_{\mathrm{src}} \le C_s H^6$; also $e_0 \ge C_e H^{-3}$. Thus the conditioning discrepancy is at most e_0 provided $K_{\mathrm{cond}} H^9 e^{-\alpha H} \le 1$. Equation (J.22) ensures this because $9\log H \le 9\sqrt{H} \le \alpha H/2$ and $\log K_{\mathrm{cond}} \le \alpha H/2$.

Let $\delta_j = \delta/20$ for each of seven source estimates, and choose

$$B_{\mathrm{src}} = \left\lceil 16 s_{\mathrm{src}}^2 K^2 \log(2T_{\mathrm{src}}/\delta_j)/e_0^2\right\rceil, \tag{J.32}$$

$$W_H = \sqrt{V} + \sqrt{2\log(1/\delta_j)}, \qquad R_{\mathrm{src}} = W_H + \sqrt{2}As_{\mathrm{src}}, \tag{J.33}$$

$$R_{\mathrm{tot}} = R_{\mathrm{src}} + \sqrt{2}nAs, \qquad \zeta_{\mathrm{tot}} = \min\left\{1, \frac{e_0}{As_{\mathrm{src}}(AR_{\mathrm{src}} + \sqrt{2}) + R_{\mathrm{src}}}\right\}. \tag{J.34}$$

For the base member choose $\gamma = \min\{1/16, e_0/(25s_{\mathrm{src}}), h\}$ and $\zeta_{\mathrm{raw}} = \zeta_{\mathrm{tot}}$. For the normalized member include $\zeta_{\mathrm{tot}}/8$ in that minimum and set $\zeta_{\mathrm{raw}} = \zeta_{\mathrm{tot}}/(2\sqrt{2})$. These choices precede every width.

Put $\xi_{\mathrm{enc}} = \gamma/16$, $R = \lceil 8N\log(2N/\delta_j)\rceil$, $p = NV$, $k_{\mathrm{aug}} = p + 1$, $s_{\mathrm{sym}} = k_{\mathrm{aug}}(k_{\mathrm{aug}} + 1)/2$, and $D_{\mathrm{feat}} = p + s_{\mathrm{sym}} + 1$. Choose the smallest d divisible by R, with head width $d_h = d/R$, satisfying

$$d \ge 2048\xi_{\mathrm{enc}}^{-2}[p\log 9 + \log(8/\delta_j)], \tag{J.35}$$

$$d_h \ge \max\left\{1, \left\lceil \frac{4096N}{R\xi_{\mathrm{enc}}^2}[V\log 9 + \log(8N/\delta_j)]\right\rceil\right\}, \tag{J.36}$$

$$d \ge 4096\gamma^{-2}[D_{\mathrm{feat}}\log 9 + \log(8/\delta_j)], \tag{J.37}$$

$$d \ge 50[\sqrt{VD_{\mathrm{feat}}} + \sqrt{2\log(1/\delta_j)}]^2/e_0^2. \tag{J.38}$$

With $J = s_{\mathrm{sym}} \log 9 + \log(8/\delta_j)$, $A_0 = 1 + \gamma^{-2} J$, and $L_0 = \log(eA_0 k_{\mathrm{aug}}/(\delta_j \gamma))$, set

$$M_{\mathrm{ff}} = \left\lceil 2^{50} A_0^2 (k_{\mathrm{aug}} + L_0)^8 \right\rceil . \tag{J.39}$$

The parameter count is $P = 2dV + 4Rd_h d + 3dM_{\mathrm{ff}} + 2M_{\mathrm{ff}} + d + RN$, plus d for the normalized gain. Define

$$M_{\mathrm{init}} = \max\{1, \sqrt{2P \log(2P/\delta_j)}\}, \qquad Z_0 = M_{\mathrm{init}} + \sqrt{R} M_{\mathrm{init}}^3, \tag{J.40}$$

$$L_{\mathrm{att}} = 2\sqrt{R} M_{\mathrm{init}}^3 [1 + 2M_{\mathrm{init}}^3 (Z_0 + 1)], \tag{J.41}$$

$$\ell_{\mathrm{att}} = \min\{1/2, \zeta_{\mathrm{raw}}/L_{\mathrm{att}}\}, \qquad g_{\mathrm{gap}} = \delta_j/(RN^2), \tag{J.42}$$

$$t_{\mathrm{att}} = \min\left\{ \frac{g_{\mathrm{gap}}}{4M_{\mathrm{init}}^4}, \frac{g_{\mathrm{gap}}}{2\log((N-1)/\ell_{\mathrm{att}})} \right\}. \tag{J.43}$$

At $M_* = M_{\mathrm{init}} + 1$, put

$$Z_* = M_* + \sqrt{R} M_*^3, \tag{J.44}$$

$$D_z = 1 + \sqrt{R}[3M_*^2 + M_*^3 N(4M_*^3 + 1/t_{\mathrm{att}})], \tag{J.45}$$

$$D_{\mathrm{raw}} = D_z[1 + 2M_*^3 (Z_* + 1)] + 3M_*^2 (Z_* + 1)^2 + 1. \tag{J.46}$$

Use $D = D_{\mathrm{raw}}$ for the base member or $D = 2\sqrt{2} D_{\mathrm{raw}} + \sqrt{24}$ for the normalized member, and $r_{\mathrm{ball}} = \min(1, 1/D)$. The source readout rate is η_{src} and every other supplied parameter has rate $\eta_{\mathrm{src}} \rho_{\mathrm{src}}$, with

$$\rho_{\mathrm{src}} = \min\left\{ 1, \frac{r_{\mathrm{ball}}}{2\sqrt{2} D R_{\mathrm{src}} s_{\mathrm{src}}}, \frac{e_0}{D^2 \left[A R_{\mathrm{src}} (A R_{\mathrm{src}} + \sqrt{2}) s_{\mathrm{src}}^2/\sqrt{2} + \sqrt{2} R_{\mathrm{src}}^2 s_{\mathrm{src}} \right]} \right\}. \tag{J.47}$$

Every rate is strictly positive. There is no momentum state, reinitialization or source-score selection. The feedback rate schedule is separately prescribed in Equation (J.83).

Checking the shared bounds for SGD. Here $\xi_{\mathrm{enc}} = \gamma/16$, $D_{\mathrm{feat}} = p + s_{\mathrm{sym}} + 1$ and $\delta_{\mathrm{bias}} = \delta_{\mathrm{enc}} = \delta_{\mathrm{ffn}} = \delta_{\mathrm{norm}} = \delta_j$. The first three width inequalities above are precisely the premises of Theorems G.1 and G.2; the displayed FFN width supplies its remaining premise. The fourth width inequality and $1 + \gamma \le 2$ give initial coefficient-readout norm at most $e_0/5$ in (J.49). The temperature bounds preserve the bias winner gap and make the soft/hard error at most ζ_{raw}. For the normalized member, $\gamma \le \zeta_{\mathrm{tot}}/8$ and $\zeta_{\mathrm{raw}} = \zeta_{\mathrm{tot}}/(2\sqrt{2})$ give $\sqrt{2}\zeta_{\mathrm{raw}} + 4\gamma \le \zeta_{\mathrm{tot}}$; for the base member the raw bound already equals ζ_{tot}. Apply Theorem J.2 with $W_{\max} = W_H$, $R_W = R_{\mathrm{src}}$, $s = s_{\mathrm{src}}$, $r = r_{\mathrm{ball}}$, $\eta = \eta_{\mathrm{src}}$ and the chosen A (6 or 12). The second term defining ρ_{src} implies $\sqrt{2}\rho_{\mathrm{src}} D R_{\mathrm{src}} s_{\mathrm{src}} \le r_{\mathrm{ball}}/2$; the third bounds (J.53) by e_0. Thus both the ball and error premises hold for the actual common-batch path. Conditional encoder and coefficient bounds are integrated before taking the union; the seven source failure allocations remain $\delta_j = \delta/20$ each.

J.2.5 The source event and its consequence

The seven estimates proved in Sections G, J.3 and J.3.2 concern bias coverage/gaps, encoder frame, FFN/output Gram matrix, whole-initialization norms, initial coefficient readout, the full readout norm and fresh complete-record sampling. Their union failure is at most $7\delta/20$. The population reference, conditioned population, sampled ideal coefficients, actual hard features, soft or normalized fixed features, and moving all-parameter network are compared on the same batches. The total uniform source logit error is at most $6e_0 \le h$; feature error is at most $2e_0/R_{\mathrm{src}}$, readout norm at most R_{src}, and backbone displacement at most $r_{\mathrm{ball}}/2$. Since $R_{\mathrm{src}} \ge 1$ and $R_{\mathrm{tot}} + 1 \le 2R_{\mathrm{src}}(1 + \sqrt{2} nAs)$,

$$2e_0/R_{\mathrm{src}} \le h/[12(R_{\mathrm{tot}} + 1)] < h/[4(R_{\mathrm{tot}} + 1)]. \tag{J.48}$$

These estimates provide all entrance conditions used by the finite feedback proof. In particular the checkpoint was produced by the actual source algorithm, and the uniform prefix bound includes incorrect generated histories. Combined with the target failure allowance $\delta/2$, the joint failure is at most $17\delta/20 < \delta$ for each fixed unknown world.

The source reference selects its first control uniformly. Averaging the first transition over all controls produces a uniform state, and subsequent noisy permutations preserve uniformity. Class masses

depend only on the mask. Thus its endpoint, first-state and task-utility probabilities are at most $1/q$ on every fixed prompt. Sequential coupling gives at most $1/q + nh$ for the actual source sampler. Arbitrary decoders require the separate average-over-binding information bound in Section J.4.7.

The source uses $T_{\mathrm{src}}B_{\mathrm{src}}$ records, $NB_{\mathrm{src}}T_{\mathrm{src}}$ raw next-token targets, $NB_{\mathrm{src}}T_{\mathrm{src}}$ prepended PADs, zero verifier calls and T_{src} updates. Separate target evaluation processes N^2 context-token appearances per record, with $N(N+1)/2$ PAD appearances. Final feedback and rate-inspection accounting appears in Section J.4.10. The explicit bounds are finite polynomials in the expanded variables $q, H, c^{-1}, \delta^{-1}$ up to logarithms, with the exponents specified in the budget formulas. The task alphabet has finite-precision symbols; the neural optimizer is analyzed in real arithmetic.

J.3 SGD READOUT AND MOVING-FEATURE BOUNDS

Use the shared initializer and network in Section G, including the same Gaussian laws, encoder frame, realized coefficient operator, soft/hard approximation and derivative bounds. The remaining estimates below concern the SGD schedule.

Initial readout events. The readout W_0 is independent of C_{all}, with entry variance $1/d$. Conditional on the coefficient-operator event, Gaussian norm concentration gives

$$\|W_0 C_{\mathrm{all}}\|_F \le \sqrt{\frac{1+\gamma}{d}}\left[\sqrt{V D_{\mathrm{feat}}} + \sqrt{2\log(1/\delta_{\mathrm{head}})}\right] \tag{J.49}$$

except with probability δ_{head}. Increasing d thus gives a small but nonzero initial coefficient error, without setting the initial function to zero. The separate full-readout Gaussian tail gives

$$\|W_0\|_F \le \sqrt{V} + \sqrt{\frac{2\log(1/\delta_{\mathrm{head}})}{d}},$$

with its own failure allocation.

J.3.1 SOURCE COUPLING WITH EVERY PARAMETER TRAINABLE

Proposition J.2 (Deterministic shared-batch coupling)**.** *Write the complete logits as $W h_\psi(x)$, where ψ contains all trainable blocks except the readout W. On the Euclidean/Frobenius ball $\|\psi - \psi_0\| \le r$, suppose, uniformly over allowed prefixes,*

$$\|h_\psi(x)\| \le A, \qquad \|D_\psi h_\psi(x)\|_{\mathrm{op}} \le D. \tag{J.50}$$

Let $\|W_0\|_F \le W_{\max}$, $\eta \le A^{-2}$, $T < \infty$, $s = \eta T > 0$, and $R_W = W_{\max} + \sqrt{2}\,As$. The actual minibatch cross-entropy algorithm uses rate η on W and $\eta\rho$ on every block in ψ, with $\rho > 0$. Each complete-record loss is the arithmetic mean over its N prescribed next-token targets, and the minibatch loss averages complete-record losses. The reference uses the same batch sequence and W_0, holding features at h_{ψ_0}. Writing the batch loss as $\mathcal{L}_t$, the simultaneous updates are

$$\begin{aligned} W_{t+1} &= W_t - \eta \nabla_W \mathcal{L}_t(W_t, \psi_t), \\ \psi_{t+1} &= \psi_t - \eta\rho \nabla_\psi \mathcal{L}_t(W_t, \psi_t), \\ V_{t+1} &= V_t - \eta \nabla_W \mathcal{L}_t(V_t, \psi_0), \qquad V_0 = W_0. \end{aligned} \tag{J.51}$$

The reference is an analysis process, not the actual algorithm. If

$$\sqrt{2}\,\rho D R_W s \le r/2, \tag{J.52}$$

the actual path stays in the radius-r ball and satisfies

$$\begin{aligned} &\sup_x \|W_T h_{\psi_T}(x) - V_T h_{\psi_0}(x)\|_2 \\ &\qquad \le \rho D^2 \left[\frac{AR_W(AR_W + \sqrt{2})s^2}{\sqrt{2}} + \sqrt{2}\,R_W^2 s\right]. \end{aligned} \tag{J.53}$$

This holds deterministically for every shared batch sequence, without an additional source-sampling or test-word-distribution premise.

Proof. The softmax residual has norm at most $\sqrt{2}$, so $\|\nabla_W \mathcal{L}_t\|_F \le \sqrt{2}\,A$, and both readouts have norm at most R_W. The full non-readout gradient satisfies $\|\nabla_\psi \mathcal{L}_t\| \le \sqrt{2}\,DR_W$. Therefore

$$\|\psi_t - \psi_0\| \le \sqrt{2}\,\rho D R_W \eta t. \tag{J.54}$$

Induction with (J.52) prevents leaving the ball. A fixed-feature batch objective is convex and A^2-smooth in W, so its step is nonexpansive at the stated rate. At the same readout,

$$\|\nabla_W \mathcal{L}_t(W,\psi) - \nabla_W \mathcal{L}_t(W,\psi_0)\|_F \le D\|\psi - \psi_0\|(AR_W + \sqrt{2}). \tag{J.55}$$

Add and subtract the two residual–feature products and use softmax Lipschitzness to obtain this inequality. Summing the nonexpansive comparison gives

$$\|W_T - V_T\|_F \le \frac{\rho D^2 R_W (AR_W + \sqrt{2}) s^2}{\sqrt{2}}. \tag{J.56}$$

The final logit difference is at most $A\|W_T - V_T\|_F + R_W D\|\psi_T - \psi_0\|$, proving (J.53). Its feature-motion term contains D^2: the parameter gradient and the conversion of parameter motion to feature motion each supply one factor of D. □

Raw-network radius and rates. If initial block norms are at most M_{init}, evaluate D at $M_{\text{init}} + 1$. If the realized initializer satisfies $\sup_x \|h_{\psi_0}(x)\| \le A_{\text{init}}$, choose

$$r \le \min\{1, D^{-1}\}, \qquad A = A_{\text{init}} + 1. \tag{J.57}$$

The mean-value bound gives $\|h_\psi\| \le A$ throughout this ball, without using the much looser A_{raw} for the readout rate. Any prescribed $\varepsilon_{\text{net}} > 0$ is attained by the explicit positive choice

$$0 < \rho \le \min\left\{1,\ \frac{r}{2\sqrt{2}\,DR_W s},\ \frac{\varepsilon_{\text{net}}}{D^2[AR_W(AR_W + \sqrt{2})s^2/\sqrt{2} + \sqrt{2}\,R_W^2 s]}\right\}. \tag{J.58}$$

The initialization and source-budget construction specifies every quantity in this rate choice. The backbone-to-readout rate ratio ρ may be any positive value satisfying the displayed bound, while readout learning proceeds on the η time scale.

Normalized readout and accuracy choices.

A dimension-independent initial readout bound. Use a separate Gaussian norm event for the complete initial readout:

$$\|W_0\|_F \le W_H := \sqrt{V} + \sqrt{2\log(1/\delta_{\text{headnorm}})}. \tag{J.59}$$

The failure probability is at most δ_{headnorm}, since the sharper tail is $\sqrt{V} + \sqrt{2\log(1/\delta_{\text{headnorm}})/d}$. The bound is independent of d. Set $R_W = W_H + \sqrt{2}\,As_{\text{src}}$. The deterministic source/update couplings need this full-readout bound, including all nullspace components, rather than the larger all-parameter envelope M_{init}.

Using M_{init} in R_W would introduce a circular choice: normalization error depends on the coefficient-Gram accuracy, while M_{init} grows with the widths selected to improve that accuracy. Equation (J.59) avoids this problem. The raw derivative and attention-temperature bounds still use M_{init}.

Given source time s_{src} and a per-error budget e_0, set $A = 12$ and

$$\begin{aligned}
\zeta_{\text{total}} &= \min\left\{1,\ \frac{e_0}{As_{\text{src}}(AR_W + \sqrt{2}) + R_W}\right\},\\
\gamma &\le \min\left\{\frac{1}{16}, \frac{e_0}{25 s_{\text{src}}}, \frac{\zeta_{\text{total}}}{8}\right\},\\
\zeta_{\text{raw}} &\le \frac{\zeta_{\text{total}}}{2\sqrt{2}}.
\end{aligned} \tag{J.60}$$

Equation (G.26) is then at most ζ_{total}. These quantities precede d, M_{ff}, since R_W uses (J.59). The same actual coefficient-operator concentration proof applies to C_{all} at the smaller γ. Choose the

temperature from the raw attention bound using ζ_{raw}. The fixed-feature coupling handles the actual normalized map as a uniformly bounded perturbation, not as an exactly polynomial map.

Choose $r_{\mathrm{ball}} = \min\{1, D_{\mathrm{norm}}^{-1}\}$ and use D_{norm} in every source/feedback rate and stopping inequality.

Take

$$\eta_{\mathrm{src}} = A^{-2} = \frac{1}{144}, \qquad T_{\mathrm{src}} = \left\lceil \frac{D^*}{2\eta_{\mathrm{src}}\varepsilon} \right\rceil. \tag{J.61}$$

At the prescribed source precision, the explicit construction gives $s_{\mathrm{src}} \le C_s H^6$. The reference kernel and comparator remain at $K = 12$. All other source choices use the same formulas with $A, D_{\mathrm{norm}}, W_H, \zeta_{\mathrm{total}}$ substituted as indicated. No normalization second-derivative assertion is required: the finite feedback proof uses a fixed-backbone readout Hessian and pays the entire backbone update through a first-derivative trajectory bound.

J.3.2 Source probability and trajectory details

Semantic functions and reference geometry. We use the same h_1, h_2, D^*, μ_{src} and s_{src} as the source construction. Here the reference CE excess is denoted $\mathcal{E}$, while $E(w, X)$ remains the endpoint metric. For a state-output index y, define

$$\begin{aligned} h_{1,y}(x) &= \mathbf{1}_{\{x_0 \in \mathcal{U},\ x_H \in \mathcal{S}\}} \left(\mathbf{1}_{\{y = \phi_{x_0}(x_H)\}} - \frac{1}{q} \right), \\ h_{2,y}(x) &= \mathbf{1}_{\{x_0 \in \mathcal{S},\ x_H \in \mathcal{U}\}} \left(\mathbf{1}_{\{y = \phi_{x_H}(x_0)\}} - \frac{1}{q} \right). \end{aligned} \tag{J.62}$$

They are zero for nonstate outputs. At the first-state prediction, lag zero is U_1 and lag H is S_0. At prediction $i \ge 2$, these lags contain S_{i-1} and U_i; at the end-of-sequence prediction they contain S_H and U_1. Both functions vanish at the initial choice prediction because lag H is padding. The reference is its mask/class component plus $\beta_1 h_1 + \beta_2 h_2$. Its unconditioned Bayes entropy per optimized token is

$$\frac{(H+1)\log q + H h_q(a)}{N}, \qquad N = 2H + 2, \tag{J.63}$$

where $h_q(a)$ is the entropy of a point mass mixed with uniform at weights a and $1 - a$. Meaningful state histories, including inconsistent histories, share the same class/padding layout. The source coefficient identity controls their logits as functions, not merely on teacher prefixes.

The categorical singleton/pair map and tensor map Φ have the same kernel. For any finite linear combination of feature vectors, sending one representation to the other preserves its Gram quadratic form. The map is consequently well-defined and isometric between their spans. Project a comparator into the span if needed; this cannot increase its norm. This transfers the source zero-start coefficient descent and comparator bounds without equating arbitrary ambient parameter norms.

The unconditioned reference readout cross-entropy is convex and K-smooth, since logits are linear in Φ and $\|\Phi\|^2 = K$. For $\eta \le 1/K$, descent is nonexpansive and decreases risk. The squared-distance telescope to any comparator gives average excess over that comparator at most $D^*/(2\eta T)$. Monotonicity gives the last-iterate version. The comparator's own excess above irreducible entropy is at most $\varepsilon_{\mathrm{src}}$, so the prescribed source horizon yields

$$\mathcal{E} \le 2\varepsilon_{\mathrm{src}}. \tag{J.64}$$

Conditioning and complete-record sampling. For a signed low-coordinate observable, write its Markov transition product and insert a diagonal coordinate projector for each retained state. Fixed controls have norm at most one; each averaged tagged control has norm $a\rho_c$. The independent U_1 has norm at most one and supplies no signed contraction. Summing retained state values contributes the conservative factor q^{r_S}. After multiplying a singleton/pair feature, probability and label terms in a reference categorical gradient coordinate touch at most three retained controls and three states. Every normalized coordinate discrepancy is therefore at most

$$\mu_{\mathrm{src}} = 2q^3 (a\rho_c)^{H-4}. \tag{J.65}$$

The squared sum of categorical feature weights is at most $V^2 K$; there are V output coordinates. Thus the Frobenius gradient discrepancy is at most $\mu_{\mathrm{src}} V^{3/2}\sqrt{K}$. Apply nonexpansiveness of the

conditional population cross-entropy map to the unconditioned reference at every step to obtain uniform logit discrepancy at most

$$s_{\mathrm{src}}\mu_{\mathrm{src}}V^{3/2}K \le e_0. \tag{J.66}$$

This controls local observables of the reference trajectory, without asserting a full-history density-ratio bound.

A complete-record coefficient gradient has norm at most $\sqrt{2K}$, regardless of correlations among its N target positions. For each fresh independent record batch, a Hilbert-space bounded-difference estimate and a union bound over predictable T_{src} iterates give gradient noise at most

$$4\sqrt{\frac{K\log(2T_{\mathrm{src}}/\delta_j)}{B_{\mathrm{src}}}}. \tag{J.67}$$

The same conditional population map is nonexpansive, so the sampling logit discrepancy is at most

$$4s_{\mathrm{src}}K\sqrt{\frac{\log(2T_{\mathrm{src}}/\delta_j)}{B_{\mathrm{src}}}} \le e_0. \tag{J.68}$$

Independent batches consist of complete records; individual within-record tokens are not assumed independent.

Transport to realized and moving features. For $h_{\mathrm{hard}} = C_{\mathrm{all}}\Phi$, put $G = C_{\mathrm{all}}^{\top}C_{\mathrm{all}}$. The same-batch readout step induces

$$Z = WC_{\mathrm{all}}, \qquad Z^{+} = Z - \eta\nabla\mathcal{L}_{\mathrm{batch}}(Z)\,G. \tag{J.69}$$

Ideal coefficient descent uses $G = I$ and starts at zero. For $\eta \le [(1+\gamma)K]^{-1}$, the first map is nonexpansive in

$$\|D\|_{G^{-1}}^2 = \mathrm{tr}(DG^{-1}D^{\top}), \tag{J.70}$$

by convex smoothness after the invertible coordinate change. Each batch gradient has norm at most $\sqrt{2K}$, so

$$\begin{aligned}
&\sup_x \|(Z_T - Z_{\mathrm{ideal},T})\Phi(x)\| \\
&\quad\le \sqrt{K}\sqrt{\frac{1+\gamma}{1-\gamma}}\left[\|W_0C_{\mathrm{all}}\|_F + s_{\mathrm{src}}\sqrt{2K}\,\gamma\right] \\
&\quad\le 5\|W_0C_{\mathrm{all}}\|_F + 25s_{\mathrm{src}}\gamma \le 2e_0.
\end{aligned} \tag{J.71}$$

For two fixed feature maps within uniform distance ζ_{total}, starting from the same full W_0, the gradients at a fixed readout differ by at most $\zeta_{\mathrm{total}}(AR_{\mathrm{src}} + \sqrt{2})$. Their fixed-feature step is nonexpansive for $\eta \le A^{-2}$. Hence the fixed soft/normalized versus hard logit discrepancy is at most

$$\zeta_{\mathrm{total}}[As_{\mathrm{src}}(AR_{\mathrm{src}} + \sqrt{2}) + R_{\mathrm{src}}] \le e_0. \tag{J.72}$$

Proposition J.2 supplies the final moving-feature error at most e_0, and backbone displacement at most $r_{\mathrm{ball}}/2$. The error denominator in (J.58) also ensures feature motion at most e_0/R_{src}. Initial soft/normalized feature error is at most $\zeta_{\mathrm{total}} \le e_0/R_{\mathrm{src}}$, so the final feature error is at most $2e_0/R_{\mathrm{src}}$. The full W_0 norm, including nullspace components, is retained.

The joint probability event. Combine the seven source-event allocations by a union bound, without requiring mutual independence after conditioning. Maxima of iid continuous biases select each lag uniformly. A binomial Chernoff bound gives coverage failure at most

$$N\exp\left(-\frac{R}{8N}\right) \le \delta_j/2. \tag{J.73}$$

A difference of two $\mathcal{N}(0,1)$ biases has density at most $1/(2\sqrt{\pi})$. The probability that any of at most $RN^2/2$ pairwise gaps has magnitude below g is therefore at most

$$\frac{RN^2 g}{2} \le \delta_j/2. \tag{J.74}$$

The encoder/feed-forward concentration proofs condition only on the independent previously drawn blocks specified above. Each of the seven source events fails with probability at most $\delta_j = \delta/20$. Conditional on all source information on their joint event, fresh target sampling fails with probability at most $\delta/2$. Thus total failure is at most

$$7\delta/20 + \delta/2 = 17\delta/20. \tag{J.75}$$

Target concentration conditions on pre-batch history, never on future adaptive rates.

J.4 Finite neural feedback and information comparisons

This appendix continues the same parameter vector produced by the source analysis. Throughout, $N = 2H + 2$ is the source-record and context length, $n = H + 2$ is the number of generated decisions, and $L = H - 1$ retains its meaning as the displayed suffix length. The fixed kernel bound is $K = 12$. We write A for the feature-norm envelope ($A = 6$ for the base member and $A = 12$ for the normalized member); an operation selected by the policy is identified explicitly as a control token below.

J.4.1 The reference reward includes repaired hidden histories

Let $0 < c \le 1/64$ and suppose the source reference has

$$\left| b_j - \left(1 - \frac{c}{H}\right) \right| \le \frac{c}{8H}, \qquad j \in \{1, 2\}.$$

Set

$$B_{\mathrm{exec}} = b_1 b_2^{H-1}, \qquad d_j = b_j + \frac{2(1 - b_j)}{q}, \qquad D_{\mathrm{exec}} = d_1 d_2^{H-1}.$$

At a valid state step of type j, the probabilities of the correct local successor and its state-bit flip are respectively

$$\alpha_j = b_j + \frac{1 - b_j}{q}, \qquad \gamma_j = \frac{1 - b_j}{q}.$$

Their sum is d_j and their difference is b_j. Multiplying the two-state even/odd transition matrices shows that all even and odd state-bit-error histories have total masses $(D_{\mathrm{exec}} + B_{\mathrm{exec}})/2$ and $(D_{\mathrm{exec}} - B_{\mathrm{exec}})/2$. Thus, conditional on the selected control and correct output classes, the exact conditional expected rewards are

$$r_\eta = \frac{D_{\mathrm{exec}} + B_{\mathrm{exec}}}{2}, \qquad r_{\bar\eta} = \frac{D_{\mathrm{exec}} - B_{\mathrm{exec}}}{2}, \qquad r_a = 0 \quad (a \notin \{\eta, \bar\eta\}), \tag{J.76}$$

where $\bar\eta$ flips the binary control coordinate of η and preserves its public coordinate. A different public control coordinate cannot reach the target public endpoint while satisfying the public-coordinate consistency checks. The opposite state-bit choice can be repaired; its reward is explicitly included in (J.76). Conditional on the correct output classes, the public-coordinate checks pass with probability D_{exec}, independently of the selected control in this fixed reference. Product and union bounds give

$$r_\eta \ge B_{\mathrm{exec}} \ge 1 - \frac{9c}{8}, \qquad r_{\bar\eta} \le \gamma_1 + (H - 1)\gamma_2 \le \frac{9c}{8q}, \qquad 1 - D_{\mathrm{exec}} \le \frac{9c}{8}. \tag{J.77}$$

The transfer argument below separately pays class leakage, EOS, and feature/parameter drift. These fixed-executor equalities are not asserted to hold exactly for the moving network.

J.4.2 Population feedback with the realized predictable rates

Write $* = \eta$ and $\bar\eta$ for the two possibly rewarded choices. For $q \ge 16$, the ideal choice feature obeys

$$\|\Phi(x)\|^2 = K, \qquad \mathbb{E}_+\Phi = \mathbb{E}_-\Phi = \mu, \qquad \langle \mu, \Phi(x) \rangle = \|\mu\|^2 = k_\mu \in [2, K]$$

at every valid choice prompt. To obtain these identities, average the categorical singleton and pair features over the uniform initial state and the $H - 1$ suffix controls. Since $H \ge 5$, every coordinate leaves an unused balanced suffix tag. At degree at most two, all non-PAD coordinates are independent and uniform within their prescribed classes, so the kernel mean against any fixed valid choice prompt

is constant. The constant kernel term gives $k_\mu \geq 2$. This mean identity concerns choice prompts, not arbitrary later prefixes.

Starting from equal conditional control logits, define the reference coefficient increments by

$$\Delta_{t+1} = \Delta_t + \eta_t g(\Delta_t), \qquad g_a = p_a(r_a - J)\mu, \qquad J = \sum_a p_a r_a, \qquad \Delta_0 = 0. \tag{J.78}$$

Here $p = \mathrm{softmax}(\Delta_t \Phi(x))$ on the control class. Use exactly the same predictable rates as the actual network; do not differentiate the rate rule. Let

$$\eta_t = \frac{\eta_0}{p_{\max,\mathrm{actual},t}}, \qquad \eta_0 = \frac{1}{64K}.$$

The maximum is over *conditional* control probabilities on one fixed canonical training prompt, inspected before the fresh batch at step t. This inspection uses no reward, binding identity, or test information. Pathwise, $\eta_t \leq \eta_0 q$. Within a uniform actual/reference control-logit discrepancy of at most $1/16$, and provided the correct reference choice is maximal,

$$e^{-1/8}\eta_0 \leq \eta_t p_*^{\mathrm{ref}} \leq e^{1/8}\eta_0. \tag{J.79}$$

Put

$$\Delta_{\mathrm{gap}} = 1 - \frac{9c}{8}\left(1 + \frac{1}{q}\right) \geq \frac{31}{32}, \qquad L_q = \log(15(q-1)).$$

At every reference step with $p_* \geq 1/q$,

$$J \geq \frac{r_*}{q} > r_{\bar{\eta}}, \qquad z_a^+ - z_a = \eta_t k_\mu p_a (r_a - J).$$

The correct logit strictly increases and every wrong logit decreases. Consequently each correct-versus-wrong gap increases, and p_* remains maximal and nondecreasing. Until $p_* \geq 15/16$, $r_* - J \geq \Delta_{\mathrm{gap}}/16$, so every such gap increases by at least $k_\mu \eta_0 e^{-1/8}\Delta_{\mathrm{gap}}/16$. Therefore the predetermined number of updates

$$T = \lceil 640 K L_q \rceil \tag{J.80}$$

suffices for $p_*^T \geq 15/16$ for every $k_\mu \in [2, K]$. Indeed, a sharper sufficient bound is

$$\left\lceil \frac{16 e^{1/8} L_q}{k_\mu \eta_0 \Delta_{\mathrm{gap}}} \right\rceil.$$

The threshold persists after crossing. Only the correct gradient coefficient is positive, and all coefficients sum to zero. Hence

$$\|g\|_F \leq \sqrt{2k_\mu}\, p_*(r_* - J) \leq \sqrt{2K}\, p_*.$$

Using (J.79), a conservative bound on cumulative reference displacement is

$$B_0 = 2\sqrt{K}\, \eta_0 T. \tag{J.81}$$

It is $O(\log q)$ because $K = 12$ is fixed.

J.4.3 Local stability outside the reference subspace

For an arbitrary coefficient array U, define

$$F(U) = \mathbb{E}_+ \sum_a p_a (U\Phi(x)) r_a.$$

The source control-class logits are equal and cancel in this conditional softmax. With $v_a = p_a(r_a - J)$, the exact logit Hessian is

$$\nabla_z^2 J = \mathrm{diag}(v) - vp^\top - pv^\top.$$

Nonnegative rewards imply $\|v\|_1 \leq 2J$ and $\|\nabla_z^2 J\| \leq 6J$. In the convex tube where every control-logit vector differs from $\Delta_t \Phi$ by at most $1/16$,

$$J(U, x) \leq p_*(U, x) + r_{\bar{\eta}} \leq \left(e^{1/8} + \frac{9c}{8}\right) p_*^{\mathrm{ref}}, \qquad \|\nabla_U^2 F\| \leq 7K p_*^{\mathrm{ref}}.$$

For the *same fixed* η_t, the update map is therefore locally Lipschitz with factor at most

$$1 + 7K\eta_t p_*^{\mathrm{ref}} \le 1 + \frac{7e^{1/8}}{64} < \frac{9}{8}. \tag{J.82}$$

No derivative of $p_{\max}$ is taken. The bound allows arbitrary prompt-dependent coefficient perturbations. The reference path may itself depend on previous actual batches through the predictable rate sequence; the displayed inequalities hold pathwise for the realized sequence. Concentration below conditions only on the pre-batch filtration. Conditioning on the entire future adaptive sequence would be invalid. The fixed bounds $K = 12$ and $k_\mu \ge 2$ keep the horizon $O(\log q)$ and the amplification polynomial in q.

J.4.4 Source-stage bounds and positive slow rates

Set

$$\begin{aligned} s &= T\eta_0 q, \qquad \Lambda = (9/8)^T, \qquad \nu = \frac{1}{4096\sqrt{K}(1+s)\Lambda}, \\ h &= \frac{\nu}{128n^2A^2}, \qquad B_{\exp} = 2\sqrt{K}(B_0+1)+1. \end{aligned} \tag{J.83}$$

The source analysis, starting from initialization and raw minibatch SGD, supplies the following actual source event:

- An invariant reference f_{src} with $|b_j - (1 - c/H)| \le c/(8H)$ for $j = 1, 2$.
- Correct-class leakage at every valid source/generation mask at most $he^{-B_{\exp}}/2$.
- Uniform actual-source/reference logit discrepancy at most h on all legal-prefix strings, including incorrect state histories.
- A representation $h_{\mathrm{source}} = C\Phi + e_{\mathrm{source}}$ with $\|C^\top C - I\| \le \gamma \le h$ and $\sup\|e_{\mathrm{source}}\| \le h/[4(R_{\mathrm{tot}}+1)]$.
- $\|W_{\mathrm{source}}\| \le R_{\mathrm{src}}$, a feature/backbone derivative envelope D on the remaining radius $r_{\mathrm{ball}}/2$, and actual feature norm at most A throughout the initialization ball.

Here C is the initialized hard-feature analysis operator C_{all}, not the bilinear feed-forward factor with the same architectural name. It is never supplied to the learner. The deterministic radii satisfy

$$R_{\mathrm{tot}} = R_{\mathrm{src}} + \sqrt{2}\, nAs.$$

The source proof derives these bounds from initialization in the order specified in Section J.2.4.

Actual REINFORCE uses fresh independent training-suffix prompts and rollouts, the single binary terminal reward, baseline zero, and the complete generated-token score sum, including EOS. Control-output readout rows have rate η_t. Every other readout row and every backbone/gain parameter have the strictly positive rate $\eta_t\rho_R$. All rows and blocks update simultaneously. Let

$$G_{\mathrm{slow}} = \sqrt{2}\, n(A + DR_{\mathrm{tot}})$$

and prescribe

$$\rho_R = \min\left\{1, \frac{r_{\mathrm{ball}}}{2(1+s)G_{\mathrm{slow}}}, \frac{h}{4(1+s)G_{\mathrm{slow}}(A+DR_{\mathrm{tot}})}, \frac{h}{4D(R_{\mathrm{tot}}+1)(1+s)G_{\mathrm{slow}}}\right\}. \tag{J.84}$$

The source has already used at most $r_{\mathrm{ball}}/2$ of the backbone ball. A stopped pathwise induction bounds target slow-block displacement by $\rho_R s G_{\mathrm{slow}}$, using at most the remaining half-radius. Its direct noncontrol-logit and backbone-logit effects total at most $h/4$; feature displacement is at most $h/[4(R_{\mathrm{tot}}+1)]$. Every head update has norm at most $\eta_t\sqrt{2}\, nA$, because all block rates are at most η_t. Consequently $\|W_t\| \le R_{\mathrm{tot}}$ even without a favorable reward batch.

Define the projected actual fast-row increment

$$U_t = (W_{\mathrm{control},t} - W_{\mathrm{control,source}})C.$$

On each legal prefix the actual logits differ from $f_{\mathrm{src}} + \mathrm{control_rows}(U_t\Phi)$ by at most $3h$. This pays the source discrepancy, the fast increment times e_{source}, and the total slow/backbone effect. Moreover,

$$\|h_t^\top C - \Phi^\top\| \le \gamma\sqrt{K} + 2h \le 6h.$$

The projection does not discard head nullspace: its contribution is paid by the ambient head norm times the feature error.

J.4.5 TRANSPORT OF THE COMPLETE EXPECTED SCORE AND BATCH NOISE

Conditional on the past, the expected-gradient comparison is

$$\|\mathbb{E}[g_{\text{actual,fast}}C \mid \text{past}] - \nabla F(U_t)\| \le 32n^2A^2h \le \frac{\nu}{4}. \tag{J.85}$$

We give the full transport argument. As long as $\|U_t - \Delta_t\| \le 1/(128\sqrt{K})$, put $B = \sqrt{K}(B_0 + 1)$. Every added control logit has magnitude at most B. With source class leakage $\ell_0 \le he^{-B_{\exp}}/2$ and the paid $3h$ logit discrepancy, the actual wrong-class probability satisfies

$$\begin{aligned}\ell_{\text{actual}} &\le \frac{\ell_0}{1-\ell_0}\exp(2B + 6h)\\ &\le \frac{h\exp(-1+6h)}{1 - he^{-B_{\exp}}} \le h,\end{aligned}$$

using $h \le 1/12$ and $B_{\exp} = 2B + 1$. Thus both the odds denominator and all source/slow logit errors remain in the calculation.

Couple actual generation to the conditional control softmax $U_t\Phi$, the fixed b_1, b_2 conditional state executor, and forced EOS. The conditional-softmax error is at most $3h$ per decision, and class exits add at most h, so sequential coupling fails with probability at most $4nh$. The projected complete actual fast score has norm at most $2nA$; the ideal choice score has norm at most $\sqrt{2K} \le 2A$. On a matched correct-class path the choice-score error is at most $(6\sqrt{2} + 4\sqrt{K})h$. Feature-projection error supplies its first term; conditional probability and the control-class deficit supply its second. Every later state/EOS fast score is at most $2Ah$. The expected-gradient difference is therefore at most

$$\begin{aligned}&[8n(n+1)A + 6\sqrt{2} + 4\sqrt{K} + 2(n-1)A]h\\ &\quad\le 20An^2h \le 32n^2A^2h,\end{aligned}$$

which proves (J.85). Matched complete trajectories receive the same actual terminal bit, including repaired hidden-state histories. On unmatched trajectories the score envelopes pay the reward change. No hidden correct-trace constraint is inserted. In particular, the nullspace term is explicitly bounded by

$$\|\Delta W\| \le 2R_{\text{tot}}, \qquad \|\Delta W e_{\text{source}}\| \le h/2;$$

source and slow effects then total less than $3h$. An ambient head norm is never identified with its projection.

For a target failure allowance δ_R, take

$$B_R = \left\lceil \frac{256n^2A^2\log(2T/\delta_R)}{\nu^2} \right\rceil. \tag{J.86}$$

Conditional on the pre-batch filtration, the independent reward-weighted scores have norm at most $M_{\text{score}} = 2nA$, and their batch mean obeys

$$\mathbb{E}\|\text{batch mean} - \mathbb{E}[\text{score}]\| \le \frac{M_{\text{score}}}{\sqrt{B_R}}.$$

Bounded differences gives deviation at most

$$\frac{M_{\text{score}}[1 + \sqrt{2\log(T/\delta_R)}]}{\sqrt{B_R}} < \nu$$

with failure probability at most δ_R/T. A union bound gives simultaneous batch noise at most ν with probability at least $1 - \delta_R$. Batches with zero rewards are included. The denominator is the original B_R, without reward filtering. No independence between an iterate or its rate and earlier batches is asserted.

J.4.6 STOPPED COMPARISON AND FINAL METRICS

Compare actual and population updates using the same η_t and put $E_t = \|U_t - \Delta_t\|$. Inside the stopped tube, (J.82), (J.85), and the batch event imply conservatively

$$E_{t+1} \le \frac{9}{8}E_t + 8\eta_t\nu, \qquad E_0 = 0, \qquad E_t \le 8s\Lambda\nu < \frac{1}{512\sqrt{K}}. \tag{J.87}$$

This is stricter than the stopping threshold $1/(128\sqrt{K})$. Together with the $3h$ source/soft/slow discrepancy, it makes the actual/reference conditional choice-logit difference less than $1/16$, closing the rate comparison used in (J.79) and (J.82). The same induction closes coefficient and parameter exits. Rates are predictable before each batch. Since $K = 12$ and $T = O(\log q)$, Λ is a fixed power of q up to absolute constants.

At the predetermined final iterate the actual conditional correct-choice probability differs by at most $\sqrt{K}E_T + 3h$ from a reference probability of at least $15/16$, on every valid choice prompt, including held-out suffixes. In particular the uniform choice-logit discrepancy is less than $1/256$, so the comparison choice probability is at least $239/256$. The fixed executor has expected reward conditional on the correct initial control at least $1 - 9c/8 \ge 503/512$ and correct first-state probability at least the same lower bound. Final trajectory coupling costs at most $4nh < 1/1024$. Thus task utility and first-state memory success exceed

$$\frac{239}{256}\frac{503}{512} - \frac{1}{1024} > \frac{7}{8}. \tag{J.88}$$

Endpoint success dominates task utility.

In the source ancestral sampler, uniform choice makes the first state uniform. Every subsequent transition kernel preserves uniformity, while class masses depend only on position. Therefore the reference has $E_{\text{source}} \le 1/q$, $S_{\text{source}} \le E_{\text{source}}$, and $M_{\text{source}} \le 1/q$. Actual source coupling bounds all three by $1/q + nh$. Since $q \ge 16$ and $nh < 1/4096$, each final-minus-source gain exceeds $3/4$. The fixed executor satisfies the public-coordinate checks with probability D_{exec} independently of its choice, so actual grammar-or-public-coordinate violation risk is at most

$$1 - D_{\text{exec}} + 4nh \le \frac{9c}{8} + 4nh.$$

J.4.7 SOURCE-ONLY, DIRECT-FEEDBACK, AND SINGLETON-QUERY LIMITS

Arbitrary source-only decoders. Let η be uniform over the q controls, independently of all source observations, initialization, learner randomness, and evaluation prompts. For each fixed prompt the map

$$\eta \longmapsto \phi_{U_H} \cdots \phi_{U_2}\phi_\eta(S_0)$$

is bijective. Condition on every source-only observation and algorithmic random choice. Any endpoint equals exactly one of these q targets, so averaged endpoint success is at most $1/q$. Malformed output only reduces success. This includes arbitrary source-only computation, decoding, and calibration. It is an average-over-binding or minimax statement, not a pointwise assertion for every binding.

Indistinguishable direct-feedback worlds. Let $\ell(u)$ be the public tag bit. Pair the worlds by

$$\theta'_{u,z} = \theta_{u,z} \oplus \ell(u), \qquad \eta' = (b_\eta \oplus \ell(u_\eta), u_\eta).$$

The adjusted first transition is identical: $\phi'_{\eta'} = \phi_\eta$. The central state-bit flip B commutes with every transition map. Each suffix operation adds $B^{\ell(u)}$, so paired target endpoints differ by B raised to the suffix tag parity. They agree on every tag-zero training suffix and are opposite on every tag-one test suffix. Grammar and the public z-consistency checks are identical for every candidate trace in both worlds, including hidden errors that repair later. Hence their entire terminal training oracles are pointwise identical.

Without source observations, couple the learner's random tape. Every adaptive training transcript remains identical, even with unlimited training queries. On a common test prompt the identical output law has summed endpoint success at most one against the two opposite targets; average paired success is at most $1/2$. The pairing is an involution and preserves uniform η. It does not imply a first-state memory lower bound, because the paired first targets agree. Hidden-mechanism intermediate verification would invalidate this pointwise oracle equivalence and is outside the information model.

Singleton-query resource bound. Even grant θ for free. Endpoint bijectivity implies that any fixed prompt/candidate terminal query accepts for at most one η; malformed or public-inconsistent candidates accept for none. Fix learner randomness and follow the all-zero transcript. At most Q adaptive queries name at most Q distinct bindings. Every world producing a positive response belongs to these at most Q values, irrespective of subsequent behavior. All remaining bindings yield the same all-zero transcript. On each independent test prompt a final output succeeds for at most one remaining value. Therefore

$$\mathbb{E}_\eta[\text{test endpoint success}] \le \min\left\{1, \frac{Q+1}{q}\right\}. \tag{J.89}$$

In particular, averaged success at least $7/8$ requires $Q \ge 7q/8 - 1$. Each sampled candidate verification counts separately. Expected-reward oracles, intermediate labels, binding-dependent prompt side information, and test-time verification are excluded. This is a membership-query law, not a logarithmic entropy argument; the neural upper need not match this bound.

Packed-mechanism information bound. The separate proof in Appendix E.2.7 establishes, for $m \ge 512$, $H \ge 5$, independent reset source records, and at most Q adaptive binary pre-test replies,

$$\sup_{\theta,\eta} \mathbb{E}[\text{endpoint error}] \ge \frac{7}{32}\left[1 - \frac{128(n_{\mathrm{rec}}H + Q + 1)}{m^2}\right]_+.$$

It permits a fixed known binding and strengthens query permissions to adaptive prompt choice. The quantity $n_{\mathrm{rec}}H + Q$ counts information bits, not heterogeneous monetary or computation costs. It supplements the independent-binding singleton bound and the source-free paired-world ambiguity; its proof is not duplicated here.

J.4.8 Opaque-table recovery and fresh-prompt binding selection

Under the raw source law, execution controls $U_1, \dots, U_{H-1}$ are jointly iid uniform: U_1 is independent, and $U_2, \dots, U_{H-1}$ form a strict subset of the tag-conditioned suffix. Retain odd execution transitions through $H-1$ and average the intervening even controls. Their inputs become independent uniform states because $q^{-1}\sum_{U\in\mathcal{U}} T_U = \Pi$. Thus a complete record supplies $\lfloor H/2 \rfloor$ iid triples in this marginal retained experiment. Take

$$M_0 = \left\lceil 8q^2 a^{-2} \log \frac{2q^3}{\delta_{\mathrm{tab}}} \right\rceil, \qquad n_{\mathrm{rec}} = \left\lceil \frac{M_0}{\lfloor H/2 \rfloor} \right\rceil. \tag{J.90}$$

The charged raw-token count satisfies

$$N n_{\mathrm{rec}} = (2H+2) n_{\mathrm{rec}} \le 6M_0 + 2H + 2.$$

The empirical mode for every opaque state/control row recovers all q^2 successors with probability at least $1 - \delta_{\mathrm{tab}}$. This estimator requires neither a coordinate chart nor downstream goal labels.

For completeness, let $M = n_{\mathrm{rec}}\lfloor H/2 \rfloor$ be the number of retained iid triples and put $t = \log(2q^3/\delta_{\mathrm{tab}})$. Each row count has mean $\mu = M/q^2 \ge 8a^{-2}t$. Binomial lower-tail concentration bounds the probability of fewer than $\mu/2$ observations by $e^{-\mu/8}$. Conditional on n observations of a row, true-minus-competitor indicator differences lie in $[-1,1]$ and have mean a. Their centered log moment generating function is at most $\lambda^2/2$, so an incorrect empirical mode, including a tie, has probability at most $e^{-na^2/2}$. Union bounding the q^2 row counts and $q^2(q-1)$ comparisons gives failure probability at most δ_{tab}. This proves recovery with full- record rounding. It does not assume independence of all H transitions within a record.

On recovery, enumerate candidate bindings on separately drawn fresh uniform training prompts. Generate each candidate's exact path using the recovered table. The binary terminal reward equals one exactly when that candidate is η: suffix invertibility and sharp transitivity of the first transition prove the equivalence. Test at most $q-1$ candidates, stopping at success or inferring the last binding if all reject. No canonical-prompt query selection is needed.

The additional costs are at most $q-1$ verifier calls, $(H+2)(q-1)$ generated symbols, and H table lookups per candidate or deployment. There are zero gradient updates and no test-time oracle. Final storage is q^2 successor indices and one binding index; training may use q^3 counts. Deployment has perfect utility and public-coordinate compliance on the recovery event. This comparator uses the same observations with a table-based computation, separating information requirements from the neural training route. The $2H+2$ granularity term and every unused raw token remain charged.

J.4.9 GREEDY SOURCE EXECUTION WITH A BINDING CONTROLLER

On the proved source event, the conditional transition parameters obey $b_j \geq 1 - 9c/(8H)$, the wrong-class mass is at most $\ell_0 = he^{-B_{\exp}}/2 \leq h/2$, and actual source logits differ from the reference by at most h uniformly over legal-pattern prefixes. An individual softmax probability changes by at most h under this logit discrepancy: integrate its coordinate gradient, whose ℓ_1 norm is $2p(1-p) \leq 1/2$. The correct state token therefore has actual probability at least

$$(1 - \ell_0)b_j - h \geq 1 - \frac{9c}{8H} - \frac{3h}{2} > \frac{1}{2}.$$

The correct EOS probability is at least $1 - \ell_0 - h > 1/2$. The strict inequalities follow from $c \leq 1/64$, $H \geq 5$ and $nh < 1/4096$. After externally selecting any valid first control, greedy decoding consequently follows its unique exact state trajectory and then EOS, by induction over legal-pattern prefixes. This uses the uniform source function bound, not an assumption about teacher-forced visits.

That candidate satisfies the public-coordinate consistency checks. Suffix invertibility and sharp transitivity make its endpoint equal the target if and only if its first control equals η. At most $q - 1$ fresh-prompt verifications therefore identify the binding, with the last candidate inferred if all others reject. The same corrected choice gives the target first state and exact execution on every test prompt. Success is deterministic on the source event, whose probability is at least $1 - 7\delta/20$ for each fixed world.

The comparator pays the original source tokens and updates. Additional costs are at most $q - 1$ verifier queries, $n(q-1)$ generated symbols including the externally supplied control, $H(q-1)$ prompt tokens and $(H+1)(q-1)$ model forward evaluations, each retaining N context tokens. There are no additional optimizer updates. The binding index and enumeration counter are additional persistent controller state; model parameters, inference context and forward workspace are separate. Each verification still performs H trusted transitions and public checks. Deployment uses $H + 1$ forwards and n output symbols, with no verification. This comparator combines the frozen source model with an external binding search.

J.4.10 COMPLETE RESOURCE ACCOUNTING AND LIKELIHOOD EQUIVALENCE

No model updates occur after the fixed source and target budgets, and no test score selects a checkpoint. A canonical rate-inspection prompt may use a fixed state and $H-1$ copies of the publicly tag-positive zero control. Before each fresh rollout batch, read its conditional control probabilities without generation or verification. This charges T inference-only evaluations, HT prompt-token appearances, NT processed context-token appearances, and nT PAD appearances, separately from zero data-selection queries and from optimizer updates.

Each source record prepends N PADs and trains on N non-PAD targets. If predictions are separately evaluated, it processes N^2 context-token occurrences with $N(N+1)/2$ PAD appearances. Each feedback rollout starts its N-token context with $n = N - H$ PADs and H real prompt tokens. Its n decisions process nN context-token occurrences, including $n(n+1)/2$ PAD appearances; no real token is truncated. Across $Q = B_R T$ rollouts, charge

$$nQ \text{ prepended PADs}, \qquad nNQ \text{ context-token occurrences},$$
$$\frac{n(n+1)Q}{2} \text{ attended PAD occurrences},$$
$$HQ \text{ prompt tokens}, \qquad nQ \text{ generated decisions}.$$

Canonical inspections are additional. Optimization uses real arithmetic and stores P real parameter scalars. Each verifier call evaluates H unknown transition-table entries for the target endpoint and H public-coordinate comparisons. The verifier reveals the prescribed bit; its private θ table remains part of the verifier's resources.

For a fixed complete sampled batch and current model, baseline-zero REINFORCE ascent is exactly the negative gradient of

$$-\frac{1}{B_R}\sum_b R_b \sum_{j=1}^{n} \log \pi_w(Y_{b,j} \mid \text{prefix}_{b,j}). \tag{J.91}$$

Hold the realized rewards and trajectory tokens fixed for this single differentiation. With identical simultaneous block rates and the same canonical inspection rule, reward-weighted likelihood makes exactly the same update. Fresh coupled batches preserve pathwise equality; independently generated fresh batches give equality in law. This does not identify arbitrary expert SFT, filtered-denominator training, or repeated off-policy replay with this algorithm. Source-only continuation remains independent of the binding, so the arbitrary-decoder average bound still applies. Test-time best-of-N scored by the task verifier adds post-test-prompt oracle calls outside the pre-test lower bounds and cannot be counted as free.