WHERE SHOULD LEARNING LIVE? REGRET ANALYSIS FOR TEXT, PARAMETER, AND JOINT ADAPTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Language-model systems can select prompts and parameter adapters repeatedly during deployment, using feedback to guide later choices. Earlier choices may change the context or environment inherited by later configurations. We study selection among preconstructed candidates and compare cumulative loss with the best fixed prompt–parameter pair used from the beginning. This excess loss separates into three terms: selection regret within the available family, the gap from restricting that family, and the benefit or harm of the actual deployment history. Selection is evaluated using the history each candidate would have produced if used throughout, which separates candidate quality from the effects of switching. When loss functions are fixed independently of selector randomness, past choices affect loss for only a bounded number of rounds, and every candidate's own-history loss is revealed, limiting switches yields sublinear expected regret within the selected family. We also quantify the costs and additional assumptions of independent probing and deployment-only feedback. Controlled experiments show that fewer switches can either lower or raise deployed loss, depending on whether inherited histories are harmful or helpful. With Qwen3.5-9B on Spider, text-only selection has lower within-family regret than joint selection, yet produces more failures because its candidate family excludes stronger configurations.

1 INTRODUCTION

Language-model agents edit code, execute tests, and use tool feedback (Yang et al., 2024b), while recent benchmarks study learning across sequences of related tasks (Asawa et al., 2026). Consider a coding agent that handles successive tasks in a persistent workspace. Before each task, a selector chooses an instruction prompt and a model checkpoint or adapter. The agent edits code, runs tools, and receives feedback such as whether its changes pass the task's tests. Feedback guides the next configuration choice, while the edited code and tool history remain in the workspace. The next configuration therefore inherits the consequences of earlier choices. *How should we compare text-only, parameter-only, and joint adaptation in such a system?*

Each opportunity to select and execute a configuration is a *deployment round*, indexed by t. For example, a round can be one coding task, one question in Spider (Yu et al., 2018), or one dialogue turn in SParC (Yu et al., 2019). We consider a sequence of T rounds, which can share state and need not be independent. The selector uses earlier feedback to choose a configuration, executes it, incurs a loss, and uses the available feedback to update future selection probabilities. Minimizing cumulative loss over these repeated choices gives an online learning problem (Cesa-Bianchi & Lugosi, 2006). Recent methods construct configurations through reflective prompt optimization (Agrawal et al., 2026), joint prompt–parameter optimization (Bo et al., 2026), and alternating harness–weight optimization (Kim et al., 2026). In our setting, candidates are constructed before evaluation. During evaluation, online learning updates their selection probabilities without constructing new prompts or adapters.

A *configuration* pairs reusable text, such as prompts, demonstrations, and retrieval contents, with deployable parameters, such as checkpoints or adapters (Brown et al., 2020; Lewis et al., 2020; Hu et al., 2022). Text-only selection keeps the base parameters, parameter-only selection keeps the base text, and joint selection can change both. Figure 1 illustrates these configuration families. The workspace, generated history, and tool sessions constitute *operational state*, which evolves even

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.

under a fixed configuration. Retaining feedback across trials can influence later behavior without changing model parameters (Shinn et al., 2023).

Our comparator is the best available joint configuration used for all T rounds from the common initial condition, with its own resulting history. This own-history comparison follows the policy-regret perspective (Arora et al., 2012). For example, a coding agent that switches to configuration b after using a can encounter different code from an agent that used b throughout.

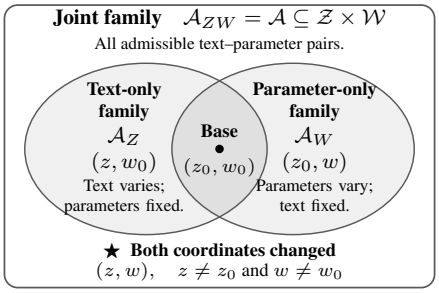


Figure 1: Text-only and parameter-only families within the joint family.

This comparison raises three separate questions. (i) *How well does the selector choose among its available candidates, when each is evaluated on the history it would have produced if used throughout?* (ii) *How much better is the best joint candidate than the best candidate in the selected family?* (iii) *Does the actual sequence of choices create a more or less favorable history than those fixed-configuration evaluations?* We call the corresponding terms *diagonal selection regret*, *family restriction*, and the *signed state effect*. Their sum is exactly the deployed excess loss. *Diagonal* refers to repeating a single configuration throughout the history, and *signed* emphasizes that inherited state can reduce as well as increase loss.

Our strongest feedback model evaluates every candidate on its own fixed-configuration history. Ordinary deployment feedback need not supply those counterfactual losses. We also study independent probes of individual candidates and updates based only on deployed outcomes. For state-dependent guarantees, we require *bounded influence*: choices sufficiently far in the past no longer affect current loss, including through hidden state. These assumptions connect our analysis to online learning with memory (Anava et al., 2015). We also compare joint sampling and coordinatewise conditional selection: they agree for additive losses under a product prior, but interactions can separate their regret even when both eventually select the optimum. The online-learning methods are established (Geulen et al., 2010; Avner et al., 2012; Cheng et al., 2023). Our analysis connects their guarantees to adaptation comparators, information requirements, and signed state effects.

Controlled studies show that reducing switches helps when mixed histories are harmful and hurts when they are beneficial, even with the same per-round selection probabilities. With Qwen3.5-9B (Qwen Team, 2026) on Spider, text-only selection incurs less within-family regret than joint selection but more deployed loss because its candidates are weaker. On recursive SParC, earlier generated queries in Structured Query Language (SQL) supply the inherited state. Lazy exponential weights (EW) substantially reduces switching with only a small change in estimated expected loss.

Our main contributions are summarized as follows.

- *Loss accounting.* We separate selection, family restriction, and state effects relative to a common fixed joint comparator; see Section 2.
- *Selection and state guarantees.* We connect selection bounds to configuration families and bound expected excess loss for a selector that limits switches under exogenous losses, bounded influence, and full counterfactual feedback; see Sections 3–4.
- *Weaker feedback.* We quantify the costs and additional assumptions of approximate evaluation, sparse probes, and deployment-only feedback; see Section 5.
- *Empirical comparisons.* Controlled studies and SQL evaluations show how candidate availability and inherited state change the interpretation of selection regret and switching; see Section 6.

2 Configurations, Comparators, and Loss Decomposition

Configurations and Candidate Families. A protocol Ξ declares the executor, available data, resource budgets, and permitted updates. An *admissible* configuration is one allowed and executable under these rules. Write an admissible configuration as $a = (z, w) \in \mathcal{A} \subseteq \mathcal{Z} \times \mathcal{W}$, where z is reusable text and w specifies deployable parameters. Learned continuous prompt tensors also belong to w. Relative to a base pair (z_0, w_0), the text-only family $\mathcal{A}_Z = \mathcal{A} \cap (\mathcal{Z} \times \{w_0\})$ holds parameters fixed. The parameter-only family $\mathcal{A}_W = \mathcal{A} \cap (\{z_0\} \times \mathcal{W})$ holds text fixed. Their

union is $\mathcal{A}_\cup = \mathcal{A}_Z \cup \mathcal{A}_W$, while the joint family is $\mathcal{A}_{ZW} = \mathcal{A}$. We select from finite candidates $\mathcal{C}_{ZW} \subseteq \mathcal{A}$ constructed before evaluation and containing the base pair, with $\mathcal{C}_e = \mathcal{C}_{ZW} \cap \mathcal{A}_e$ for $e \in \{Z, W, \cup, ZW\}$.

One Deployment Round. At round $t \in \{1, 2, \ldots, T\}$, the selector chooses $A_t \in \mathcal{C}_e$ using the feedback available from earlier rounds. The executor applies A_t to task input x_t in operational state s_t, producing outputs and tool actions, incurring a normalized loss $\ell_t \in [0, 1]$, and advancing the state to s_{t+1}. Loss can indicate task failure, while state can retain generated code or dialogue history. The permitted feedback updates future selection probabilities, while execution changes the operational state. Incurring a loss does not imply observing every candidate's loss. The formal executor and state transition are provided in Appendix A.1.

Fixed-Configuration Histories and Feedback. Fix a common initial condition and an exogenous random tape, independent of selector randomness, determining losses $F_t(a_{1:t}) \in [0, 1]$ for histories $a_{1:t} = (a_1, \ldots, a_t)$. For the deployed history, $\ell_t = F_t(A_{1:t})$. Each alternative history reruns the state and any deployment-responsive tasks. The *diagonal loss* of a fixed configuration a is

$$d_t(a) = F_t(a, \ldots, a). \tag{1}$$

This is its round-t loss if it had been used from the beginning. After switching from a to b, for example, the round-two deployed loss is $F_2(a, b)$, whereas the diagonal loss of the current choice is $d_2(b) = F_2(b, b)$. Full diagonal feedback reveals every $d_t(a)$ after round t, sparse feedback probes individual candidates independently of deployment, and deployment-only feedback reveals the realized loss ℓ_t. The accounting identity requires no observation of counterfactual losses, while the learning guarantees state their feedback assumptions separately.

Comparator and Loss Accounting. Let $D_T(a) = \sum_{t=1}^T d_t(a)$ be the cumulative loss of using a throughout. The best fixed value within family e is $V^e_{T,\mathcal{C}} = \min_{a \in \mathcal{C}_e} D_T(a)$. Our common comparator has value $V^{ZW}_{T,\mathcal{C}}$: hindsight selects one joint candidate for the entire horizon, with its own induced history. Write $L^{\mathrm{dep}}_T = \sum_{t=1}^T F_t(A_{1:t})$ for the learner's cumulative deployed loss and $L^{\mathrm{diag}}_T = \sum_{t=1}^T d_t(A_t)$ for its choices evaluated on their respective fixed-configuration histories. The latter sum need not describe one deployment trajectory. Adding and subtracting it and the best within-family value gives the decomposition.

Proposition 1 (Selection, restriction, and state accounting)**.** *For every trajectory with $A_t \in \mathcal{C}_e$,*

$$L^{\mathrm{dep}}_T - V^{ZW}_{T,\mathcal{C}} = \underbrace{L^{\mathrm{diag}}_T - V^e_{T,\mathcal{C}}}_{\textit{diagonal selection}} + \underbrace{V^e_{T,\mathcal{C}} - V^{ZW}_{T,\mathcal{C}}}_{\textit{family restriction}} + \underbrace{L^{\mathrm{dep}}_T - L^{\mathrm{diag}}_T}_{\textit{signed state effect}}. \tag{2}$$

Only family restriction is necessarily nonnegative; the identity also holds in expectation.

Selection compares choices within the available family, restriction charges its gap to the joint family, and state accounts for the actual history's benefit or harm. Stateless systems have zero state effect. For fixed diagonal tables, identical action marginals give equal expected diagonal loss, while different deployment histories can help or hurt. Appendix A.1.2 details sampling residuals and comparator averaging. The proofs appear in Appendix B.

3 Selection and Family Restriction

Selection and Prior Complexity. For stateless full-information selection, exponential weights with a positive prior μ_e on $\mathcal{C}_e$ and learning rate $\eta > 0$ uses the following distribution:

$$q_t(a) = \mu_e(a) e^{-\eta \sum_{r<t} d_r(a)} / \sum_{b \in \mathcal{C}_e} \mu_e(b) e^{-\eta \sum_{r<t} d_r(b)}. \tag{3}$$

Sample $A_t \sim q_t$, then observe every $d_t(a)$. The following theorem also covers standard-Borel configuration spaces with prior μ fixed before evaluation and measurable losses $L_t : \mathcal{A} \to [0, 1]$ fixed before the current draw and fully revealed afterward. Here $p_t(\mathrm{d}a) \propto e^{-\eta \sum_{r<t} L_r(a)} \mu(\mathrm{d}a)$. Write $\overline{L}_t = \int L_t \, \mathrm{d}p_t$ and $R_T(Q) = \sum_t \int L_t \, \mathrm{d}(p_t - Q)$, and let $C_Q = \mathrm{KL}(Q \,\|\, \mu)$ denote the Kullback–Leibler (KL) divergence. The comparator holds one draw from fixed Q throughout, with $R_T(a) = R_T(\delta_a)$.

Theorem 1 (Prior-complexity regret)**.** *For every $\eta > 0$ and fixed $Q \ll \mu$ with $C_Q < \infty$, prior-initialized exponential weights satisfies the following:*

$$R_T(Q) \le \mathrm{KL}(Q \,\|\, \mu)/\eta + \eta T/8. \tag{4}$$

For $C_Q > 0$, $\eta = \sqrt{8C_Q/T}$ gives the following:

$$R_T(Q) \le \sqrt{TC_Q/2}. \tag{5}$$

Setting $L_t = d_t$ and $\mu = \mu_e$, the best finite candidate $a_e^\star$ has $C_{\delta_{a_e^\star}} = \log(1/\mu_e(a_e^\star))$. Thus,

$$\mathbb{E}L_T^{\mathrm{dep}} - V_{T,\mathcal{C}}^{ZW} \le \log(1/\mu_e(a_e^\star))/\eta + \eta T/8 + V_{T,\mathcal{C}}^{e} - V_{T,\mathcal{C}}^{ZW}. \tag{6}$$

Uniform priors give a selection-regret bound of $\sqrt{T\log|\mathcal{C}_e|/2}$, but restriction can remain linear.

Appendix A.2.2 separates text and parameter prior complexity, treats continuous comparators by localization, and discusses computational limits.

Restriction and Mode Selection. To extend the finite-candidate comparison to all admissible configurations, let $V_T^e = \inf_{a\in\mathcal{A}_e}\sum_t L_t(a)$. Define $\Delta_e^{\mathrm{mode}} = V_T^e - V_T^{\cup}$, $\Delta^{\mathrm{joint}} = V_T^{\cup} - V_T^{ZW}$, and $\mathsf{Apx}_T^e(Q_e) = \sum_t \int L_t \,\mathrm{d}Q_e - V_T^e$. These nonnegative gaps distinguish comparator approximation, mode restriction, and access to joint configurations.

Proposition 2 (Selection–approximation–restriction decomposition)**.** *For any learner and comparator Q_e on $\mathcal{A}_e$, $e \in \{Z, W\}$, the regret $R_T^{ZW} = \sum_t \overline{L}_t - V_T^{ZW}$ satisfies the following:*

$$R_T^{ZW} = R_T(Q_e) + \mathsf{Apx}_T^e(Q_e) + \Delta_e^{\mathrm{mode}} + \Delta^{\mathrm{joint}}. \tag{7}$$

EW on $\mathcal{A}_e$ with prior μ_e bounds $R_T(Q_e)$ by $\mathrm{KL}(Q_e \,\|\, \mu_e)/\eta + \eta T/8$ whenever this KL is finite.

A mixture prior also permits online mode selection, charging the comparator mode an additional prior cost; see Corollary 5. Candidate construction remains separate: Appendix A.2.5 distinguishes localization from discovery without claiming convergence of prompt search or adapter training.

Joint and Coordinatewise Selection. Fix finite Cartesian candidates $\mathcal{C} = \mathcal{Z}_c \times \mathcal{W}_c$, a positive product prior, and exogenous stateless losses $d_t : \mathcal{C} \to [0,1]$. Both selectors observe every loss and start from $A_1 \sim \mu$. For deterministic $\eta_t > 0$, let $q_t(z,w) \propto \pi(z)\rho(w)e^{-\eta_t \sum_{r<t} d_r(z,w)}$. Joint sampling draws from q_t, while one coordinate sweep, for $t \ge 2$, uses fresh draws:

$$Z_t \sim q_t(\cdot \mid W_{t-1}), \qquad W_t \sim q_t(\cdot \mid Z_t). \tag{8}$$

Only the final pair is deployed. Write $R_T^{\mathrm{joint}}, R_T^{\mathrm{coord}}$ for expected regret against the same best fixed candidate. In this comparison, restriction and state effects vanish.

Proposition 3 (Equivalence under additive losses)**.** *If $d_t(z,w) = g_t(z) + h_t(w)$, one sweep has law q_t for every deterministic schedule, so $R_T^{\mathrm{coord}} = R_T^{\mathrm{joint}}$.*

Proposition 4 (Separation under a fixed learning rate)**.** *On $\mathcal{C} = \{0,1\}^2$ with uniform prior, let $d_t = f$, where $f(0,0) = \Delta$, $f(1,0) = f(0,1) = \Delta + b$, $f(1,1) = 0$, and $\Delta, b > 0$, $\Delta + b \le 1$. For fixed $\eta_t = \eta > 0$ and every $T \ge 1$,*

$$R_T^{\mathrm{joint}} \le 1 + 3/\eta, \qquad R_T^{\mathrm{coord}} \ge (\Delta T/4)\exp\!\big(-2/(e^{\eta b} - 1)\big). \tag{9}$$

Corollary 1 (Separation despite eventual optimal selection)**.** *For the same construction with $\Delta > b$, use $\eta_t = \log(4(t-1))/(b(t-1))$ for $t \ge 2$. Then $R_T^{\mathrm{joint}} = O(1)$ and $R_T^{\mathrm{coord}} = \Omega(\sqrt{T})$, although both selectors deploy $(1,1)$ on every sufficiently late round almost surely.*

These results concern the specified conditional sampler. More sweeps or global refreshes can change its behavior. Appendix A.2.4 provides refined constants, and Section 6 examines these settings.

4 State Effects and Policy Regret

Policy regret groups selection and state in (2): $R_T^{\mathrm{pol}}(Q) = \mathbb{E}L_T^{\mathrm{dep}} - \mathbb{E}\int D_T(a)\,\mathrm{d}Q(a)$, where the comparator draws once from Q and deploys it from the common initial condition (Arora et al., 2012). Point regret takes $Q = \delta_a$. Evaluating comparators on the learner's realized state instead gives external regret.

Proposition 5 (Zero external regret and linear policy regret)**.** *A deterministic two-configuration persistent-state system admits a trajectory with zero realized-state external regret and policy regret $T-1$ against a fixed configuration.*

Proposition 6 (Linear policy regret)**.** *For every optimizer whose first action uses only common pre-round history and private coins, and every $T \ge 2$, some deterministic two-configuration environment gives expected policy regret at least $(T-1)/2$, when its index and transition kernel are unknown before that action.*

Assumption 1 (Component-specific bounded influence)**.** *For integers $m_Z, m_W \geq 0$, measurable functions f_t, fixed before learner randomization, give loss $f_t(z_{\max\{1,t-m_Z\}:t}; w_{\max\{1,t-m_W\}:t}) \in [0,1]$. Older coordinates have no influence, including through hidden state, while interactions within the windows are unrestricted.*

Put $m = m_{\max} = \max\{m_Z, m_W\}$, truncating horizons at $T-1$. The diagonal $d_t(z,w)$ repeats each coordinate throughout its window. Influence measures persistence after replacement.

Assumption 2 (Diagonal counterfactual evaluation oracle)**.** *After round t, every own-trajectory loss $d_t(a)$, $a \in \mathcal{A}$, is revealed.*

Lazy Exponential Weights. Apply the EW update from Section 3 to d_t, starting with $q_1 = \mu$ and $A_1 \sim \mu$. After round $t < T$, retain A_t with probability $e^{-\eta d_t(A_t)}$ and otherwise refresh from q_{t+1} (Geulen et al., 2010; Kalai & Vempala, 2005). Refreshes preserve $A_t \sim q_t$ and occur at most $\eta(T-1)$ times in expectation. Each refresh contaminates at most m rounds. This requires own-trajectory evaluation.

Proposition 7 (Matched marginals isolate state effects)**.** *For fixed exogenous losses, family, prior, and learning rate, independent and lazy draws with the same diagonal feedback and EW marginals have equal expected diagonal losses. Hence,*

$$\mathbb{E}L_T^{\mathrm{dep,lazy}} - \mathbb{E}L_T^{\mathrm{dep,ind}} = \mathbb{E}[L_T^{\mathrm{dep,lazy}} - L_T^{\mathrm{diag,lazy}}] - \mathbb{E}[L_T^{\mathrm{dep,ind}} - L_T^{\mathrm{diag,ind}}]. \tag{10}$$

Thus fewer switches can help or hurt. The identity also holds conditional on shared exogenous approximate feedback.

Theorem 2 (Lazy policy regret under bounded influence)**.** *Under Assumptions 1–2, for every $\eta > 0$ and $Q \ll \mu$ with finite $C_Q = \mathrm{KL}(Q \,\|\, \mu)$,*

$$R_T^{\mathrm{pol}}(Q) \leq C_Q/\eta + \eta\,[T/8 + m_{\max}(T-1)]\,. \tag{11}$$

For $C_Q > 0$, comparator-dependent tuning gives the following:

$$\eta^\star = \sqrt{C_Q/(T/8 + m_{\max}(T-1))}, \tag{12}$$

$$R_T^{\mathrm{pol}}(Q) \leq 2\sqrt{C_Q[T/8 + m_{\max}(T-1)]} \leq 2\sqrt{TC_Q(m_{\max} + 1/8)} = O\left(\sqrt{(1+m_{\max})TC_Q}\right). \tag{13}$$

For $C_Q = 0$, draw once from μ without refreshing; regret is zero.

The $m\eta(T-1)$ term controls state, while the remaining terms control selection. These established low-switch methods (Geulen et al., 2010; Anava et al., 2015; Zhang et al., 2022) connect prior complexity to influence. Uniform priors give $C_Q = \log K$ for every point comparator.

Proposition 8 (Mode-restricted lazy influence rates)**.** *Under Assumptions 1–2, run lazy EW with priors $\mu_Z = \pi \otimes \delta_{w_0}$ and $\mu_W = \delta_{z_0} \otimes \rho$, supported on $\mathcal{A}_Z$ and $\mathcal{A}_W$, respectively. For fixed comparators $Q_Z \ll \pi$ and $Q_W \ll \rho$, let $C_Z = \mathrm{KL}(Q_Z \,\|\, \pi)$ and $C_W = \mathrm{KL}(Q_W \,\|\, \rho)$ be finite. For every $\eta > 0$, text-only and parameter-only selection satisfy*

$$R_T^{\mathrm{pol}}(Q_Z \otimes \delta_{w_0}) \leq C_Z/\eta + \eta[T/8 + m_Z(T-1)], \tag{14}$$

$$R_T^{\mathrm{pol}}(\delta_{z_0} \otimes Q_W) \leq C_W/\eta + \eta[T/8 + m_W(T-1)]. \tag{15}$$

Freezing one coordinate removes its switching contribution: the state charge depends on the influence horizon of the coordinate being changed. For positive complexity, comparator-tuned rates are $O(\sqrt{(1+m_Z)TC_Z})$ and $O(\sqrt{(1+m_W)TC_W})$. For the best fixed candidate within each finite family, add family restriction to compare with the joint optimum. Different upper bounds alone do not establish which mode performs better. Appendix A.3.2 gives localization and coupling details, while Appendix A.3.3 treats whole-policy selection across reset episodes. Proofs are in Appendix B.3.

5 Feedback Requirements and Extensions

Counterfactual feedback evaluates each candidate on the history it would have produced if used throughout. Deployment feedback reports only the loss actually incurred. We now quantify what changes as feedback becomes less informative. Fix a finite family $\mathcal{C}$ of K candidates with positive prior μ, and a comparator distribution Q fixed independently of learner and probe randomness. Write $m = m_{\max}$ and $C_Q = \mathrm{KL}(Q \,\|\, \mu)$ as in Section 4.

Approximate Full Counterfactual Feedback. Theorem 2 assumes exact diagonal losses. An exogenous full estimate sequence in $[0,1]$ with uniform round-t error ϵ_t adds at most $2\sum_t \epsilon_t$ to its bound. Conditionally unbiased estimates in the same range preserve the expected bound; refer to Appendix A.4.1. The entire evaluator must be independent of deployment randomness. Even a common estimation offset can alter retention and state effects while leaving EW marginals unchanged.

Sparse Independent Probes. Under the bounded-influence condition in Assumption 1, an autonomous evaluator can instead probe one candidate $J_t \sim q_t$ after each round. Its outcome $Z_t \in [0,1]$ must have conditional mean $d_t(J_t)$ given the exogenous response tree, earlier evaluation history, and J_t. Conditional on that tree, the entire evaluator is independent of deployment randomness, consumes no deployed actions or outcomes, and is unchanged by deployment interventions. Probing also leaves the tree unchanged; see Assumption 3. Initialize $q_1 = \mu$ and $A_1 \sim \mu$, and update EW using $\widehat{d}_t(a) = \mathbf{1}\{J_t = a\}Z_t/q_t(a)$. Retain A_t with probability $e^{-\eta \widehat{d}_t(A_t)}$ and otherwise refresh from q_{t+1} using independent deployment randomness.

Theorem 3 (Sparse-diagonal policy regret)**.** *Under Assumption 3, with probe law $r_t = q_t$, every fixed Q and $\eta > 0$ satisfy the following:*

$$R_T^{\mathrm{pol}}(Q) \le C_Q/\eta + \eta\,[KT/2 + m(T-1)]\,. \tag{16}$$

For $C_Q > 0$, $\eta = \sqrt{C_Q/[KT/2 + m(T-1)]}$ gives

$$R_T^{\mathrm{pol}}(Q) \le 2\sqrt{C_Q[KT/2 + m(T-1)]}. \tag{17}$$

For $C_Q = 0$, drawing once from μ gives zero expected regret.

One probe per round reduces KT counterfactual evaluations to T probes, replacing the full-feedback selection term $\eta T/8$ by $\eta KT/2$ while preserving the state term. Simulation or cloning costs remain, and independence from only the current deployment draw is insufficient. Appendix A.4.5 gives the formal process and refined bounds.

Deployment-Only Feedback. Without a counterfactual evaluator, hold each chosen configuration for a block of b rounds, where $1 \le b \le T$ is an integer. There are $J = \lfloor T/b \rfloor$ complete blocks and $r = T - bJ$ remaining rounds. Block j starts at $s_j = (j-1)b + 1$ after the realized prefix $H_j = A_{1:s_j-1}$. We write U explicitly for the external random tape introduced in Section 2.

Definition 1 (Block counterfactual)**.** *The average loss from repeating a after the actual prefix, rerunning environmental responses, is*

$$g_j(a) = (1/b)\sum_{u=1}^{b} F_{s_j+u-1}(H_j, a^u; U), \tag{18}$$

where a^u denotes u repetitions of a.

Starting from $p_1 = \mu$, deploy a fresh private draw $B_j \sim p_j$ throughout block j and update EW using $\widehat{g}_j(a) = \mathbf{1}\{B_j = a\}Z_j/p_j(a)$, where $Z_j = g_j(B_j)$ is the observed block-average loss. Retain the last configuration for the remaining r rounds. The signed *causal reaction* compares repeating a comparator after the learner's prefixes with using it from the beginning, over the complete blocks:

$$\Gamma_b(Q) = \mathbb{E}\sum_{a\in\mathcal{C}} Q(a)\left[b\sum_{j=1}^{J} g_j(a) - \sum_{t=1}^{bJ} d_t(a)\right]. \tag{19}$$

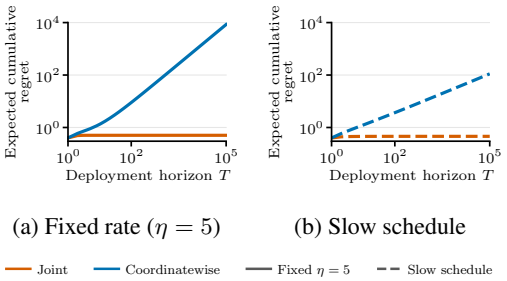
It includes changes in tasks, user responses, and operational state.

Theorem 4 (Oracle-free causal policy-regret inequality)**.** *For causal responses $F_t \in [0,1]$ driven by an external tape independent of private learner coins, the block algorithm satisfies, for every fixed Q and $\eta > 0$,*

$$R_T^{\mathrm{pol}}(Q) \le bC_Q/\eta + \eta bKJ/2 + \Gamma_b(Q) + r. \tag{20}$$

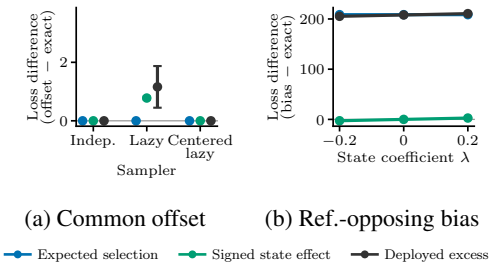
No oracle for g_j, d_t, or Γ_b is required.

The first two terms bound selection from observed block losses, while $\Gamma_b(Q)$ accounts for the different histories and r covers the unfinished block. This retains the causal discrepancy in policy-regret mini-batching (Arora et al., 2012), and uncontrolled reaction can remain linear. If the *complete* causal system has influence horizon m, including task selection and evaluation, then $|\Gamma_b(Q)| \le \min\{m, b\}(J-1)$: only the first m rounds of each later block can remember its prefix. Optimizing b and η yields $O((mKC_Q)^{1/3}T^{2/3})$ for fixed positive m, K, C_Q and sufficiently large



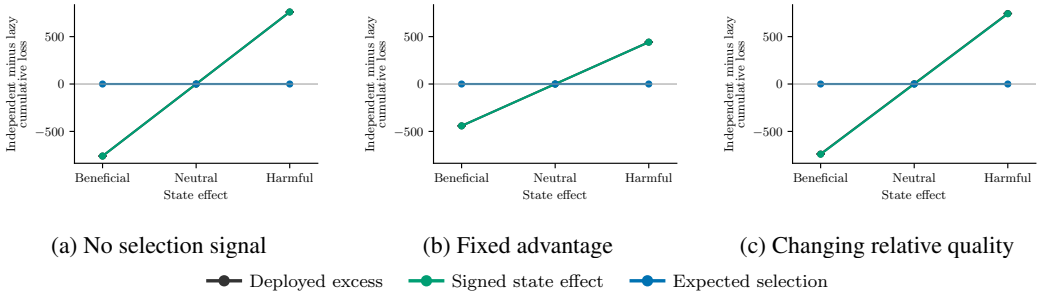
(a) Fixed rate ($\eta = 5$) (b) Slow schedule

Figure 2: Exact expected regret for joint and coordinatewise selection with fixed and slow learning rates.



(a) Common offset (b) Ref.-opposing bias

Figure 3: Approximate-minus-exact loss components on fixed-advantage tables: (a) $\lambda = 0.2$; (b) lazy EW.



(a) No selection signal (b) Fixed advantage (c) Changing relative quality

Figure 4: Independent-minus-lazy loss contrasts at $T = 4{,}096$, with $\lambda \in \{-0.2, 0, 0.2\}$. Reversing the mechanism reverses the selector ordering while expected diagonal selection remains unchanged.

T; see Corollary 12. Use $b = 1$ for $m = 0$, and a single prior draw for $C_Q = 0$. Forgetting in a cache alone does not establish this complete-system condition.

For random response trees, independence and unbiasedness must hold conditionally to compare with the tree's best fixed candidate. Query-adaptive autonomous evaluation supports fixed or best expected-loss comparators, not comparators selected using probe noise. Appendix A.4 provides supporting lemmas and extensions.

6 Empirical Analysis

Experimental Setup. Controlled studies isolate selection, family restriction, and signed state effects on deployed loss. SQL experiments use Qwen3.5-9B (Qwen Team, 2026) on Spider (Yu et al., 2018) and SParC (Yu et al., 2019). Excess loss is measured relative to each table's or benchmark's best fixed joint configuration unless a paired contrast is specified. Summaries and error bars use means $\pm$ one standard error (SE). Refer to Appendices C and C.1 for detailed designs, numerical results, and uncertainty.

Joint and Coordinatewise Selection. Joint EW and text-then-parameter sweeps share the uniform prior, preceding loss rows, and rate schedule on the additive and barrier tables of Section 3. Family restriction, state effects, and the best fixed loss are zero, so cumulative loss equals regret. The additive control gives equal regret, whereas one coordinatewise sweep incurs substantially more regret on the barrier under both schedules, as in Figure 2. Under the slow schedule, high final-round optimal-selection probability coexists with this accumulated gap. Extra sweeps reduce the gap at additional computation cost without new feedback; see Figure 8 and Appendix C.2 for details.

Selection and Signed State Effects. Fixed-advantage, changing-quality, and no-signal diagonal tables use $f_t = d_t(A_t) + \lambda I_t$, where I_t indicates mixed configurations in the relevant coordinate histories. Negative, zero, and positive λ give beneficial, neutral, and harmful state effects. State and approximate-feedback figures use $(m_Z, m_W) = (4, 0)$; see Appendix C.3 for details. Independent and lazy EW have identical expected diagonal losses, as presented in Figure 9, but lazy EW has lower deployed loss under harmful state and higher loss under beneficial state; see Figure 4. Fewer switches can therefore help or hurt. Without a selection signal, draw-once has zero excess while both EW selectors incur harmful state costs, as shown in Figure 10.

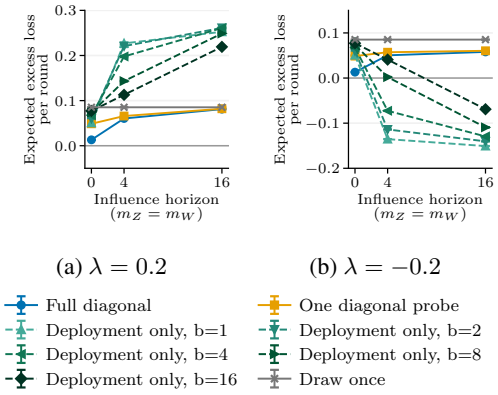
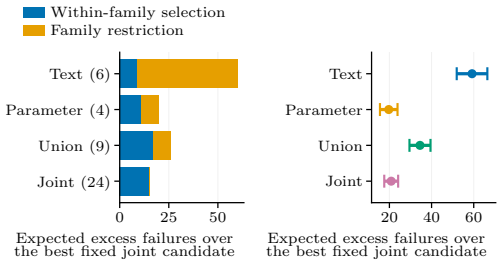


Figure 5: Restricted feedback on fixed-gap losses at $T = 1{,}024$. Zero marks the best fixed comparator.



(a) Full candidate families (b) Equal count ($K = 4$)

Figure 6: Spider full-diagonal EW excess failures relative to each seed's joint optimum. Panel (a) shows selection regret and family restriction, with family sizes in parentheses. Panel (b) shows all base-anchored $K = 4$ subsets, averaged within each seed.

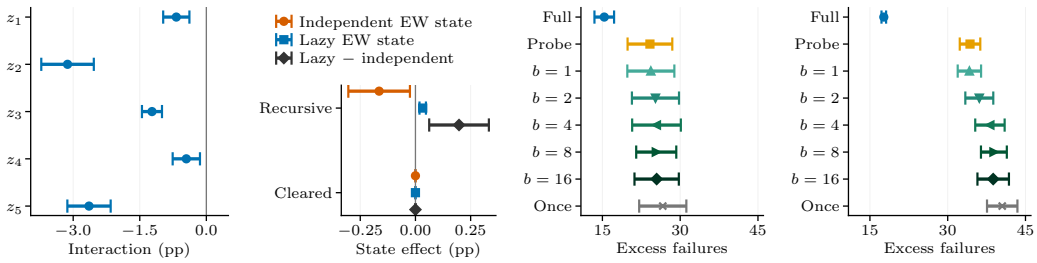
Feedback Accuracy, Switching, and Evaluation Cost. We perturb diagonal feedback at $T = 4{,}096$, fixing the true table, prior, rate, and learner random inputs. A common offset preserves selection distributions but changes ordinary lazy EW's switching and state costs, while the trajectories of independent EW and centered lazy EW remain unchanged. A reference-opposing bias changes both selection and state, as shown in Figure 3; refer to Appendix C.4 for experimental details and additional results. Increasing the Bernoulli sampling budget reduces feedback error without monotone deployed-loss improvement in the displayed conditions; see Figure 12.

Restricted Feedback on Controlled Losses. On the four-candidate joint family, we compare full-diagonal and one-probe lazy EW, deployment-only EW with blocks $b \in \{1, 2, 4, 8, 16\}$, and draw-once. All use uniform priors and regime-specific rates, so both information and learning rate vary; refer to Appendix C.5. At $T = 1{,}024$ and $m_Z = m_W = 4$, increasing b from 1 to 16 lowers excess under harmful state but raises it under beneficial state, as shown in Figure 5. These results illustrate how blocks sacrifice updates and suppress both kinds of state effects. Negative excess means a changing history beats every fixed configuration under this synthetic law.

Selection and Family Restriction on Spider. Spider compares text, parameter, union, and joint families with 6, 4, 9, and 24 configurations against each seed's best fixed joint candidate. Full-diagonal EW uses a fixed question order and a uniform prior. Independent and lazy expectations coincide because state is absent; refer to Appendices C.6 and C.7 for details. Text EW has lower selection regret than joint EW, but its larger family restriction produces higher expected deployed loss, as shown in Figure 6(a). Joint EW also has lower mean loss than parameter EW. Averaging all base-anchored $K = 4$ subsets gives similar parameter and joint excess losses, as presented in Figure 6(b). The full-pool ordering therefore need not survive equal candidate counts. Demonstrations also show stronger mean improvement beyond additive prompt–adapter effects than verification, as presented in Figure 7(a); refer to Appendix C.7 for details.

State Effects on SParC. SParC isolates history effects under matched configuration marginals. Recursive history retains earlier generated SQL, while cleared history removes it, retaining the schema and raw utterances. Dialogue resets bound influence by four rounds. Both couplings share diagonal losses and an influence-dependent rate, which is retained in the cleared control. Mixed-history outcomes measure state effects without updating selection; see Appendix C.8. Independent EW has lower estimated recursive-history loss despite identical expected diagonal loss, as shown in Figure 7(b). Lazy EW substantially reduces switching with only a small change in estimated expected loss; see Table 7. Fewer mixed histories remove both beneficial and harmful effects. These data do not establish that fewer switches worsen performance. Clearing eliminates state effects and equalizes expected deployed losses but changes the diagonal table and comparator, so cross-condition totals cannot identify state effects.

Restricted Feedback on Precomputed SQL Losses. We apply the controlled feedback regimes to precomputed Spider and cleared-SParC loss matrices, keeping outcomes, order, prior, and joint comparator fixed. We use zero-influence learning rates; see Appendix C.9. Full feedback has the lowest mean excess on both tasks, whereas one-probe feedback queries fewer diagonal entries but incurs more excess, as presented in Figures 7(c) and 7(d). One-probe and deployment-only $b = 1$



(a) Spider interactions (b) SParC state (c) Spider feedback (d) SParC feedback

Figure 7: Diagnostics with prompt–adapter interactions at adapter learning rate 10^{-4} and SParC state effects, followed by restricted-feedback excess failures. Negative interactions indicate improvement beyond additive effects. Full and Probe denote full-diagonal and one-probe lazy EW, b the deployment block length, and Once draw-once.

means are close; refer to Appendix C.9. The longest blocks reduce switches but raise mean excess relative to $b = 1$ on both tasks; see Table 8.

7 Related Work

Parameter Adaptation and Model Editing. Weight adaptation includes full tuning, adapters, low-rank updates (Houlsby et al., 2019; Hu et al., 2022), continuous prompt/prefix tuning (Li & Liang, 2021; Lester et al., 2021), and targeted editing (Meng et al., 2022). Updates can cause interference and forgetting (Kirkpatrick et al., 2017), with low-rank and full tuning differing even at similar performance (Shuttleworth et al., 2025). Comparisons with in-context learning report task-dependent accuracy, generalization, and cost tradeoffs (Liu et al., 2022; Mosbach et al., 2023; Ghosh et al., 2026; Razumovskaia et al., 2025), while theory characterizes when prompting can match weight tuning (Genewein et al., 2025). Joint construction includes BetterTogether (Soylu et al., 2024), MetaTuner (Bo et al., 2026), fast–slow updates (Tiwari et al., 2026), and WHALE's alternation of online rejection-sampling tuning with Meta-Harness search (Kim et al., 2026). Our conditional sampler studies selection within a fixed family, with candidate construction.

Expert Advice, Mixture Priors, and Limited Switching. Hedge and external regret compare fixed experts (Hannan, 1957; Foster & Vohra, 1999; Littlestone & Warmuth, 1994; Freund & Schapire, 1997; Cesa-Bianchi & Lugosi, 2006), and KL change-of-measure extends the analysis to measurable priors. Bandit, tracking, and sleeping/growing-expert models change feedback, comparators, or availability (Auer et al., 2002; Bubeck & Cesa-Bianchi, 2012; Herbster & Warmuth, 1998; Cortes et al., 2019; Mourtada & Maillard, 2017). Lazy and limited-switch methods control regret and action changes (Kalai & Vempala, 2005; Geulen et al., 2010; Altschuler & Talwar, 2018). Switching costs admit $\sqrt{T}$ full-information and $\widetilde{\Theta}(T^{2/3})$ adversarial-bandit rates; bounded-memory adversaries can force $T^{2/3}$ under ordinary full information (Cesa-Bianchi et al., 2013; Dekel et al., 2014). Zhang et al. (2022) combine prior divergence and movement penalties in comparator-adaptive finite-expert bounds. Under exogenous losses, coupling converts marginal movement into expected switches.

Appendix E provides additional related work.

8 Conclusion

We decomposed deployed loss over the best fixed joint configuration into selection, family restriction, and signed state effects. Low regret within a candidate family, eventual optimal selection, and infrequent switching can each occur alongside higher cumulative loss. Under the stated influence and feedback assumptions, EW with infrequent switching bounds selection regret and state effects, while weaker feedback introduces additional costs and requires further assumptions. Our controlled and SQL experiments illustrate how these distinctions arise within the evaluated candidate families.

AI USE STATEMENT

We used GPT-5.6 Sol (OpenAI, 2026a) and GPT-6 Astra (OpenAI, 2026b) as our primary generative AI models. We used these models to refine research ideas, derive mathematical proofs, implement experiments, and draft and polish the manuscript. The human authors take responsibility for the final content of this work, including all claims, proofs, experimental results, code, and text produced with AI assistance.

ETHICS STATEMENT

We do not identify direct ethical concerns arising from the theoretical analysis and benchmark experiments presented in this work. However, methods for adapting language-model systems could be used in applications with harmful consequences. We acknowledge this potential and emphasize the importance of responsible use and appropriate safeguards.

REFERENCES

L. A. Agrawal, S. Tan, D. Soylu, N. Ziems, R. Khare, K. Opsahl-Ong, A. Singhvi, H. Shandilya, M. J. Ryan, M. Jiang, C. Potts, K. Sen, A. G. Dimakis, I. Stoica, D. Klein, M. Zaharia, and O. Khattab. GEPA: Reflective prompt evolution can outperform reinforcement learning. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2026.

V. Akinwande, Y. Jiang, D. Sam, and J. Z. Kolter. Understanding prompt engineering may not require rethinking generalization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024.

J. Altschuler and K. Talwar. Online learning over a finite action set with limited switching. In *Proceedings of the Conference on Learning Theory (COLT)*, 2018.

O. Anava, E. Hazan, and S. Mannor. Online learning for adversaries with memory: Price of past mistakes. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2015.

R. Arora, O. Dekel, and A. Tewari. Online bandit learning against an adaptive adversary: From regret to policy regret. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2012.

P. Asawa, C. M. Glaze, G. Orlanski, R. Ramakrishnan, B. Xu, A. Biswal, V. S. Chen, F. Sala, M. Zaharia, and J. E. Gonzalez. Continual learning bench: Evaluating frontier AI systems in real-world stateful environments, 2026. arXiv preprint arXiv:2606.05661.

P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2002.

O. Avner, S. Mannor, and O. Shamir. Decoupling exploration and exploitation in multi-armed bandits. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2012.

X. Bo, R. Li, Z. Sun, Q. Dai, Z. Zhang, Z. Tian, X. Chen, and Z. Dong. Prompt and parameter co-optimization for large language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2026.

T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

S. Bubeck and N. Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning*, 5(1):1–122, 2012.

N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

N. Cesa-Bianchi, O. Dekel, and O. Shamir. Online learning with switching costs and other adaptive adversaries. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2013.

D. Cheng, X. Zhou, and B. Ji. Understanding the role of feedback in online learning with switching costs. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2023.

C. Cortes, G. DeSalvo, C. Gentile, M. Mohri, and S. Yang. Online learning with sleeping experts and feedback graphs. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2019.

D. P. de Farias and N. Megiddo. Combining expert advice in reactive environments. *Journal of the ACM*, 53(5):762–799, 2006.

O. Dekel, J. Ding, T. Koren, and Y. Peres. Bandits with switching costs: $T^{2/3}$ regret. In *Proceedings of the Annual ACM Symposium on Theory of Computing (STOC)*, 2014.

M. Deng, J. Wang, C.-P. Hsieh, Y. Wang, H. Guo, T. Shu, M. Song, E. Xing, and Z. Hu. RLPrompt: Optimizing discrete text prompts with reinforcement learning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2022.

M. Dudík, J. Langford, and L. Li. Doubly robust policy evaluation and learning. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2011.

E. Even-Dar, S. M. Kakade, and Y. Mansour. Online Markov decision processes. *Mathematics of Operations Research*, 34(3):726–736, 2009.

J. Fang, X. Deng, H. Xu, Z. Jiang, Y. Tang, Z. Xu, S. Deng, Y. Yao, M. Wang, S. Qiao, H. Chen, and N. Zhang. LightMem: Lightweight and efficient memory-augmented generation. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2026.

U. Feige, T. Koren, and M. Tennenholtz. Chasing ghosts: Competing with stateful policies. *SIAM Journal on Computing*, 46(1):190–223, 2017.

C. Fernando, D. S. Banarse, H. Michalewski, S. Osindero, and T. Rocktäschel. Promptbreeder: Self-referential self-improvement via prompt evolution. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2024.

D. P. Foster and R. V. Vohra. Regret in the on-line decision problem. *Games and Economic Behavior*, 29(1–2):7–35, 1999.

Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.

T. Genewein, K. Li, J. Grau-Moya, A. Ruoss, L. Orseau, and M. Hutter. Understanding prompt tuning and in-context learning via meta-learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

S. Geulen, B. Vöcking, and M. Winkler. Regret minimization for online buffering problems using the weighted majority algorithm. In *Proceedings of the Conference on Learning Theory (COLT)*, 2010.

B. Ghosh, S. Das, T. Speicher, Q. Wu, M. A. Khan, D. Garg, K. P. Gummadi, and E. Terzi. Fine-tuning vs. in-context learning in large language models: A formal language learning perspective. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2026.

P. D. Grünwald. *The Minimum Description Length Principle*. MIT Press, 2007.

Q. Guo, R. Wang, J. Guo, B. Li, K. Song, X. Tan, G. Liu, J. Bian, and Y. Yang. Connecting large language models with evolutionary algorithms yields powerful prompt optimizers. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024.

J. Hannan. Approximation to Bayes risk in repeated play. In M. Dresher, A. W. Tucker, and P. Wolfe (eds.), *Contributions to the Theory of Games, Volume III*, pp. 97–139. Princeton University Press, Princeton, NJ, 1957.

M. Herbster and M. K. Warmuth. Tracking the best expert. *Machine Learning*, 32:151–178, 1998.

W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. De Laroussilhe, A. Gesmundo, M. Attariyan, and S. Gelly. Parameter-efficient transfer learning for NLP. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2019.

E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. LoRA: Low-rank adaptation of large language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2022.

N. Jiang and L. Li. Doubly robust off-policy value evaluation for reinforcement learning. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2016.

A. Kalai and S. Vempala. Efficient algorithms for online decision problems. *Journal of Computer and System Sciences*, 71(3):291–307, 2005.

O. Khattab, A. Singhvi, P. Maheshwari, Z. Zhang, K. Santhanam, S. Vardhamanan, S. Haq, A. Sharma, T. T. Joshi, H. Moazam, H. Miller, M. Zaharia, and C. Potts. DSPy: Compiling declarative language model calls into self-improving pipelines. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024.

H. Kim, Y. Lee, G. Lee, C. Finn, and K. Lee. WHALE: A simple recipe for joint harness-weight optimization, 2026. arXiv preprint arXiv:2609.00196.

J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska, D. Hassabis, C. Clopath, D. Kumaran, and R. Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017.

Y. Lee, J. Boen, and C. Finn. Feedback descent: Open-ended text optimization via pairwise comparison. In *ICLR Workshop on AI with Recursive Self-Improvement*, 2026a.

Y. Lee, R. S. Nair, Q. Zhang, K. Lee, O. Khattab, and C. Finn. Meta-Harness: End-to-end optimization of model harnesses. In *Proceedings of the Conference on Language Modeling (COLM)*, 2026b.

B. Lester, R. Al-Rfou, and N. Constant. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2021.

P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. t. Yih, T. Rocktäschel, S. Riedel, and D. Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

X. L. Li and P. Liang. Prefix-Tuning: Optimizing continuous prompts for generation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, 2021.

N. Littlestone and M. K. Warmuth. The weighted majority algorithm. *Information and Computation*, 108(2):212–261, 1994.

H. Liu, D. Tam, M. Muqeeth, J. Mohta, T. Huang, M. Bansal, and C. A. Raffel. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

D. A. McAllester. PAC-Bayesian model averaging. In *Proceedings of the Conference on Learning Theory (COLT)*, 1999.

K. Meng, D. Bau, A. Andonian, and Y. Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

S. Min, X. Lyu, A. Holtzman, M. Artetxe, M. Lewis, H. Hajishirzi, and L. Zettlemoyer. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2022.

M. Mosbach, T. Pimentel, S. Ravfogel, D. Klakow, and Y. Elazar. Few-shot fine-tuning vs. in-context learning: A fair comparison and evaluation. In *Findings of the Association for Computational Linguistics: ACL*, 2023.

J. Mourtada and O.-A. Maillard. Efficient tracking of a growing number of experts. In *Proceedings of the International Conference on Algorithmic Learning Theory (ALT)*, 2017.

G. Neu, A. Antos, A. György, and C. Szepesvári. Online Markov decision processes under bandit feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2010.

OpenAI. GPT-5.6: Frontier intelligence that scales with your ambition. `https://openai.com/index/gpt-5-6/`, 2026a. Accessed September 25, 2026.

OpenAI. GPT-6 Astra: A new generation of intelligence. `https://openai.com/index/gpt-6-astra/`, 2026b. Accessed September 25, 2026.

K. Opsahl-Ong, M. J. Ryan, J. Purtell, D. Broman, C. Potts, M. Zaharia, and O. Khattab. Optimizing instructions and demonstrations for multi-stage language model programs. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024.

F. Orabona and D. Pál. Coin betting and parameter-free online learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

C. Packer, S. Wooders, K. Lin, V. Fang, S. G. Patil, I. Stoica, and J. E. Gonzalez. MemGPT: Towards LLMs as operating systems, 2023. arXiv preprint arXiv:2310.08560.

R. Pryzant, D. Iter, J. Li, Y. Lee, C. Zhu, and M. Zeng. Automatic prompt optimization with "gradient descent" and beam search. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023.

Qwen Team. Qwen3.5-9B model card. `https://huggingface.co/Qwen/Qwen3.5-9B`, 2026. Accessed August 30, 2026.

E. Razumovskaia, I. Vulić, and A. Korhonen. Analyzing and adapting large language models for few-shot multilingual NLU: Are we there yet? *Transactions of the Association for Computational Linguistics*, 13:1096–1120, 2025.

O. Rubin, J. Herzig, and J. Berant. Learning to retrieve prompts for in-context learning. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2022.

T. Schick, J. Dwivedi-Yu, R. Dessì, R. Raileanu, M. Lomeli, E. Hambro, L. Zettlemoyer, N. Cancedda, and T. Scialom. Toolformer: Language models can teach themselves to use tools. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

T. Shin, Y. Razeghi, R. L. Logan IV, E. Wallace, and S. Singh. AutoPrompt: Eliciting knowledge from language models with automatically generated prompts. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020.

N. Shinn, F. Cassano, A. Gopinath, K. Narasimhan, and S. Yao. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

R. Shuttleworth, J. Andreas, A. Torralba, and P. Sharma. LoRA vs full fine-tuning: An illusion of equivalence. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

D. Soylu, C. Potts, and O. Khattab. Fine-tuning and prompt optimization: Two great steps that work better together. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024.

R. Tiwari, K. Sareen, L. A. Agrawal, J. E. Gonzalez, M. Zaharia, K. Keutzer, I. S. Dhillon, R. Agarwal, and D. Khatri. Learning, fast and slow: Towards LLMs that adapt continually, 2026. arXiv preprint arXiv:2605.12484.

D. Wu, H. Wang, W. Yu, Y. Zhang, K.-W. Chang, and D. Yu. LongMemEval: Benchmarking chat assistants on long-term interactive memory. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025.

W. Xu, Z. Liang, K. Mei, H. Gao, J. Tan, and Y. Zhang. A-Mem: Agentic memory for LLM agents. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

C. Yang, X. Wang, Y. Lu, H. Liu, Q. V. Le, D. Zhou, and X. Chen. Large language models as optimizers. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024a.

J. Yang, C. E. Jimenez, A. Wettig, K. Lieret, S. Yao, K. Narasimhan, and O. Press. SWE-agent: Agent-computer interfaces enable automated software engineering. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024b.

S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. Narasimhan, and Y. Cao. ReAct: Synergizing reasoning and acting in language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023.

T. Yu, R. Zhang, K. Yang, M. Yasunaga, D. Wang, Z. Li, J. Ma, I. Li, Q. Yao, S. Roman, Z. Zhang, and D. Radev. Spider: A large-scale human-labeled dataset for complex and cross-domain semantic parsing and text-to-SQL task. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2018.

T. Yu, R. Zhang, M. Yasunaga, Y. C. Tan, X. V. Lin, S. Li, H. Er, I. Li, B. Pang, T. Chen, E. Ji, S. Dixit, D. Proctor, S. Shim, J. Kraft, V. Zhang, C. Xiong, R. Socher, and D. Radev. SParC: Cross-domain semantic parsing in context. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2019.

M. Yüksekgönül, F. Bianchi, J. Boen, S. Liu, P. Lu, Z. Huang, C. Guestrin, and J. Zou. Optimizing generative AI by backpropagating language model feedback. *Nature*, 639:609–616, 2025.

L. Zhang, T. Ergen, L. Logeswaran, M. Lee, and D. Jurgens. SPRIG: Improving large language model performance by system prompt optimization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2026.

Z. Zhang, A. Cutkosky, and I. Ch. Paschalidis. Optimal comparator adaptive online learning with switching cost. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

R. Zhong, T. Yu, and D. Klein. Semantic evaluation for text-to-SQL with distilled test suites. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020.

Y. Zhou, A. I. Muresanu, Z. Han, K. Paster, S. Pitis, H. Chan, and J. Ba. Large language models are human-level prompt engineers. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023.

A SUPPLEMENTARY THEORY

This appendix supplies the detailed setup, exact constants, supporting statements, and extensions for Sections 2–5. Proofs of the main and supplementary results appear in Appendix B.

A.1 CONFIGURATIONS, COMPARATORS, AND LOSS DECOMPOSITION

This section describes the configuration spaces, admissible families, and execution traces used to define the loss decomposition in (2).

A.1.1 ADMISSIBLE CONFIGURATIONS AND EXECUTION

Let θ_0 denote the base model. The admissible configuration families depend on how updates are produced and deployed.

Definition 2 (Adaptation protocol and admissible families)**.** *An* adaptation protocol Ξ *declares the base model, executor and interfaces, available data, resource budgets, and permitted text- and parameter-update procedures. It induces standard Borel spaces* $(\mathcal{Z}_\Xi, \mathcal{F}_\mathcal{Z})$ *and* $(\mathcal{W}_\Xi, \mathcal{F}_\mathcal{W})$ *of admissible text and parameter configurations, together with a measurable compatibility family* $\mathcal{A}_\Xi \subseteq \mathcal{Z}_\Xi \times \mathcal{W}_\Xi$ *of deployed pairs containing the base pair. Cross-mode simulation is admissible only if* Ξ *permits and charges its interfaces, storage, and computation.*

Fix Ξ and suppress its subscript. We specify evaluator and feedback assumptions separately. Standard-Borel spaces accommodate finite, countable, and usual continuous configurations. The *text configuration* z, with base object z_0, comprises reusable objects with canonical token serializations: prompts, retained memory, demonstrations, retrieval contents, and harness instructions. Literal serializations are countable, while continuous relaxations require a Ξ-declared quantizer or serializer for deployment evaluation.

A parameter configuration $w \in \mathcal{W}$ specifies model parameters $\theta(w)$, including learned auxiliary tensors, with $\theta(w_0) = \theta_0$. It may be continuous or atomic, including checkpoints, adapters, patches, and learned prompt/prefix tensors. In particular, dense tensors belong to w. The measurable executor exec combines $(z, \theta(w))$, task input, and possibly persistent operational state s: environment state, tool sessions, caches, and runtime objects not selected as reusable configuration.

For a current configuration $a = (z, w)$, define the admissible one-coordinate update sets:

$$\mathcal{U}_Z(a) = \mathcal{A} \cap (\mathcal{Z} \times \{w\}), \tag{21}$$

$$\mathcal{U}_W(a) = \mathcal{A} \cap (\{z\} \times \mathcal{W}). \tag{22}$$

These updates hold the other coordinate at its current value, whereas joint updates select from $\mathcal{A}$. The comparison families in Section 2 instead fix coordinates at the base pair $a_0 = (z_0, w_0)$. For non-Cartesian compatibility families, priors must be supported on $\mathcal{A}$. The prior-based results remain valid, but product KL identities need not hold. Joint configurations may change both base coordinates regardless of installation order.

Given task input x_t at round t, execution produces a trace, loss, and next state:

$$\tau_t \sim \mathrm{exec}(M_{\theta(w_t)}, z_t, s_t, x_t), \tag{23}$$

$$\ell_t = \ell(x_t, \tau_t) \in [0, 1], \tag{24}$$

$$s_{t+1} \sim P_t(\cdot \mid s_t, z_t, w_t, x_t, \tau_t). \tag{25}$$

Here τ_t records outputs and intermediate actions, and P_t governs state transitions. The normalized loss may combine quality and per-use resources. Appendix D treats installation and path-dependent transition costs separately.

A.1.2 FIXED COMPARATORS AND LOSS DECOMPOSITION

We use the candidate families, causal losses, and cumulative comparator values defined in Section 2. Exogenous tasks are shared across counterfactual histories, while deployment-responsive tasks, users, and environments are rerun under each history. The accounting identity (2) conditions on this response tree. We refine its restriction and averaging conventions below. For text-only or parameter-only selection, restriction can be further split at the union family into $V^e_{T,\mathcal{C}} - V^\cup_{T,\mathcal{C}}$ and $V^\cup_{T,\mathcal{C}} - V^{ZW}_{T,\mathcal{C}}$. Neither difference alone measures a nonadditive text–parameter interaction.

For fixed diagonal tables and action marginals q_t, $\mathbb{E}L_T^{\mathrm{diag}} = \sum_t \int d_t \, \mathrm{d}q_t$. Substituting this expectation into a realized trajectory's decomposition requires the *sampling residual*:

$$L_T^{\mathrm{diag}} - \mathbb{E}L_T^{\mathrm{diag}}. \tag{26}$$

Experiments retain this residual and distinguish averaged hindsight optima from expected-loss optima: generally, $\mathbb{E}\min_a D_T(a) \neq \min_a \mathbb{E}D_T(a)$.

A.1.3 Feasibility and Configuration Reachability

Feasibility asks whether a requested operational effect is attainable, separately from its loss.

Definition 3 (Protocol-relative update reachability)**.** *Fix an operating condition $c = (t, s, x)$. For configuration $b = (z', w')$, let $\Gamma_c(b)$ be the law of the execution trace and next state (τ, s^+) obtained by executing $M_{\theta(w')}$ with z' from (s, x) and applying P_t. From current configuration a, the reachable sets are*

$$\mathfrak{R}_Z(c; a) = \{\Gamma_c(b) : b \in \mathcal{U}_Z(a)\}, \tag{27}$$

$$\mathfrak{R}_W(c; a) = \{\Gamma_c(b) : b \in \mathcal{U}_W(a)\}. \tag{28}$$

A target effect is feasible for a mode when its law belongs to that mode's reachable set, while a task-specific discrepancy defines approximate feasibility.

Two updates have the same effect at c when their trace/state laws agree, while approximate feasibility uses a declared discrepancy tolerance. Neither reachable set need contain the other, and both depend on Ξ and potentially c. Finite search failure does not prove nonreachability, and budget-violating effects are inadmissible. For multi-round targets, replace Γ_c by the declared policy's rollout law.

These definitions imply neither disjoint neural effects nor persistence determined by artifact type.

A.2 Selection and Family Restriction

We bound stateless full-information selection in (2), then distinguish restriction, coordinatewise selection, and candidate discovery.

A.2.1 Finite Candidate Selection

Section 3 states the finite-family update (3) and selection-plus-restriction bound (6). The proof is in Appendix B.2. With a uniform prior over $N_e > 1$ candidates, choosing $\eta = \sqrt{8 \log N_e / T}$ bounds selection regret by $\sqrt{T \log N_e / 2}$ without knowing the best candidate. Restriction can remain linear, although expanding a family can reduce it while increasing selection complexity, independently of the mode label. A singleton family incurs no diagonal selection regret.

A.2.2 Measurable Priors and Comparator Complexity

To include continuous parameters, let $(\mathcal{A}, \mathcal{F}_{\mathcal{A}})$ be a standard-Borel space, not necessarily a product, with prior μ fixed before the losses. Deployment of $a = (z, w)$ has measurable expected loss

$$L_t(a) = L_t(z, w) = \mathbb{E}_{\tau \sim \mathrm{exec}(M_{\theta(w)}, z, x_t)}[\ell(x_t, \tau)] \in [0, 1]. \tag{29}$$

This evaluator averages model, tool, and harness randomness in (24), whereas Section 2 fixes exogenous randomness. The bounds apply to either table under its comparator averaging convention.

Loss L_t may depend on earlier history but is fixed before $a_t \sim p_t$. The learner incurs $L_t(a_t)$ and observes all of L_t. Given pre-draw information $\mathcal{H}_{t-1}$, its conditional mean loss is

$$\overline{L}_t = \mathbb{E}[L_t(a_t) \mid \mathcal{H}_{t-1}] = \int_{\mathcal{A}} L_t(a) \, \mathrm{d}p_t(a). \tag{30}$$

Averaging over the learner's draw defines conditional pseudo-regret, and an outer expectation gives expected pseudo-regret.

Definition 4 (Point and distributional external regret)**.** *For a fixed reference configuration $a^\star \in \mathcal{A}$, define*

$$R_T(a^\star) = \sum_{t=1}^{T} \overline{L}_t - \sum_{t=1}^{T} L_t(a^\star). \tag{31}$$

For a fixed probability measure Q on $\mathcal{A}$, define

$$R_T(Q) = \sum_{t=1}^{T} \overline{L}_t - \sum_{t=1}^{T} \int_{\mathcal{A}} L_t(a)\,\mathrm{d}Q(a). \tag{32}$$

The comparator holds one draw from fixed Q throughout the sequence, with $R_T(a^\star) = R_T(\delta_{a^\star})$. Overall point regret is

$$R_T^{\mathrm{ext}} = \sup_{a^\star \in \mathcal{A}} R_T(a^\star) = \sum_{t=1}^{T} \overline{L}_t - \inf_{a \in \mathcal{A}} \sum_{t=1}^{T} L_t(a). \tag{33}$$

The infimum need not be attained.

Definition 4 excludes action-dependent state trajectories. For probability measures Q, P, define

$$\mathrm{KL}(Q \,\|\, P) = \int \log\left(\frac{\mathrm{d}Q}{\mathrm{d}P}\right) \mathrm{d}Q, \tag{34}$$

when $Q \ll P$, and $+\infty$ otherwise. Here $Q \ll P$ denotes absolute continuity and $\mathrm{d}Q/\mathrm{d}P$ the relative density. For $P = \mu$, KL measures prior-relative comparator complexity.

Exponential weights initialized by μ uses the normalizer:

$$Z_t = \int_{\mathcal{A}} \exp\left(-\eta \sum_{r=1}^{t-1} L_r(a)\right) \mathrm{d}\mu(a), \tag{35}$$

and sampling measure p_t with density

$$\frac{\mathrm{d}p_t}{\mathrm{d}\mu}(a) = \frac{\exp(-\eta \sum_{r=1}^{t-1} L_r(a))}{Z_t}. \tag{36}$$

Here $Z_1 = 1$ and $p_1 = \mu$. This update is measurable-space Hedge (Littlestone & Warmuth, 1994; Freund & Schapire, 1997; Cesa-Bianchi & Lugosi, 2006).

Theorem 1 gives the arbitrary-rate and optimized bounds. The proofs of the selection results below appear in Appendix B.2.

Corollary 2 (Product-prior complexity decomposition)**.** *For $\mathcal{A} = \mathcal{Z} \times \mathcal{W}$, $\mu = \pi \otimes \rho$, and finite-KL Q with marginals Q_Z, Q_W,*

$$\mathrm{KL}(Q \,\|\, \pi \otimes \rho) = \mathrm{KL}(Q_Z \,\|\, \pi) + \mathrm{KL}(Q_W \,\|\, \rho) + I_Q(Z; W). \tag{37}$$

Here $I_Q = \mathrm{KL}(Q \,\|\, Q_Z \otimes Q_W)$ measures comparator dependence, not loss interaction.

For a product comparator $Q = Q_Z \otimes Q_W$, $I_Q(Z; W) = 0$. A frozen language model can induce the text prior, while the parameter prior may encode rank, precision, sparsity, or provenance.

Corollary 3 (Prior-atom description-length regret)**.** *Under the product prior $\mu = \pi \otimes \rho$, for positive-mass atoms $z^\star, w^\star$, put $C_\pi = -\log \pi(\{z^\star\})$ and $C_\rho = -\log \rho(\{w^\star\})$. Then $R_T(z^\star, w^\star) \le (C_\pi + C_\rho)/\eta + \eta T/8$, optimized to $\sqrt{T(C_\pi + C_\rho)/2}$ when the complexity is positive.*

For an autoregressive language-model prior with almost-sure termination or bounded serialization length, C_π is a text description length. Zero-mass points have infinite pointwise KL, so they require a nearby positive-mass comparator and a regularity condition.

Corollary 4 (Local point regret on continuous spaces)**.** *Let $\mathcal{A} = \mathcal{Z} \times \mathcal{W}$, $\mu = \pi \otimes \rho$, and let d_Z, d_W be metrics with measurable balls. Suppose every loss satisfies*

$$|L_t(z, w) - L_t(z', w')| \le G_Z d_Z(z, z') + G_W d_W(w, w'), \tag{38}$$

for nonnegative G_Z, G_W uniform over rounds. Fix $a^\star = (z^\star, w^\star)$ and radii $\delta_Z, \delta_W > 0$ whose balls $\mathbb{B}_Z = B(z^\star, \delta_Z)$ and $\mathbb{B}_W = B(w^\star, \delta_W)$ have positive respective prior mass. Set $C_{\delta_Z, \delta_W}(a^\star) = -\log \pi(\mathbb{B}_Z) - \log \rho(\mathbb{B}_W)$. For every $\eta > 0$,

$$R_T(a^\star) \le \frac{C_{\delta_Z, \delta_W}(a^\star)}{\eta} + \frac{\eta T}{8} + T(G_Z \delta_Z + G_W \delta_W). \tag{39}$$

Shrinking the balls reduces localization loss but may increase complexity. For positive local complexity, optimizing η gives

$$R_T(a^\star) \le \sqrt{\frac{T C_{\delta_Z, \delta_W}(a^\star)}{2}} + T(G_Z \delta_Z + G_W \delta_W). \tag{40}$$

For mixed spaces, replace one ball by an atom. The bound is sublinear if $C_{\delta_Z, \delta_W}(a^\star) = o(T)$ and $G_Z \delta_Z + G_W \delta_W = o(1)$. Without regularity, however, the distributional theorem does not control zero-mass points.

Learning-Rate and Computational Interpretation. The arbitrary-η result holds simultaneously over finite-KL comparators, whereas optimized rates are comparator-tuned. Grids, doubling, or parameter-free methods remove the need for advance knowledge at an additional regret cost (Orabona & Pál, 2016), although preserving low-refresh guarantees is nontrivial. The guarantee is statistical and does not require Gibbs normalization, sampling, or candidate proposal to be computationally efficient.

A.2.3 Mode Restriction and the Value of Joint Configurations

The all-admissible family values and gaps in Proposition 2 depend on Ξ, the evaluator, and T, independently of hard-target feasibility; see Definition 3. Joint access need not measure interaction: $L(z, w) = 1-0.4z-0.4w$ on $\{0, 1\}^2$ with base $(0, 0)$ has a positive joint-effect gap but zero factorial interaction. These gaps also differ from finite-candidate restriction in the Spider comparison and per-use target discrepancy ε_e; see Appendix D for the latter.

For mode-restricted selection under the Cartesian specialization, use slice priors:

$$\mu_Z = \pi \otimes \delta_{w_0} \quad \text{on } \mathcal{A}_Z, \tag{41}$$

$$\mu_W = \delta_{z_0} \otimes \rho \quad \text{on } \mathcal{A}_W. \tag{42}$$

These priors avoid conditioning on a possibly zero-mass product slice. Proposition 2 gives the exact selection–approximation–restriction identity and its EW bound. Appendix A.2.5 further splits comparator approximation into localization and discovery.

Corollary 5 (Mixture-prior mode selection)**.** *Let $\mathcal{E} \subseteq \{Z, W, ZW\}$ be nonempty, with priors μ_e supported on $\mathcal{A}_e$ and positive weights α_e summing to one. Extend each mode prior by zero to the union and define*

$$\mu_{\mathrm{mix}} = \sum_{e \in \mathcal{E}} \alpha_e \mu_e. \tag{43}$$

EW with this mixture prior satisfies, for fixed finite-KL $Q_e \ll \mu_e$,

$$R_T(Q_e) \le (\mathrm{KL}(Q_e \,\|\, \mu_e) + \log(1/\alpha_e))/\eta + \eta T/8. \tag{44}$$

For $e \in \{Z, W\}$, add the proposition's approximation and restriction gaps to compare with the joint optimum.

Overlapping modes contribute additively to the mixture. Choosing $\mu_{ZW} = \pi \otimes \rho$ preserves (37).

To keep mode identity observable, use the tagged lift:

$$\widetilde{\mathcal{A}}_{\mathcal{E}} = \bigsqcup_{e \in \mathcal{E}} (\{e\} \times \mathcal{A}_e), \tag{45}$$

$$\widetilde{\mu}_{\mathrm{mix}} = \sum_{e \in \mathcal{E}} \alpha_e (\delta_e \otimes \mu_e), \tag{46}$$

with projected loss $\widetilde{L}_t(e, a) = L_t(a)$. EW samples (e, a) and deploys a, retaining mode identity where families overlap. The lifted comparator has exact KL

$$\mathrm{KL}(\delta_e \otimes Q_e \,\|\, \widetilde{\mu}_{\mathrm{mix}}) = \mathrm{KL}(Q_e \,\|\, \mu_e) + \log(1/\alpha_e). \tag{47}$$

Without tags, overlap can reduce KL, so this certificate need not identify a posterior mode cost.

Equal weights charge at most $\log 2$ for two modes or $\log 3$ with the joint mode, while approximation or restriction can remain $\Theta(T)$. Full-information integration neither proposes candidates nor certifies feasibility or deployment cost. Bandit variants additionally introduce action-set dependence (Auer et al., 2002; Bubeck & Cesa-Bianchi, 2012).

A.2.4 Joint and Coordinatewise Selection

Use the finite Cartesian family, product prior, exogenous loss table, full feedback, and common initialization of Section 3. For a positive deterministic learning-rate schedule η_t, define the common target distribution:

$$q_t(z, w) = \frac{\pi(z)\rho(w) \exp\big(-\eta_t \sum_{r<t} d_r(z, w)\big)}{\sum_{(u,v)\in\mathcal{C}} \pi(u)\rho(v) \exp\big(-\eta_t \sum_{r<t} d_r(u, v)\big)}. \tag{48}$$

The constant choice $\eta_t = \eta$ recovers (3), whereas a varying η_t rescales the entire cumulative loss. A sweep preserves q_t when initialized from it, but the preceding configuration need not have that law.

Additivity and a product prior factorize the EW target; see Proposition 3. Bounded multiplicative losses $d_t(z, w) = g_t(z)h_t(w)$ generally do not factorize it, but nonadditivity alone need not separate regret.

Exact Constants for the Fixed-Rate Separation. For the barrier loss f of Proposition 4, the exact constants are as follows. For every fixed $\eta_t = \eta > 0$ and every $T \geq 1$, the selectors above satisfy

$$R_T^{\text{joint}} \leq \frac{3\Delta + 2b}{4} + \frac{\Delta}{e^{\eta\Delta} - 1} + \frac{2(\Delta + b)}{e^{\eta(\Delta+b)} - 1} \leq 1 + \frac{3}{\eta}, \tag{49}$$

$$R_T^{\text{coord}} \geq \frac{\Delta T}{4} \exp\left(-\frac{2}{e^{\eta b} - 1}\right). \tag{50}$$

The linear lower bound is not uniform as η vanishes. A slower schedule separates regret even when both selectors eventually reach the optimum.

Exact Constants for the Convergent Separation. Under Corollary 1, the value of $\eta_1 > 0$ does not affect uniform initialization. For $\gamma = \Delta/b > 1$ and $T \geq 2$,

$$R_T^{\text{joint}} \leq \frac{3\Delta + 2b}{4} + \Delta 4^{-\gamma}\left(1 + \frac{1}{\gamma - 1}\right) + 2(\Delta + b)4^{-(\gamma+1)}\left(1 + \frac{1}{\gamma}\right), \tag{51}$$

$$R_T^{\text{coord}} \geq \frac{\Delta e^{-1/2}}{4} \sum_{n=1}^{T-1} n^{-1/2} = \Omega(\sqrt{T}). \tag{52}$$

These separations concern the specified conditional sampler, not arbitrary alternating optimization. More sweeps, global refreshes, or deliberate transitions through higher-loss configurations can change its behavior. Exact joint sampling may also be costly, and these results establish no computational advantage. The controlled selection study in Section 6 tests interaction, sweep count, and learning rate by exact probability propagation.

A.2.5 Candidate Proposal, Discovery, and Selection

Candidate construction uses scores, traces, gradients, or critiques (Zhou et al., 2023; Deng et al., 2022; Pryzant et al., 2023; Yang et al., 2024a; Guo et al., 2024; Fernando et al., 2024; Yüksekgönül et al., 2025; Lee et al., 2026a;b). We separate discovery from selection on a fixed family. Changing families instead require tracking, sleeping-expert, or growing-expert guarantees (Herbster & Warmuth, 1998; Cortes et al., 2019; Mourtada & Maillard, 2017).

Let $\widehat{L}_{\text{train}}$ and $\widehat{L}_{\text{valid}}$ denote empirical losses on the declared training and validation data. A bilevel joint proposal fits, for each proposed harness z,

$$\widehat{w}(z) \approx \operatorname*{arg\,min}_{w:(z,w)\in\mathcal{A}} \widehat{L}_{\text{train}}(z, w), \tag{53}$$

then chooses

$$\widehat{z} \approx \operatorname*{arg\,min}_{z:\,\exists w\ (z,w)\in\mathcal{A}} \widehat{L}_{\text{valid}}\big(z, \widehat{w}(z)\big), \tag{54}$$

$$\widehat{a} = \big(\widehat{z}, \widehat{w}(\widehat{z})\big). \tag{55}$$

Outer validation search may be discrete and inner fitting gradient-based, but these approximate minimizations do not establish finite-budget convergence. Alternating procedures use (21)–(22): BetterTogether alternates prompt optimization and fine-tuning (Soylu et al., 2024), while WHALE alternates model updates and executable-harness search (Kim et al., 2026). They motivate, but differ from, the fixed-family conditional sampler in Section 3. Our guarantees apply after the family and prior are fixed.

For nonempty measurable $\mathcal{C}_e = \mathcal{C} \cap \mathcal{A}_e$, $e \in \{Z, W\}$, define

$$V_{T,\mathcal{C}}^e = \inf_{a\in\mathcal{C}_e} \sum_{t=1}^{T} L_t(a), \tag{56}$$

$$\mathsf{Disc}_T^e(\mathcal{C}_e) = V_{T,\mathcal{C}}^e - V_T^e \geq 0, \tag{57}$$

$$\Lambda_{T,\mathcal{C}}^e(Q_e) = \sum_{t=1}^{T} \int_{\mathcal{C}_e} L_t(a)\,\mathrm{d}Q_e(a) - V_{T,\mathcal{C}}^e \geq 0, \tag{58}$$

where Q_e is supported on $\mathcal{C}_e$. Discovery charges the candidate family's gap to the all-admissible optimum V_T^e, while localization charges Q_e's gap to the best candidate in the family. Substituting

$\mathsf{Apx}^e_T(Q_e) = \Lambda^e_{T,\mathcal{C}}(Q_e) + \mathsf{Disc}^e_T(\mathcal{C}_e)$ into (7) gives

$$\sum_{t=1}^{T} \overline{L}_t - V_T^{ZW} = \underbrace{\left[\sum_{t=1}^{T} \overline{L}_t - \sum_{t=1}^{T} \int_{\mathcal{C}_e} L_t(a)\,\mathrm{d}Q_e(a)\right]}_{\text{selection}} + \underbrace{\Lambda^e_{T,\mathcal{C}}(Q_e)}_{\text{localization}} + \underbrace{\mathsf{Disc}^e_T(\mathcal{C}_e)}_{\text{discovery}} + \underbrace{\Delta^{\text{mode}}_e(T) + \Delta^{\text{joint}}(T)}_{\text{mode and joint restriction}}. \tag{59}$$

A family prior controls only selection. Localization vanishes for a point mass on the best atomic candidate, while discovery remains relative to the proposals and budget. Non-Cartesian families require a joint prior and KL.

A.3 State Effects and Policy Regret

Bounded influence and full diagonal feedback control the state term in (2), and resets permit whole-episode policy selection. Section 5 weakens the feedback assumptions.

A.3.1 Persistent State and Policy Regret

Section 4 defines policy regret using each fixed comparator's own trajectory. Propositions 5 and 6 show why evaluating all configurations on the learner's state can conceal irreversible damage and why unrestricted persistent state precludes sublinear guarantees. The constructions and proofs are in Appendix B.3. Relabeling gives the parameter-only obstruction when that mode admits the same transitions, although actual text and parameter updates need not be symmetric.

A.3.2 Policy Regret under Bounded Influence

Bound each coordinate's influence while allowing interactions within its history window (Arora et al., 2012; Anava et al., 2015). Fix a prior μ on $\mathcal{A} = \mathcal{Z} \times \mathcal{W}$ or a measurable compatibility subset supporting every refresh.

The functions in Assumption 1 have domains

$$f_t : \mathcal{Z}^{\min\{m_Z+1,t\}} \times \mathcal{W}^{\min\{m_W+1,t\}} \to [0,1]. \tag{60}$$

Windows include the current coordinate, so zero horizon excludes carryover. Bundles use their largest horizon. For $e \in \{Z, W\}$, write m_e for its coordinate horizon, truncated at $T-1$ without changing the windows.

For $a = (z, w) \in \mathcal{A}$, define the diagonal loss:

$$d_t(a) = f_t\left(\underbrace{z, \ldots, z}_{\min\{m_Z+1,t\}} \;;\; \underbrace{w, \ldots, w}_{\min\{m_W+1,t\}}\right). \tag{61}$$

This specializes (1). Assumption 2 requires valid simulation or cloning of each constant history. Ordinary full information on the learner's realized history does not reveal these losses.

Lazy Exponential Weights. Apply exponential weights (36) to the diagonal oracle, coupling draws by measurable Shrinking Dartboard to control state (Geulen et al., 2010; Kalai & Vempala, 2005). For $\eta > 0$, set $L_t = d_t$ in (36) to define the target marginal q_t, and draw $A_1 = (z_1, w_1) \sim \mu$. After deploying $A_t = (z_t, w_t)$ and observing d_t, update

$$\zeta_t = \int_{\mathcal{A}} e^{-\eta d_t(a)}\,\mathrm{d}q_t(a), \tag{62}$$

$$q_{t+1}(\mathrm{d}a) = \frac{e^{-\eta d_t(a)}}{\zeta_t} q_t(\mathrm{d}a). \tag{63}$$

For $t < T$, retain A_t with probability $e^{-\eta d_t(A_t)}$ and otherwise *refresh* from q_{t+1}, with indicator Y_t. Refreshes may redraw A_t, so the number of switches is at most $\sum_{t<T} Y_t$. Under $A_t \sim q_t$, ζ_t is the mean retention probability. At $\eta = 0$, draw once from μ and never refresh.

Lemma 1 (Lazy exponential-weights coupling)**.** *The lazy algorithm has marginals $A_t \sim q_t$. For $t = 1, \dots, T-1$,*

$$\mathbb{E}[Y_t] \le \min\left\{\eta\,\mathbb{E}[d_t(A_t)], 1 - e^{-\eta}\right\} \le \eta, \tag{64}$$

$$\mathbb{E}\left[\sum_{t=1}^{T-1} Y_t\right] \le \eta(T-1). \tag{65}$$

To control the signed state effect isolated by Proposition 7, count departures from constant histories.

Lemma 2 (Contamination accounting)**.** *Under Assumption 1, for any realized sequence $A_t = (z_t, w_t)$, put*

$$\chi_s = \max\left\{m_Z \mathbf{1}\{z_s \ne z_{s-1}\}, m_W \mathbf{1}\{w_s \ne w_{s-1}\}\right\}, \qquad s = 2, \dots, T. \tag{66}$$

The charge χ_s uses the larger horizon for simultaneous changes. Let N_T^{cont} count rounds with a nonconstant coordinate history. Pathwise,

$$\left|L_T^{\mathrm{dep}} - L_T^{\mathrm{diag}}\right| \le N_T^{\mathrm{cont}} \le \sum_{s=2}^{T} \chi_s. \tag{67}$$

Overlap can make the summed charges loose. For the lazy algorithm,

$$\sum_{s=2}^{T} \chi_s \le m_{\max} \sum_{t=1}^{T-1} Y_t. \tag{68}$$

Theorem 2 combines these selection and contamination bounds, with the exact learning rate in (12).

For mixture prior (43), with each μ_e supported on its slice $\mathcal{A}_e$ and $Q_e \ll \mu_e$, replace C_Q by

$$\overline{C}_e^{\mathrm{mix}} = \mathrm{KL}(Q_e \,\|\, \mu_e) + \log(1/\alpha_e), \tag{69}$$

which is exact on the tagged lift and may exceed untagged KL. Whole-pair refreshes still charge $m_{\max}$, even when comparator complexity concentrates on one mode.

Corollary 6 (Lazy local point policy regret)**.** *Under Theorem 2, suppose $\mu = \pi \otimes \rho$ and the diagonal losses satisfy the Lipschitz condition of Corollary 4. For its metrics, positive-mass balls, and local complexity, every $\eta > 0$ gives*

$$R_T^{\mathrm{pol}}(a^\star) \le C_{\delta_Z, \delta_W}(a^\star)/\eta + \eta\,[T/8 + m_{\max}(T-1)] + T(G_Z \delta_Z + G_W \delta_W). \tag{70}$$

For positive local complexity, optimizing η gives

$$R_T^{\mathrm{pol}}(a^\star) \le 2\sqrt{C_{\delta_Z, \delta_W}(a^\star)[T/8 + m_{\max}(T-1)]} + T(G_Z \delta_Z + G_W \delta_W). \tag{71}$$

At zero local complexity, one conditioned-prior draw leaves only localization loss, whereas a product-prior atom needs no localization.

Proposition 8 in Section 4 gives the mode-restricted bounds for the supported slice priors (41)–(42). Zero KL permits a fixed prior draw, atoms give description lengths, and balls add localization. Proofs are in Appendix B.3.

A.3.3 Episodic Optimization with Resets

Consider K episodes of H rounds, each in a fresh sandbox with common resets

$$s_{k,1} = s_0, \tag{72}$$

$$a_{k,0} = a_0 = (z_0, w_0). \tag{73}$$

A policy φ maps within-episode history to $a_{k,h} = (z_{k,h}, w_{k,h})$ through admissible updates, whose modes may differ in effect; see Definition 3. Fix a standard-Borel policy family $(\mathcal{P}, \mathcal{F}_{\mathcal{P}})$ with prior Π_0.

For expected evaluation loss $\ell_{k,h}(s, a) \in [0, 1]$, define episode loss

$$J_k(\varphi) = \mathbb{E}_\varphi\left[\sum_{h=1}^{H} \ell_{k,h}(s_{k,h}, a_{k,h})\right]. \tag{74}$$

The expectation includes execution, transitions, and policy randomness from the common reset. Assume measurable $J_k : \mathcal{P} \to [0, H]$ are fixed before learner randomization, with exogenous tasks.

Given pre-selection history $\mathcal{H}_{k-1}$, including completed-episode feedback, define

$$\overline{J}_k = \mathbb{E}[J_k(\varphi_k) \mid \mathcal{H}_{k-1}]. \tag{75}$$

This expectation also averages over policy selection, as in (30). For a fixed policy $\varphi^\star$ or comparator probability measure $Q_\mathcal{P}$ on $\mathcal{P}$, define

$$R_K^{\mathrm{epi}}(\varphi^\star) = \sum_{k=1}^K \overline{J}_k - \sum_{k=1}^K J_k(\varphi^\star), \tag{76}$$

$$R_K^{\mathrm{epi}}(Q_\mathcal{P}) = \sum_{k=1}^K \overline{J}_k - \sum_{k=1}^K \int_\mathcal{P} J_k(\varphi)\,\mathrm{d}Q_\mathcal{P}(\varphi). \tag{77}$$

The comparator holds one policy draw across episodes, while allowing within-episode configuration changes. Thus (74) extends the own-trajectory comparison to policies.

Specializing (36) to $\mathcal{P}$, sample the episode-k policy from

$$\frac{\mathrm{d}q_k}{\mathrm{d}\Pi_0}(\varphi) = \frac{\exp(-\eta \sum_{i=1}^{k-1} J_i(\varphi))}{\int_\mathcal{P} \exp(-\eta \sum_{i=1}^{k-1} J_i(\psi))\,\mathrm{d}\Pi_0(\psi)}. \tag{78}$$

Execute it for the episode, then reveal J_k for every policy, including unexecuted ones.

Theorem 5 (Episodic policy prior-complexity regret)**.** *Under the preceding reset, exogeneity, measurability, and full-feedback assumptions, for every $\eta > 0$ and fixed $Q_\mathcal{P} \ll \Pi_0$ with finite $C_\mathcal{P} = \mathrm{KL}(Q_\mathcal{P} \| \Pi_0)$, episode-level EW satisfies*

$$R_K^{\mathrm{epi}}(Q_\mathcal{P}) \le \frac{C_\mathcal{P}}{\eta} + \frac{\eta K H^2}{8}. \tag{79}$$

For $C_\mathcal{P} > 0$, $\eta = \sqrt{8C_\mathcal{P}/(KH^2)}$ gives $R_K^{\mathrm{epi}} \le H\sqrt{KC_\mathcal{P}/2}$.

The proof is in Appendix B.3. Predictably reactive episode losses instead yield realized-sequence external pseudo-regret.

As in Corollary 3, atoms give description length $-\log \Pi_0(\{\varphi^\star\})$. For measurable metric balls and episode losses uniformly $G_\mathcal{P}$-Lipschitz in policy distance, conditioning on the radius-δ ball $B(\varphi^\star, \delta)$ gives complexity $-\log \Pi_0(B(\varphi^\star, \delta))$ and localization loss $KG_\mathcal{P}\delta$, as in Corollary 4.

Restricted policy families may still incur the gaps in Section 3.

A.4 Feedback Requirements and Extensions

The feedback interface determines control of selection and state in (2). Full diagonal feedback evaluates each candidate's own trajectory, sparse probing reduces evaluations, and deployment-only learning requires control of causal reaction. Appendix A.4.10 specifies when the guarantees retain the fixed joint-family comparator. These guarantees neither construct candidates nor rank unrestricted families.

Sections 4–5 present the main guarantees for full counterfactual feedback, sparse probes, and deployment-only feedback. Here K counts candidates, not episodes as in Appendix A.3.3. We use $C_Q = \mathrm{KL}(Q\|\mu)$ and $m = m_{\max}$ for comparator complexity and the maximum influence horizon, respectively. The comparator-tuned rates concern expectations, while the exact bounds cover zero complexity, zero influence, and finite-horizon boundaries.

Exponential weights and importance weighting (Auer et al., 2002), low-switch couplings (Geulen et al., 2010; Anava et al., 2015), and policy-regret mini-batching (Arora et al., 2012) supply the algorithmic ingredients. Decoupled exploration and deployment (Avner et al., 2012) and limited-observation Shrinking Dartboard (Cheng et al., 2023) provide further precedents.

A.4.1 Accuracy of Full Diagonal Evaluation

These results relax Assumption 2 conditional on an exogenous response tree. For random trees, the relevant independence, mean, or error assumptions must hold conditionally. Unconditional unbiasedness does not justify a tree-dependent hindsight comparator; see Appendix A.4.10.

Corollary 7 (Approximate diagonal evaluation)**.** *Under Assumption 1, let the entire estimate sequence be independent of all learner randomness, with $\widehat{d}_t \in [0,1]$ and $\sup_a |\widehat{d}_t(a) - d_t(a)| \le \epsilon_t$*

almost surely for fixed $\epsilon_t \geq 0$. Lazy EW using these estimates satisfies, for every $\eta > 0$ and $Q \ll \mu$ with finite C_Q,

$$R_T^{\mathrm{pol}}(Q) \leq \frac{C_Q}{\eta} + \eta \left[\frac{T}{8} + m_{\max}(T-1) \right] + 2 \sum_{t=1}^{T} \epsilon_t. \tag{80}$$

Independence from only the current draw is insufficient. Off-policy estimates require additional coverage, variance, and validity assumptions (Dudík et al., 2011; Jiang & Li, 2016). Proofs of the feedback results are in Appendix B.4.

Corollary 8 (Finite-family sampling cost for diagonal estimation)**.** *Let μ be supported on a K-candidate family $\mathcal{C} \subseteq \mathcal{A}$, restricting the learner and comparators $Q \ll \mu$ to it. Fix $\delta \in (0,1)$. Estimate each $d_t(a)$ by the mean of $n \geq 1$ independent $[0,1]$ samples with mean $d_t(a)$, using a sampling mechanism exogenous to all learner randomness. Define*

$$\epsilon_n = \sqrt{\frac{\log(2KT/\delta)}{2n}}. \tag{81}$$

With probability at least $1-\delta$, all KT estimates satisfy $|\widehat{d}_t(a) - d_t(a)| \leq \epsilon_n$ simultaneously. Under the remaining assumptions of Corollary 7, policy regret averaged over estimation and learner randomness satisfies

$$R_T^{\mathrm{pol}}(Q) \leq \frac{C_Q}{\eta} + \frac{\eta T}{8} + m_{\max}\eta(T-1) + 2T\epsilon_n + \delta T. \tag{82}$$

The estimator uses KTn oracle samples.

Uniform accuracy ϵ is guaranteed by $n \geq \log(2KT/\delta)/(2\epsilon^2)$, and the envelope can be capped at one. Oracle costs must include simulation or cloning, while logged language-model counterfactuals need separate unbiasedness justification. With fixed samples per pair, the envelope need not vanish, so this certificate alone does not imply sublinear regret.

Corollary 9 (Fresh unbiased diagonal feedback)**.** *Fix a finite compatible K-candidate family with positive prior μ. Under Assumption 1, let the entire estimate sequence be independent of all learner randomness, with $\widehat{d}_t \in [0,1]$ and*

$$\mathbb{E}[\widehat{d}_t(a) \mid \widehat{d}_{1:t-1}] = d_t(a). \tag{83}$$

For lazy EW using these estimates, every $\eta > 0$ and fixed Q satisfy, averaging over estimation and learner randomness,

$$R_T^{\mathrm{pol}}(Q) \leq \frac{C_Q}{\eta} + \eta \left[\frac{T}{8} + m_{\max}(T-1) \right]. \tag{84}$$

Fresh independent oracle samples satisfy Corollary 9 at $n = 1$. Larger n improves uniform accuracy but is unnecessary for this coarse bound. Biased or trajectory-dependent estimates do not inherit it. The $2\sum_t \epsilon_t$ error bridge compares true and estimated learner and comparator losses for the *same* policy, not learners trained on different feedback. An offset $\widehat{d}_t(a) = d_t(a) + c_t \in [0,1]$ preserves EW marginals but multiplies retention (63) by $e^{-\eta c_t}$, changing state effects; see the feedback accuracy study in Section 6.

A.4.2 Causal Responses and Comparator Conventions

Fix a nonempty finite compatible family $\mathcal{C} \subseteq \mathcal{A}$, $K = |\mathcal{C}|$, and prior $\mu(a) > 0$. Neither a Cartesian family nor a factorized prior is required.

Definition 5 (Causal response and expected policy regret)**.** *Let U be an external random tape independent of the learner's private randomization, and let $F_t : \mathcal{C}^t \times \mathcal{U} \to [0,1]$ be the measurable causal response functions from Section 2. Realized loss is $F_t(A_{1:t}; U)$. For a constant history $a^t = (a, \ldots, a)$, write $d_t(a; U) = F_t(a^t; U)$, making the tape explicit in* (1)*. Policy regret is as defined in Section 4, averaging over U and learner randomness, with Q fixed independently of learner observations and randomization.*

Each intervention reruns task selection, user responses, and environment state through F_t. This defines counterfactuals but supplies no evaluation oracle. The environment may observe actions and outcomes, not future private coins. For observed announcements or messages, their rules and intervention-dependent values must also be specified.

A deterministic tree's best fixed candidate is admissible. Random environments distinguish best expected loss from realized hindsight:

$$\min_a \mathbb{E} D_T(a) \neq \mathbb{E} \min_a D_T(a). \tag{85}$$

Conditioning on an exogenous tree can justify hindsight comparison; see Appendix A.4.10. Conditioning on noisy probes, however, cannot justify a probe-dependent comparator in an unbiasedness argument.

A.4.3 Predictability Does Not Justify Conditional Lazy Marginals

Predictability and the bounded influence specified in Assumption 1 do not supply conditional lazy marginals: a retained action can correlate with losses chosen from its past despite matching unconditional marginals.

Proposition 9 (Two-round counterexample)**.** *A predictable second-round loss vector can satisfy $\mathbb{E}\ell_2(A_2) \neq \mathbb{E}\langle q_2, \ell_2\rangle$ under lazy sampling, even when A_2 has unconditional marginal q_2.*

For $F_2(a_1, a_2) = \mathbf{1}\{a_1 = a_2\}$, both diagonal losses are $d_2(a) = 1$. The counterexample invalidates substituting current-action losses on the realized prefix, not Theorem 2's diagonal oracle.

Proposition 10 (External regret versus few switches)**.** *Let $\ell_1 \equiv 0$ and $\ell_t(a) = \mathbf{1}\{a = A_{t-1}\}$ for $t \geq 2$, and let*

$$S_T = \sum_{t=2}^{T} \mathbf{1}\{A_t \neq A_{t-1}\}. \tag{86}$$

Every learner satisfies

$$\mathbb{E}\left[\sum_t \ell_t(A_t) - \min_{a\in\{0,1\}} \sum_t \ell_t(a)\right] \geq \frac{T-1}{2} - \mathbb{E}S_T. \tag{87}$$

Constant-trajectory comparators lose $T - 1$, excluding a policy-regret lower bound but motivating fresh block draws under reactive feedback; see Appendix A.4.7.

A.4.4 Exponential Weights with Nonnegative Estimates

Importance-weighted estimates may exceed one, invalidating the bounded-feedback term $\eta T/8$. A standard second-order potential argument instead controls selection by their second moments (Auer et al., 2002; Cheng et al., 2023).

Lemma 3 (Prior-EW with nonnegative estimates)**.** *Let $x_t(a) \geq 0$ be finite, let $\eta > 0$, and define*

$$q_t(a) = \frac{\mu(a)\exp(-\eta\sum_{s<t} x_s(a))}{\sum_c \mu(c)\exp(-\eta\sum_{s<t} x_s(c))}. \tag{88}$$

For every distribution Q on $\mathcal{C}$, pathwise,

$$\sum_{t=1}^{T}\langle q_t - Q, x_t\rangle \leq \frac{C_Q}{\eta} + \frac{\eta}{2}\sum_{t=1}^{T}\langle q_t, x_t^2\rangle. \tag{89}$$

The sequence x_t may be adaptively generated.

Actual-loss bounds additionally require regime-specific sampling identities, unbiasedness, and comparator validity.

A.4.5 Sparse Independent Diagonal Probes

Let A_t denote deployment, J_t the probe, Z_t its outcome, and $R_t \in \{0, 1\}$ the lazy refresh indicator.

Assumption 3 (Independent probes of an exogenous diagonal table)**.** *Conditional on an exogenous response tree U, the following requirements hold. Assumption 1 holds. An autonomous evaluator uses randomness independent of all deployment coins and consumes no deployed actions or outcomes. Its pre-probe sigma-field $\mathcal{G}_{t-1}$ includes U and earlier probes, and may include evaluation context hidden from the learner. The selector q_t uses only previous estimates, and positive r_t is $\mathcal{G}_{t-1}$-measurable. Drawing $J_t \sim r_t$, the evaluator returns $Z_t \in [0, 1]$ with*

$$\mathbb{E}[Z_t \mid \mathcal{G}_{t-1}, J_t = a] = d_t(a). \tag{90}$$

Probing leaves the response tree unchanged, and deployment intervention leaves the entire evaluation process unchanged.

Independent simulation noise on an exogenous table qualifies, whereas deployment-reactive evaluation does not. Proposition 11 permits probe-driven adaptation with narrower comparators.

Under Assumption 3, initialize the selector with $q_1 = \mu$ and $A_1 \sim \mu$. At round t, execute A_t, obtain the independent probe (J_t, Z_t), and form

$$\widehat{d}_t(a) = \frac{\mathbf{1}\{J_t = a\} Z_t}{r_t(a)}. \tag{91}$$

Update

$$q_{t+1}(a) = \frac{q_t(a) e^{-\eta \widehat{d}_t(a)}}{\sum_c q_t(c) e^{-\eta \widehat{d}_t(c)}}. \tag{92}$$

For $t < T$, retain $A_{t+1} = A_t$ with probability $e^{-\eta \widehat{d}_t(A_t)}$ and otherwise draw afresh from q_{t+1}. Draws use the independent deployment tape and return no deployment-loss feedback. A refresh need not switch configurations.

Lemma 4 (Probe mean and second moment)**.** *Under Assumption 3,*

$$\mathbb{E}[\widehat{d}_t(a) \mid \mathcal{G}_{t-1}] = d_t(a), \tag{93}$$

$$\mathbb{E}[\langle q_t, \widehat{d}_t^2 \rangle \mid \mathcal{G}_{t-1}] = V_t, \tag{94}$$

where

$$V_t = \sum_a \frac{q_t(a)}{r_t(a)} \mathbb{E}[Z_t^2 \mid \mathcal{G}_{t-1}, J_t = a]. \tag{95}$$

In particular, $r_t = q_t$ gives $V_t \le K$.

Lemma 5 (Conditional lazy marginals and refreshes)**.** *Under Assumption 3, conditional on the full auxiliary tape and the fixed response tree, A_t has law q_t. Moreover,*

$$\mathbb{E} R_t \le \eta \mathbb{E}\langle q_t, d_t \rangle \le \eta, \qquad 1 \le t < T. \tag{96}$$

Conditioning requires the entire autonomous process, not current-coin independence; see Proposition 9.

Refined Bound for Arbitrary Probe Distributions. Theorem 3 in Section 5 specializes a more detailed inequality. Under Assumption 3, for $m = m_{\max}$, every fixed Q and $\eta > 0$ satisfy

$$R_T^{\mathrm{pol}}(Q) \le \frac{C_Q}{\eta} + \frac{\eta}{2} \sum_{t=1}^{T} \mathbb{E} V_t + m\eta \sum_{t=1}^{T-1} \mathbb{E}\langle q_t, d_t \rangle. \tag{97}$$

An infinite expected second moment makes this bound vacuous. For $r_t = q_t$, all moments are finite, and (16)–(17) follow. Text-only and parameter-only families use m_Z and m_W, and non-Cartesian families use their supported joint prior. Tuning is comparator-dependent, but uniform priors give $C_Q = \log K$ for every point comparator.

Cloning, replay, constant-history initialization, and simulation costs remain. Finite influence supplies no simulator, and the expected bound guarantees neither low variance nor high-probability performance.

Proposition 11 (Autonomous query-adaptive evaluation)**.** *The sparse-probe bounds hold for fixed Q if the auxiliary process chooses response functions from earlier probes and q_t, independently of deployment coins and without consuming deployed actions or outcomes. The current diagonal vector and r_t must be $\mathcal{G}_{t-1}$-measurable and satisfy* (90)*. Conditional on the complete auxiliary process, responses must have diagonal d_t and the stated coordinate influence bounds. Deployment intervention must leave that entire process unchanged.*

This allows probe-driven curricula, not deployment reactions or realized query-dependent comparators. The best expected-loss candidate is fixed before probes and remains admissible under (85). The budget and bias extensions below inherit this fixed-Q restriction for query-adaptive trees.

A.4.6 Probe Budgets and Biased Evaluation

Corollary 10 (Multiple probes per round)**.** *Hold the current evaluator context and q_t fixed while drawing $k \ge 1$ conditionally independent probes with $r_t = q_t$. If $\widehat{d}_t$ is the average of their importance-weighted vectors, then*

$$\mathbb{E}[\langle q_t, \widehat{d}_t^2 \rangle \mid \mathcal{G}_{t-1}] \le 1 + \frac{K-1}{k}. \tag{98}$$

Consequently K in (16) *can be replaced by $1 + (K-1)/k$, while the expected refresh bound is unchanged.*

Sampling is with replacement, whereas querying every candidate once has a different estimator law.

Corollary 11 (Expected query budget)**.** *Use an independent Bernoulli gate G_t with probability $p \in (0,1]$ and, when it opens, one probe drawn from q_t. Set the estimate to zero when the gate is closed. The resulting estimator,*

$$\widehat{d}_t(a) = \frac{G_t \mathbf{1}\{J_t = a\} Z_t}{p q_t(a)}, \tag{99}$$

is unbiased, with weighted second moment at most K/p. Writing $B_{\mathrm{qry}} = pT$ for the expected *number of probes, the optimized bound for positive C_Q is*

$$R_T^{\mathrm{pol}}(Q) \le 2\sqrt{C_Q\left(\frac{KT^2}{2B_{\mathrm{qry}}} + m(T-1)\right)}. \tag{100}$$

A hard cap makes later inclusion probabilities zero, invalidating Corollary 11's argument. Fixed-budget designs therefore need separate analysis (Cheng et al., 2023).

Simulator bias separates the true selection target from the mean driving retention.

Proposition 12 (Biased-probe bound)**.** *Retain probe autonomy, but suppose*

$$\mathbb{E}[Z_t \mid \mathcal{G}_{t-1}, J_t = a] = \bar{d}_t(a), \tag{101}$$

$$\sup_a |\bar{d}_t(a) - d_t(a)| \le \epsilon_t, \tag{102}$$

where $\epsilon_t \ge 0$ is predictable, $Z_t \in [0,1]$, and V_t is the actual weighted second moment of the estimator. Then

$$R_T^{\mathrm{pol}}(Q) \le \frac{C_Q}{\eta} + \frac{\eta}{2}\sum_{t=1}^{T}\mathbb{E}V_t + 2\sum_{t=1}^{T}\mathbb{E}\epsilon_t + m\eta\sum_{t=1}^{T-1}\mathbb{E}\langle q_t, \bar{d}_t\rangle. \tag{103}$$

In particular, for $r_t = q_t$,

$$R_T^{\mathrm{pol}}(Q) \le \frac{C_Q}{\eta} + \frac{\eta KT}{2} + m\eta(T-1) + 2\sum_{t=1}^{T}\mathbb{E}\epsilon_t. \tag{104}$$

Replacing $\bar{d}_t$ by d_t in the refresh term requires an additional $m\eta\sum_{t<T}\mathbb{E}\epsilon_t$: even when $d_t \equiv 0$, positive simulator means can trigger refreshes.

A.4.7 Deployment-Only Feedback and Causal Reaction

Under Definition 5, the learner observes only actual deployment losses. Use the block notation and counterfactual losses of Definition 1 in Section 5. Conditional on U and the past, g_j precedes the draw and includes within-block reactions, but only its executed coordinate is observed.

Block Exponential Weights. Initialize $p_1 = \mu$. For each block j, deploy a fresh private draw $B_j \sim p_j$ for b rounds, observe $Z_j = g_j(B_j)$, and set

$$\widehat{g}_j(a) = \frac{\mathbf{1}\{B_j = a\} Z_j}{p_j(a)}, \tag{105}$$

$$p_{j+1}(a) = \frac{p_j(a) e^{-\eta \widehat{g}_j(a)}}{\sum_c p_j(c) e^{-\eta \widehat{g}_j(c)}}. \tag{106}$$

Retain the last configuration for the r remaining rounds, giving at most $J-1$ switches. Block draws are fresh, not lazy. Exact probabilities stay positive, but numerical underflow remedies involving clipping or exploration require corresponding bias or exploration terms.

Fresh draws justify sampling identities against adaptive prefixes.

Lemma 6 (Bandit selection on block losses)**.** *For every fixed Q and $\eta > 0$,*

$$\mathbb{E}\sum_{j=1}^{J}\left[g_j(B_j) - \langle Q, g_j\rangle\right] \le \frac{C_Q}{\eta} + \frac{\eta KJ}{2}. \tag{107}$$

Adding the causal reaction (19) and the remainder charge to Lemma 6 gives Theorem 4 in Section 5.

A.4.8 Conditions Controlling Causal Reaction

Assumption 4 (Total component-specific finite influence)**.** *For every fixed U, the complete causal response F_t depends only on the coordinate windows in Assumption 1, including all task-selection and evaluator rules, from a common initial condition.*

Forgetting in a cache is insufficient if task selection remembers the first action indefinitely, because that dependence belongs in F_t.

Lemma 7 (Prefix charge under finite influence)**.** *Under Assumption 4, with $m = m_{\max}$,*

$$|\Gamma_b(Q)| \le \min\{m, b\}(J-1). \tag{108}$$

Corollary 12 (Deployment-only bounded-influence rate)**.** *Under Assumption 4, for every $\eta > 0$,*

$$R_T^{\mathrm{pol}}(Q) \le \frac{bC_Q}{\eta} + \frac{\eta bKJ}{2} + \min\{m, b\}(J-1) + r. \tag{109}$$

For $C_Q > 0$, choosing $\eta = \sqrt{2C_Q/(KJ)}$ gives

$$\begin{aligned} R_T^{\mathrm{pol}}(Q) &\le b\sqrt{2KJC_Q} + \min\{m, b\}(J-1) + r \\ &\le \sqrt{2KC_QTb} + \frac{mT}{b} + b. \end{aligned} \tag{110}$$

For fixed positive m, K, C_Q and sufficiently large T, integer $b \asymp (m^2T/(KC_Q))^{1/3}$, truncated to $[1, T]$, gives

$$R_T^{\mathrm{pol}}(Q) = O((mKC_Q)^{1/3}T^{2/3}). \tag{111}$$

For $m = 0$, use $b = 1$, and for $C_Q = 0$, draw once from μ.

Use (109) when parameters grow with T. The zero-influence rate is $\sqrt{2KTC_Q}$, and a singleton family has zero regret. For zero KL, fresh block resampling need not have zero regret, unlike a single prior draw. Text-only, parameter-only, and joint or cross-mode blocks use m_Z, m_W, and m, respectively.

Equation (109) already includes prefix effects through $b\sum_j g_j(B_j)$, so another contamination term would double-count. The established $T^{2/3}$ rate (Arora et al., 2012) cannot universally be improved to $\sqrt{T}$ under deployment-only feedback (Dekel et al., 2014). Appendix B.4 gives the reduction to bandits with switching costs. This leaves optimal m and prior dependence in (111) open. Sparse probes provide stronger information.

To separate initial warm-up from persistent reaction, fix an integer $h_0 \ge 0$ and define

$$\mathcal{A}_b(Q) = \mathbb{E}\sum_{j=2}^{J}\sum_{a} Q(a) \sum_{u=h_0+1}^{b} \left| F_{s_j+u-1}(H_j, a^u; U) - F_{s_j+u-1}(a^{s_j+u-1}; U) \right|, \tag{112}$$

with an empty sum if $b \le h_0$. Without assuming total finite influence,

$$R_T^{\mathrm{pol}}(Q) \le \frac{bC_Q}{\eta} + \frac{\eta bKJ}{2} + \min\{h_0, b\}(J-1) + \mathcal{A}_b(Q) + r. \tag{113}$$

Sublinear regret requires $\mathcal{A}_b(Q) = o(T)$ and appropriate b, η, neither of which the definition supplies.

A suffix-sensitivity bound can control reaction without exact forgetting.

Proposition 13 (A suffix-sensitivity envelope)**.** *Suppose deterministic numbers $\beta_u \in [0, 1]$ satisfy, for every admissible prefix h of length $s-1$, every a, U, and every relevant $u \ge 1$,*

$$\left| F_{s+u-1}(h, a^u; U) - F_{s+u-1}(a^{s+u-1}; U) \right| \le \beta_u. \tag{114}$$

With $H_b = \sum_{u=1}^{b} \beta_u$, the block algorithm satisfies

$$R_T^{\mathrm{pol}}(Q) \le \frac{bC_Q}{\eta} + \frac{\eta bKJ}{2} + H_b(J-1) + r. \tag{115}$$

If $H = \sum_{u \ge 1} \beta_u < \infty$, one may replace H_b by H.

The envelope must hold for all protocol prefixes, not merely observed paths. Contractive state with Lipschitz losses supplies one such envelope.

Assumption 5 (Contractive environment state)**.** *An environment state e_t lies in a metric space of diameter at most D, with*

$$e_{t+1} = \Psi_t(e_t, A_t; U_t), \tag{116}$$

$$\operatorname{dist}(\Psi_t(e, a; u), \Psi_t(e', a; u)) \le \rho \operatorname{dist}(e, e'), \quad 0 \le \rho < 1. \tag{117}$$

The $[0,1]$ loss depends on e_t, coordinate windows of lengths $m_Z + 1, m_W + 1$, and exogenous randomness. For identical windows and randomness, it is G-Lipschitz in e_t. Learner and comparator share the initial state, maps, and admissible external noise tape.

Corollary 13 (Contractive reaction without a counterfactual oracle)**.** *Under Assumption 5, let $m = m_{\max}$ and*

$$H_b = \min\{m, b\} + \sum_{u=m+1}^{b} \min\{1, GD\rho^{u-1}\}, \tag{118}$$

$$H = m + \frac{GD\rho^m}{1-\rho}. \tag{119}$$

Then (115) *holds, with $H_b \le \min\{b, H\}$. For fixed positive H, K, C_Q, the corresponding tuned order is $O((HKC_Q)^{1/3}T^{2/3})$.*

For $\rho = 0, m = 0$, set $\rho^0 = 1$: initial state mismatch can affect loss before contraction. For $H = 0$, use $b = 1$, not the cube-root formula.

For example, $e_{t+1} = \rho e_t + (1-\rho)v(A_t)$ on a bounded convex set models a deployment-dependent task mixture with Lipschitz expected loss. Its discrete sampled-task losses, however, need not be pathwise Lipschitz. Summable reaction replaces finite influence, relating to online control under mixing (Even-Dar et al., 2009), although Markov mixing alone does not imply pathwise contraction.

A.4.9 Identification from Logged Deployment Data

An ordinary one-step importance-weighted outcome,

$$\frac{\mathbf{1}\{A_t = a\} Z_t}{\beta_t(a \mid H_{t-1})}, \tag{120}$$

intervenes only on the current action, so it need not estimate own-trajectory loss (1). For $F_2(a_1, a_2) = \mathbf{1}\{a_1 \ne a_2\}$ after $a_1 = 1$, choosing $a_2 = 0$ loses one, whereas $d_2(0) = 0$. Sequential evaluation requires sequential intervention (Jiang & Li, 2016).

Lemma 8 (Window importance sampling in a sufficient model)**.** *Consider a fixed deterministic response table with total joint influence horizon m (for coordinate influence horizons, $m = m_{\max}$ is sufficient), $t \ge m+1$, and a behavior policy with known positive propensities $\beta_s(a \mid H_{s-1}) \ge \epsilon > 0$. Let the logged outcome $Z_t \in [0,1]$ equal the realized loss, or be conditionally unbiased for it given the full history through the current action. Put*

$$W_t^a = \prod_{s=t-m}^{t} \frac{\mathbf{1}\{A_s = a\}}{\beta_s(a \mid H_{s-1})}, \tag{121}$$

$$\widetilde{d}_t(a) = W_t^a Z_t. \tag{122}$$

Let $\mathcal{H}_{t-m-1}$ denote the logged history before the window. Then

$$\mathbb{E}[\widetilde{d}_t(a) \mid \mathcal{H}_{t-m-1}] = d_t(a), \tag{123}$$

$$\mathbb{E}[\widetilde{d}_t(a)^2 \mid \mathcal{H}_{t-m-1}] \le \epsilon^{-(m+1)}. \tag{124}$$

Uniform logging gives a second-moment bound K^{m+1} for this estimator, not a minimax lower bound. Stochastic systems also require consistency, sequential ignorability, and window identification of the full own trajectory. Overlapping windows need not be unbiased conditional on preceding estimates, and lazy logging violates Assumption 3. These issues require independent episodes, valid sample splitting, or delayed/blockwise analysis. Without intervention support, neither models nor doubly robust estimation alone identify diagonal values.

A.4.10 Transferring Bounds to the Fixed Joint Comparator

For a deterministic response tree, let $a_e^\star$ minimize the fixed-candidate loss D_T from Section 2 over $\mathcal{C}_e$, attaining $V_{T,\mathcal{C}}^e$. Adding family restriction to its policy regret gives

$$\mathbb{E}L_T^{\mathrm{dep}} - V_{T,\mathcal{C}}^{ZW} = R_T^{\mathrm{pol}}(a_e^\star) + V_{T,\mathcal{C}}^e - V_{T,\mathcal{C}}^{ZW}. \tag{125}$$

This groups selection and state in (2). Distributional comparators additionally incur their approximation gap to the mode optimum. The identity does not certify candidate discovery or reachability.

Conditioning on the exogenous tree fixes its best candidate before probing, making it admissible in Theorem 3 under Assumption 3. Theorem 4 likewise permits $Q(U)$ fixed conditional on U, and arbitrary-η bounds average conditional complexity and reaction over U. Unobserved realized complexity may prevent implementable tuning, but uniform priors give $\log K$ for every point comparator without requiring knowledge of its identity.

Proposition 11 permits fixed Q or a best expected-loss candidate, not a query-dependent comparator correlated with probe noise. Conditioning on the full transcript establishes deployment marginals, not conditional unbiasedness of realized estimates, and the optima in (85) remain distinct. Sparse probing therefore needs a valid counterfactual evaluator, and sublinear deployment-only regret needs controlled reaction (19). Implementation must also account for numerical stabilization and evaluation costs.

B PROOFS

This appendix proves the results stated in the main text and Appendix A. The proofs follow the same topic order: configurations and loss decomposition, selection and family restriction, state effects, and feedback requirements. The final subsection proves the cost result in Appendix D.

B.1 CONFIGURATIONS, COMPARATORS, AND LOSS DECOMPOSITION

Proof of Proposition 1. Add and subtract L_T^{diag} and $V_{T,\mathcal{C}}^e$. Family inclusion implies $V_{T,\mathcal{C}}^e \geq V_{T,\mathcal{C}}^{ZW}$. A changing selection can outperform every fixed candidate, and a mixed history can improve or worsen its diagonal loss. Thus, neither remaining term has a fixed sign. □

B.2 SELECTION AND FAMILY RESTRICTION

Proof of (6). Apply Theorem 1 on $\mathcal{C}_e$ with $L_t = d_t$, prior μ_e, and $Q = \delta_{a_e^\star}$. Its KL term is $\log(1/\mu_e(a_e^\star))$, and its regret is $\mathbb{E}L_T^{\mathrm{diag}} - V_{T,\mathcal{C}}^e$. Statelessness makes the state term in (2) zero, so adding family restriction gives (6). □

Proof of Theorem 1. Hoeffding's lemma (Hoeffding, 1963) and telescoping give

$$\log Z_{T+1} \leq -\eta \sum_{t=1}^{T} \overline{L}_t + \frac{\eta^2 T}{8}. \tag{126}$$

For $q = \mathrm{d}Q/\mathrm{d}\mu$, Jensen's inequality gives

$$\log Z_{T+1} \geq -\eta \int_{\mathcal{A}} \sum_{t=1}^{T} L_t(a)\,\mathrm{d}Q(a) - \mathrm{KL}(Q \,\|\, \mu). \tag{127}$$

Combining the bounds gives (4). Optimizing this bound for positive KL gives (5). If $\mathrm{KL}(Q \,\|\, \mu) = 0$, then $Q = \mu$, and the $\eta = 0$ extension $p_t = \mu$ has zero distributional pseudo-regret. □

Proof of Corollary 2. Apply Radon–Nikodym factorization and the KL chain rule through $Q_Z \otimes Q_W$. □

Proof of Corollary 3. Apply Theorem 1 to $Q = \delta_{(z^\star, w^\star)}$. For this comparator, (37) makes its KL the sum of the negative log atom masses. Zero complexity means the prior is concentrated on the reference. □

Proof of Corollary 4. Use $Q = \pi(\cdot \mid \mathbb{B}_Z) \otimes \rho(\cdot \mid \mathbb{B}_W)$, whose KL is $C_{\delta_Z, \delta_W}(a^\star)$. Equation (38) adds at most $G_Z\delta_Z + G_W\delta_W$ per round. Applying Theorem 1 and optimizing then gives the stated bound. □

Proof of Proposition 2. Add and subtract the family values and the Q_e-average cumulative loss defined in Section 3. The definitions of Δ_e^{mode}, Δ^{joint}, and $\mathsf{Apx}_T^e(Q_e)$ identify the resulting gaps, and Theorem 1 bounds the selection term on $\mathcal{A}_e$. □

Proof of Corollary 5. Since $\mu_{\mathrm{mix}} \geq \alpha_e \mu_e$ as measures, $Q_e \ll \mu_{\mathrm{mix}}$ and integration gives

$$\mathrm{KL}(Q_e \,\|\, \mu_{\mathrm{mix}}) \leq \mathrm{KL}(Q_e \,\|\, \mu_e) + \log(1/\alpha_e). \tag{128}$$

Apply Theorem 1 and (7). □

Proof of Proposition 3. The cumulative loss and product prior yield the following factorization of (48):

$$q_t(z, w) = q_t^Z(z) q_t^W(w). \tag{129}$$

The first conditional draw therefore has law q_t^Z independently of the preceding w, and the second has law q_t^W independently of the new z. The resulting joint law is q_t, including the common initialization at $t = 1$. Stateless losses then give equal expected cumulative loss. □

Proof of Proposition 4. The best fixed configuration is $(1, 1)$, with zero loss. For $n = t - 1 \geq 1$, the joint normalizer includes its unit unnormalized weight, giving

$$\mathbb{E}f(A_t^{\text{joint}}) \leq \Delta e^{-\eta n \Delta} + 2(\Delta + b)e^{-\eta n(\Delta+b)}. \tag{130}$$

The uniform initialization has expected loss $(3\Delta + 2b)/4$. Summing the geometric series proves the first bound, with its final inequality following from $e^x - 1 \geq x$.

The coordinatewise selector starts at $(0, 0)$ with probability $1/4$. Conditional on remaining there, each coordinate draw at round $t = n+1$ retains zero with probability $(1+e^{-\eta n b})^{-1}$. The event E_T of starting at $(0, 0)$ and retaining zero at every subsequent coordinate draw therefore has probability:

$$\begin{aligned}
\Pr(E_T) &= \frac{1}{4} \prod_{n=1}^{T-1} (1 + e^{-\eta n b})^{-2} \\
&\geq \frac{1}{4} \exp\left(-2 \sum_{n=1}^{T-1} e^{-\eta n b} \right) \\
&\geq \frac{1}{4} \exp\left(-\frac{2}{e^{\eta b} - 1} \right).
\end{aligned} \tag{131}$$

Here $\log(1 + x) \leq x$ gives the first inequality. Loss equals Δ on all T rounds of E_T and is nonnegative otherwise, proving (50). □

Proof of Corollary 1. Writing $n = t - 1$, the three suboptimal unnormalized joint weights are $(4n)^{-\gamma}$ and two copies of $(4n)^{-(\gamma+1)}$. Bound the normalizer below by one and sum the resulting expected losses, using $\sum_{n \geq 1} n^{-p} \leq 1 + 1/(p - 1)$ for $p > 1$. These weights also bound the probabilities of nonoptimal joint draws by a summable sequence, so the first Borel–Cantelli lemma gives eventual optimal selection.

For the coordinatewise selector, the event E_t from the proof of Proposition 4 satisfies the following:

$$\begin{aligned}
\Pr(E_t) &= \frac{1}{4} \prod_{k=1}^{n} \left(1 + \frac{1}{4k} \right)^{-2} \\
&\geq \frac{1}{4} \exp\left(-\frac{1}{2} \sum_{k=1}^{n} \frac{1}{k} \right) \\
&\geq \frac{e^{-1/2}}{4\sqrt{n}}.
\end{aligned} \tag{132}$$

Multiply by Δ and sum over $t = 2, \ldots, T$ to obtain (52).

To establish eventual optimal selection, put $s_n = (4n)^{-(\gamma+1)}$. From any starting pair, the next complete sweep ends at $(1, 1)$ with conditional probability at least

$$\frac{1}{(4n + 1)(1 + s_n)} \geq \frac{1}{10n}. \tag{133}$$

The product of the corresponding avoidance bounds over any infinite tail is zero, so $(1, 1)$ is visited infinitely often almost surely. A sweep starting at $(1, 1)$ departs from it with probability at most $2s_n$. These departure probabilities are summable, so Borel–Cantelli gives only finitely many departures. Infinitely many visits and finitely many departures imply that the coordinatewise selector eventually remains at $(1, 1)$. □

B.3 State Effects and Policy Regret

Proof of Proposition 5. Take text configurations $\mathcal{Z} = \{0, 1\}$ with initial, good, and bad states, zero round-one loss, and

$$z_1 = 0 \Rightarrow s_2 = s_{\text{good}}, \tag{134}$$

$$z_1 = 1 \Rightarrow s_2 = s_{\text{bad}}. \tag{135}$$

The good and bad states are absorbing, with losses for $t \geq 2$ given by

$$\ell_t(s, z) = \begin{cases} 0, & s = s_{\text{good}}, \\ 1, & s = s_{\text{bad}}, \end{cases} \tag{136}$$

independent of z. Choosing $z_1 = 1$ loses $T-1$, as does either fixed action on that bad-state path, giving external regret:

$$\sum_{t=1}^{T} \ell_t(s_t, z_t) - \min_{z\in\{0,1\}} \sum_{t=1}^{T} \ell_t(s_t, z) = (T-1) - (T-1) = 0. \tag{137}$$

The constant counterfactual $z = 0$ enters the good state and loses zero, so

$$R_T^{\mathrm{pol}}(0) = (T-1) - 0 = T-1. \tag{138}$$

This completes the proof. □

Proof of Proposition 6. Fix an optimizer Alg and text-only environments E_ξ, $\xi \in \{0,1\}$, sharing their initial state and pre-round history. Their indistinguishability makes $p = \Pr_{\mathsf{Alg}}(z_1 = 1)$ common to both. Reuse the absorbing states and losses from the proof of Proposition 5, with first transition

$$s_2 = \begin{cases} s_{\mathrm{good}}, & z_1 = \xi, \\ s_{\mathrm{bad}}, & z_1 \neq \xi. \end{cases} \tag{139}$$

By (136), the matching fixed action $z = \xi$ loses zero, and the optimizer's regrets are

$$R_T^{\mathrm{pol}}(0; E_0) = p(T-1). \tag{140}$$

$$R_T^{\mathrm{pol}}(1; E_1) = (1-p)(T-1). \tag{141}$$

Since $\max\{p, 1-p\} \geq 1/2$,

$$\max\left\{R_T^{\mathrm{pol}}(0; E_0), R_T^{\mathrm{pol}}(1; E_1)\right\} \geq \frac{T-1}{2}. \tag{142}$$

The hard environment is chosen after fixing the optimizer but before realizing its random seed. Revealing ξ after the irreversible transition cannot help. □

Proof of Lemma 1. For measurable $D \subseteq \mathcal{A}$, the conditional transition kernel is

$$K_t(a, D) = e^{-\eta d_t(a)} \mathbf{1}_D(a) + \left(1 - e^{-\eta d_t(a)}\right) q_{t+1}(D). \tag{143}$$

Integrating against q_t gives $\zeta_t q_{t+1}(D) + (1-\zeta_t) q_{t+1}(D) = q_{t+1}(D)$, proving the marginals by induction. Use $\mathbb{E}[Y_t] = \int (1 - e^{-\eta d_t})\, \mathrm{d}q_t$, $1 - e^{-x} \leq x$, and $d_t \leq 1$, then sum. □

Proof of Proposition 7. Both actions have law q_t by Lemma 1, hence expected diagonal loss $\int d_t \, \mathrm{d}q_t$. The restriction terms also agree, so subtracting (2) in expectation proves the identity. □

Proof of Lemma 2. A coordinate switch at s can contaminate only $s, \ldots, \min\{T, s + m_e - 1\}$, at most m_e rounds. Simultaneous intervals share a start, so their union has length at most the larger horizon. Deployed and diagonal losses coincide outside these intervals and differ by at most one inside, proving (67) by the triangle inequality and a union bound. Without a refresh, $A_s = A_{s-1}$, so $\chi_s \leq m_{\max} Y_{s-1}$ gives (68). □

Proof of Theorem 2. Lemma 1 and Theorem 1 applied to (61) give

$$\sum_{t=1}^{T} \mathbb{E}[d_t(A_t)] - \int_{\mathcal{A}} \sum_{t=1}^{T} d_t(a)\, \mathrm{d}Q(a) \leq \frac{C_Q}{\eta} + \frac{\eta T}{8}. \tag{144}$$

Adding the signed state term as in (2), then applying (67), (68), and (65), gives the sharper chain

$$\begin{aligned} R_T^{\mathrm{pol}}(Q) &\leq \frac{C_Q}{\eta} + \frac{\eta T}{8} + \mathbb{E}\left[\sum_{s=2}^{T} \chi_s\right] \\ &\leq \frac{C_Q}{\eta} + \frac{\eta T}{8} + m_{\max} \mathbb{E}\left[\sum_{t=1}^{T-1} Y_t\right] \\ &\leq \frac{C_Q}{\eta} + \frac{\eta T}{8} + m_{\max}(T-1)(1 - e^{-\eta}) \\ &\leq \frac{C_Q}{\eta} + \frac{\eta T}{8} + m_{\max}\eta(T-1). \end{aligned} \tag{145}$$

Optimizing gives (12) and (13) for $C_Q > 0$. At $C_Q = 0$, equality in KL gives $Q = \mu$, and a fixed prior draw incurs $\sum_t d_t(A_1)$, matching the comparator in expectation. □

Proof of Corollary 6. The conditioned prior in the proof of Corollary 4 has KL $C_{\delta_Z,\delta_W}(a^\star)$ and diagonal loss at most $T(G_Z\delta_Z + G_W\delta_W)$ above $a^\star$ by Lipschitzness. Apply Theorem 2 and optimize. □

Proof of Proposition 8. Constant parameter history contributes no contamination in Lemma 2, so Theorem 2 applies on $\mathcal{Z}$ with prior π and horizon m_Z. The parameter-only proof is symmetric. □

Proof of Theorem 5. Apply Theorem 1 on $\mathcal{P}$ to losses $J_k/H \in [0,1]$ with learning rate ηH, which gives the weights in (78). Multiplying its regret bound by H proves (79), and optimizing over η gives the stated rate. □

B.4 Feedback Requirements and Extensions

Proof of Corollary 7. Apply Lemma 1 and Theorem 1 with the fixed estimates $\widehat{d}_t$ as losses. Bounded estimates give an expected refresh count at most $\eta(T-1)$, so Lemma 2 bounds the expected absolute state term by $m_{\max}\eta(T-1)$. For every t,

$$\mathbb{E}[d_t(A_t)] - \int d_t \,\mathrm{d}Q \le \mathbb{E}[\widehat{d}_t(A_t)] - \int \widehat{d}_t \,\mathrm{d}Q + 2\epsilon_t. \tag{146}$$

Sum and apply the KL bound to $\widehat{d}_{1:T}$. □

Proof of Corollary 8. Hoeffding's inequality bounds each failure probability by $2\exp(-2n\epsilon_n^2)$, so a union bound over KT pairs gives δ. For any estimate table in the success event, apply Corollary 7 with $\epsilon_t = \epsilon_n$. Otherwise, true policy regret is at most T, since both cumulative losses lie in $[0,T]$. Taking expectations gives (82). □

Proof of Corollary 9. For a fixed exogenous estimate table, Lemma 1 and Theorem 1 apply to $\widehat{d}_t$. Since q_t depends only on $\widehat{d}_{1:t-1}$, conditional unbiasedness gives

$$\mathbb{E}\left[\sum_a (q_t(a) - Q(a))\widehat{d}_t(a)\right] = \mathbb{E}\left[\sum_a (q_t(a) - Q(a))d_t(a)\right]. \tag{147}$$

Taking expectations of the estimated-loss bound controls true diagonal selection by $C_Q/\eta + \eta T/8$. Bounded estimates give at most $\eta(T-1)$ expected refreshes, allowing Lemma 2 to bound the state term. □

Proof of Proposition 9. Take $\mathcal{C} = \{0,1\}$, $q_1 = (1/2, 1/2)$, and $\ell_1 \equiv 0$. Lazy sampling retains $A_2 = A_1$ surely, while $q_2 = (1/2, 1/2)$. After observing A_1, but before round two, let

$$\ell_2(a) = \mathbf{1}\{a = A_1\}. \tag{148}$$

Then $\ell_2(A_2) = 1$, whereas $\langle q_2, \ell_2\rangle = 1/2$. Conditional on the realized past, the actual law is $\mathcal{L}(A_2 \mid A_1) = \delta_{A_1}$, not q_2. □

Proof of Proposition 10. The learner's loss is $T - 1 - S_T$. Since $\ell_t(0) + \ell_t(1) = 1$, the smaller fixed-action cumulative loss on this realized sequence is at most $(T-1)/2$. Subtract and take expectations. □

Proof of Lemma 3. Let $Z_t = \sum_a \mu(a) e^{-\eta \sum_{s<t} x_s(a)}$. For $u \ge 0$, $e^{-u} \le 1 - u + u^2/2$. Therefore

$$\log \frac{Z_{t+1}}{Z_t} \le -\eta\langle q_t, x_t\rangle + \frac{\eta^2}{2}\langle q_t, x_t^2\rangle. \tag{149}$$

Here $1 - u + u^2/2 > 0$, so using $\log(1+v) \le v$ is valid. The Jensen lower bound (127) applies with x_t replacing L_t because that step does not require losses bounded by one. Sum the upper bound, use $Z_1 = 1$, combine with the lower bound, and divide by η. □

Proof of Lemma 4. For the mean, the probability $r_t(a)$ cancels the denominator in (91). For the second moment it cancels only one of its two squared factors. Sum with weights $q_t(a)$ and use $Z_t^2 \le 1$. □

Proof of Lemma 5. Conditioning fixes all estimates and leaves deployment coins independent. The kernel in (143), with $\widehat{d}_t$ replacing d_t, therefore preserves q_{t+1} by the calculation in the proof of Lemma 1. That calculation needs only nonnegative finite losses, which keep retention probabilities in $[0, 1]$. Induction gives the marginals, and

$$\begin{aligned}\mathbb{E}[R_t \mid \text{auxiliary transcript}] &= \sum_a q_t(a)(1 - e^{-\eta \widehat{d}_t(a)}) \\ &\le \eta \langle q_t, \widehat{d}_t \rangle. \end{aligned} \tag{150}$$

Average using Lemma 4: although estimates can exceed one, their conditional means lie in $[0, 1]$. □

Proof of Theorem 3. Apply Lemma 3 to $x_t = \widehat{d}_t$. Since q_t and fixed Q are measurable before the probe, conditional unbiasedness gives the following:

$$\mathbb{E} \sum_t \langle q_t - Q, d_t \rangle \le \frac{C_Q}{\eta} + \frac{\eta}{2} \sum_t \mathbb{E} V_t. \tag{151}$$

Lemma 5 gives $\mathbb{E}[d_t(A_t) \mid \text{auxiliary transcript}] = \langle q_t, d_t \rangle$, identifying expected true diagonal selection regret.

On this response tree, Lemma 2 gives, with $S_T = \sum_{t=2}^T \mathbf{1}\{A_t \ne A_{t-1}\}$,

$$\left| \sum_t F_t(A_{1:t}) - \sum_t d_t(A_t) \right| \le m S_T \le m \sum_{t<T} R_t. \tag{152}$$

Add this state charge once and apply (96). Finally use $V_t \le K$, $\langle q_t, d_t \rangle \le 1$, and optimize $C_Q/\eta + \eta B$. □

Proof of Proposition 11. Conditioning on the autonomous auxiliary process fixes all estimates without conditioning on deployment coins, so Lemma 5 and the pathwise contamination bound still apply. For the selection term, condition only on $\mathcal{G}_{t-1}$ before each probe. Both q_t and fixed Q are measurable there, which justifies the unbiasedness step in the proof of Theorem 3. The second-moment and refresh calculations are unchanged. □

Proof of Corollary 10. For a single probe vector X_t, conditional independence gives

$$\begin{aligned}\mathbb{E}[\langle q_t, \widehat{d}_t^2 \rangle \mid \mathcal{G}_{t-1}] &= (1 - 1/k)\langle q_t, d_t^2 \rangle + (1/k)\mathbb{E}[\langle q_t, X_t^2 \rangle \mid \mathcal{G}_{t-1}] \\ &\le 1 - 1/k + K/k. \end{aligned} \tag{153}$$

The nonnegative unbiased average satisfies Theorem 3 with this second moment. □

Proof of Corollary 11. Condition on the pre-gate context. Multiplying the inclusion probability $pq_t(a)$ by the estimator cancels the denominator, whereas one factor remains for its square. Thus the mean is $d_t(a)$ and the moment is at most K/p. Apply Theorem 3 and optimize. □

Proof of Proposition 12. The second-order argument bounds selection relative to $\bar{d}_t$. The difference from true diagonal selection is

$$\langle q_t - Q, d_t - \bar{d}_t \rangle \le \|q_t - Q\|_1 \epsilon_t \le 2\epsilon_t. \tag{154}$$

The lazy refresh calculation gives $\mathbb{E} R_t \le \eta \mathbb{E}\langle q_t, \bar{d}_t \rangle$, not the same expression with d_t. Add the state charge and use $\bar{d}_t \in [0, 1]$. □

Proof of Lemma 6. Condition on U and the learner's history before the fresh block draw. Then g_j, p_j, Q are fixed, and $B_j \sim p_j$. Hence

$$\mathbb{E}[\widehat{g}_j(a) \mid U, \text{past}] = g_j(a), \tag{155}$$

$$\mathbb{E}[\langle p_j, \widehat{g}_j^2 \rangle \mid U, \text{past}] = \sum_a g_j(a)^2 \le K. \tag{156}$$

Apply Lemma 3, use $\mathbb{E}[g_j(B_j) \mid U, \text{past}] = \langle p_j, g_j \rangle$, and average over the past and U. □

Proof of Theorem 4. Actual loss on complete blocks is exactly $b\sum_j g_j(B_j)$. Adding and subtracting $b\sum_j \langle Q, g_j\rangle$ gives

$$\mathbb{E}\sum_{t=1}^{bJ} F_t(A_{1:t};U) - \mathbb{E}\sum_a Q(a)\sum_{t=1}^{bJ} d_t(a;U) = b\mathbb{E}\sum_{j=1}^{J}[g_j(B_j) - \langle Q, g_j\rangle] + \Gamma_b(Q). \tag{157}$$

Apply Lemma 6. The difference on the remaining r rounds is at most r, since both losses lie in $[0,1]$. □

Proof of Lemma 7. The first block's common empty prefix gives zero discrepancy. In later blocks, $u \geq m+1$ repetitions make both coordinate windows constant, equating the losses in (19). At most $\min\{m, b\}$ earlier differences remain, each bounded by one. Summing their absolute values and averaging over Q and randomness gives the bound. □

Proof of Corollary 12. Combine Theorem 4 with Lemma 7, optimize the first two terms, and use $J \leq T/b$ and $r < b$. □

Deployment-only lower bound. The memory-one subclass

$$F_t(a_{1:t}) = \frac{1}{2}\big(\ell_t(a_t) + \mathbf{1}\{a_t \neq a_{t-1}\}\big), \quad t \geq 2, \tag{158}$$

$$F_1(a_1) = \frac{1}{2}\ell_1(a_1), \tag{159}$$

contains adversarial bandits with switching cost, up to a fixed scaling. The lower bound of Dekel et al. (2014) therefore gives a $\widetilde{\Omega}(T^{2/3})$ obstruction for fixed action count, without establishing optimal m or prior dependence in (111). □

Proof of (113). The first block has the same empty prefix under deployment and the comparator, so its reaction is zero. In the remaining blocks, split the first $\min\{h_0, b\}$ rounds from the remainder in (19). Bound each early absolute discrepancy by one and the remainder by (112), then apply Theorem 4. □

Proof of Proposition 13. The first block still has identical prefixes. In every other block, sum the β_u bounds in (19), then apply Theorem 4. □

Proof of Corollary 13. After $u-1$ common actions, contraction reduces any initial distance at most D to $D\rho^{u-1}$. For $u > m$, coordinate windows agree, bounding loss discrepancy by $\min\{1, GD\rho^{u-1}\}$. For the remaining rounds $u \leq m$, use the bound of one. Apply Proposition 13 and sum the geometric tail, using that each summand is also at most one, to obtain (119). Optimize as in (110) with H replacing m. □

Proof of Lemma 8. The product of propensity ratios implements the intervention fixing every action in the window to a. On that intervention, total finite memory identifies current loss with $d_t(a)$ independently of the earlier prefix. Sequential conditioning gives the mean identity and $\mathbb{E}W_t^a = 1$. Since $0 \leq W_t^a \leq \epsilon^{-(m+1)}$ and $0 \leq Z_t \leq 1$,

$$(W_t^a Z_t)^2 \leq \epsilon^{-(m+1)} W_t^a. \tag{160}$$

Take the same conditional expectation. □

B.5 Adaptation Costs and Amortization

Proof of Proposition 14. Subtract (173) from (174) and divide by the per-use difference, which is positive by (175). The same comparison immediately gives the nonpositive-threshold case. □

C Experimental Details and Numerical Results

This appendix provides the experimental specifications and numerical results supporting Section 6. Summaries and error bars use means ± one SE.

C.1 Replication and Uncertainty Calculations

For independent replication means $x_1, \dots, x_N$ with average $\bar{x}$, we calculate

$$\mathrm{SE}(\bar{x}) = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N(N-1)}}. \tag{161}$$

Paired contrasts are formed within each replication before this calculation. The persistent-state and approximate-feedback studies have ten independent environment tables per loss family, with 64 paired learner repetitions per table and condition. Their uncertainty includes variation across tables after averaging learner repetitions. The joint-versus-coordinatewise study propagates distributions exactly and has no sampling uncertainty.

The controlled restricted-feedback study uses four fixed environments and 64 selector repetitions per environment and condition. SQL restricted-feedback replay uses 256 selector repetitions in each of three training banks, with independent selector streams across banks. For E equally weighted fixed environments or banks, let s_e^2 be the sample variance across R_e selector repetitions in stratum e. The conditional Monte Carlo standard error (MCSE) of the selector mean combines their independent mean variances as

$$\mathrm{MCSE}(\bar{x}) = \frac{1}{E}\sqrt{\sum_{e=1}^{E}\frac{s_e^2}{R_e}}. \tag{162}$$

Dispersion between the fixed environments or banks is not added to this conditional variance.

SQL training-seed summaries use the three separately trained candidate banks as replications on the same fixed benchmark. They assume independent training randomization conditional on the shared construction data, benchmark, and base candidates. The SEs themselves, however, are imprecisely estimated from only three replications. Matched candidate subsets are averaged within a seed before calculating uncertainty because overlapping subsets are not additional replications. For Spider, the six base-model prompt configurations share fixed outcomes across seeds. Interaction SEs do not quantify uncertainty in selected family minima or variation over new databases. Full-diagonal Spider expectations and full-feedback and draw-once replay expectations are exact within each bank, but may vary between training banks.

The original SParC deployment study shares selector randomness across training banks. For selector repetition r, we first average the bank-specific quantity x_{br} over the three banks,

$$x_r = \frac{1}{3}\sum_{b=1}^{3} x_{br}, \tag{163}$$

then compute the SE over the 16 paired repetition averages. Coupling differences are taken within each paired repetition before averaging. This preserves cross-bank dependence rather than treating the banks as independent selector runs. Training-seed uncertainty conditions on the benchmark, while conditional MCSEs additionally condition on the evaluated banks or environments. These SE summaries do not provide confidence levels or significance tests.

C.2 Joint and Coordinatewise Selection

The coordinate-selection study uses the common family $\{0,1\}^2$, a uniform product prior, and exact full-information feedback. The additive control has losses $(0.4, 0.2, 0.2, 0)$ in configuration order $((0,0),(0,1),(1,0),(1,1))$. The barrier condition has losses $(0.4, 0.6, 0.6, 0)$, corresponding to $\Delta = 0.4$ and $b = 0.2$ in Proposition 4. Both tables are constant across rounds and share the base loss and optimal configuration, but their intermediate losses differ. Exact joint EW and the coordinatewise sampler share the prior, loss table, and learning-rate schedule. Both selectors initialize uniformly, and the coordinatewise sampler deploys only after a complete text-then-parameter sweep. The best fixed comparator $(1,1)$ has zero loss, so expected cumulative loss equals regret. Family restriction and state effects are also zero.

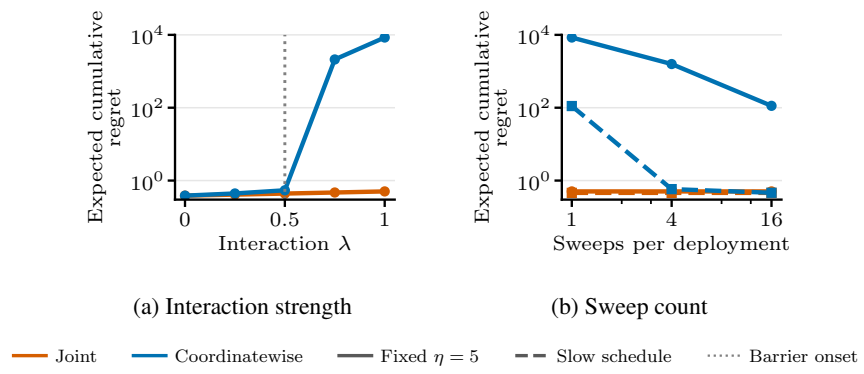


Figure 8: Sensitivity of exact expected regret to interaction strength and coordinatewise sweep count at $T = 100{,}000$, with a shared logarithmic scale. The dotted line in Panel (a) marks the onset of a strict coordinatewise local minimum.

For the four-state sweep kernel K_t and coordinatewise distribution p_t, we compute

$$p_1 = \mu, \tag{164}$$

$$p_t = p_{t-1} K_t^s, \qquad t \geq 2, \tag{165}$$

where s is the number of sweeps before deployment and every sweep in round t uses the same target q_t. Joint loss is computed directly from q_t in (48), and both targets use only preceding loss rows. Expectations are evaluated exactly in floating-point arithmetic, without selector Monte Carlo error. Conditional probabilities are normalized within each coordinate slice, avoiding division by globally underflowed target probabilities. Independent tests enumerate short conditional-draw trajectories and check additive equivalence, target stationarity, and the stated bounds. The 21 distinct conditions propagate 1,522,200 rounds in total. Across all conditions, probability normalization and agreement of the two methods' regrets on additive losses held within floating-point tolerance.

The primary comparisons run through $T = 100{,}000$ using either fixed $\eta = 5$ or the slow schedule of Corollary 1 with the known synthetic barrier $b = 0.2$. The additive control gives equal expected regret under both schedules. For the barrier loss with fixed $\eta = 5$, joint regret is 0.504 at $T = 100{,}000$, whereas one-sweep coordinatewise regret is $8{,}454.705$; see Figure 2. Under the slow schedule, joint regret is 0.460, while coordinatewise regret grows from 35.264 at $T = 10{,}000$ to 111.154 at $T = 100{,}000$. Over those same horizons, the coordinatewise probability of selecting $(1, 1)$ rises from 99.56% to 99.86%. The finite-horizon curves illustrate the separations. The almost-sure eventual fixation and regret lower bounds, however, follow from Section 3, rather than from these curves.

The interaction sweep fixes $T = 100{,}000$, $\eta = 5$, and one sweep per deployment. We retain $f(0, 0) = 0.4$ and $f(1, 1) = 0$ and set

$$f(0, 1) = f(1, 0) = 0.2 + 0.4\lambda, \qquad \lambda \in \{0, 0.25, 0.5, 0.75, 1\}. \tag{166}$$

The table is additive at $\lambda = 0$ and has a strict barrier at $(0, 0)$ only for $\lambda > 0.5$. At $T = 100{,}000$, coordinatewise regret is 0.443 and 0.541 for $\lambda = 0.25$ and 0.5, but rises to $2{,}112.357$ for $\lambda = 0.75$. The tested nonadditive tables therefore do not all exhibit a large regret gap; see Figure 8, Panel (a). We also compare $s \in \{1, 4, 16\}$ sweeps under the fixed and slow schedules; see Figure 8, Panel (b). All intermediate draws remain undeployed, and extra sweeps receive no new feedback. At the same horizon, four and sixteen sweeps reduce fixed-rate coordinatewise regret to $1{,}581.021$ and 112.628, respectively. Under the slow schedule, the corresponding regrets fall to 0.582 and 0.462, close to the joint value 0.460. These finite-horizon reductions do not imply that a fixed number of sweeps eliminates the fixed-rate trapping mechanism. Additional checks use fixed rates $\eta \in \{1, 2, 5\}$ and the horizon-tuned rate $\eta = \sqrt{8 \log(4)/T}$ at $T \in \{100, 1{,}000, 10{,}000, 100{,}000\}$, restarting each horizon-tuned run from the common prior. The fixed-rate coordinatewise regrets at $T = 100{,}000$ are 906.794, $2{,}854.616$, and $8{,}454.705$, respectively. With the horizon-tuned rate, joint and coordinatewise regrets are instead 131.838 and 132.180, a gap of 0.342. However, the fixed-rate linear separation cannot be transferred uniformly to this horizon-dependent schedule. These comparisons use the same number of deployment rounds but do not equalize sampling computation. Their scope

is selection among four fixed configurations, rather than prompt search, parameter training, or arbitrary alternating optimization.

C.3 Selection and Signed State Effects

The persistent-state study uses configurations $\{0,1\} \times \{0,1\}$. Ten independent environment seeds each generate a master table of length $4{,}096$ for each of three loss-table families, supplying prefixes at $T \in \{256, 1{,}024, 4{,}096\}$. The first has identical diagonal losses across configurations. The second has fixed means $(0.62, 0.56, 0.54, 0.46)$ in configuration order $((0,0),(0,1),(1,0),(1,1))$, with bounded exogenous variation and a learning signal that does not shrink with the horizon. The third changes the favored configuration over fixed 128-round phases. All tables and schedules precede learner randomization.

Let I_t indicate whether any relevant coordinate window contains different values, so that it is zero when both histories are constant. A zero-memory coordinate contains only its current value and cannot trigger I_t. The deployed loss follows the synthetic loss law in Section 6, with $\lambda \in \{-0.2, 0, 0.2\}$. Interior diagonal losses keep $f_t \in [0,1]$. Negative, zero, and positive λ make mixed histories beneficial, neutral, and harmful, respectively. This synthetic mechanism makes no claim about switching actual prompts or adapters.

The joint configuration family is evaluated across all loss-table families, horizons, $(m_Z, m_W) \in \{(0,0),(4,0),(0,4)\}$, and state coefficients. A restriction panel evaluates text-only and parameter-only families on the fixed-advantage and changing-quality tables at $T = 1{,}024$ under both positive-memory conditions. Every panel compares independent and lazy EW with draw-once, which retains one initial sample, and with best fixed configurations. Both EW selectors use the uniform prior and the rate from (12),

$$\eta = \sqrt{\frac{\log |\mathcal{C}_e|}{T/8 + m_{\max}(T-1)}}. \tag{167}$$

This uses the family size, declared horizon, and influence bound, without the best candidate's identity.

Across the three loss-table families, each environment seed has 315 algorithm conditions and 64 learner tapes (sequences of random inputs) per condition, totaling 201,600 trajectories over ten seeds. Shared tapes pair algorithm contrasts, which are averaged over tapes before forming the ten independent table means per loss-table family. Each trajectory records excess deployed loss over the common joint comparator, expected diagonal selection regret, family restriction, signed state effect, and the sampling residual in (26).

Draw-once removes learning and switches, while zero memory removes state effects. When there is no selection signal, the coupling alone determines differences. These controls test the accounting interpretation, while theorem-inequality checks verify implementation.

All trajectories satisfied the sampled decomposition to floating-point tolerance (maximum residual 2.8×10^{-11}), and matched EW selectors had exactly equal computed expected diagonal selection. At $T = 4{,}096$, $(m_Z, m_W) = (4,0)$, and $\lambda = 0.2$, both had expected diagonal selection regret 144.797 on the fixed-advantage tables. Their mean state effects were 445.631 and 4.038, respectively, and sampled excess deployed losses were 590.328 ± 0.994 and 148.543 ± 1.675. The sampling residual in (26) reconciles empirical totals with exact expected diagonal selection. Figure 9 illustrates the decomposition on fixed-advantage tables at $T = 1{,}024$.

Reversing the state coefficient reversed the selector ordering: on fixed-advantage tables at $T = 4{,}096$ and $(m_Z, m_W) = (4,0)$, the paired independent-minus-lazy excess-loss contrast was 441.785 ± 1.404 for $\lambda = 0.2$, -441.401 ± 2.264 for $\lambda = -0.2$, and 0.192 ± 1.646 for $\lambda = 0$.

Relative to the common fixed joint-family optimum, mean text-only and parameter-only family restrictions on fixed-advantage tables at $T = 1{,}024$ were 82.039 and 102.701. Under independent selection with harmful state effects, changing influence from $(4,0)$ to $(0,4)$ changed text-versus-parameter mean excess loss from 292.850 versus 127.504 to 112.627 versus 312.816. Candidate losses and expected diagonal selection were unchanged: these conditional reversals reflect the assigned influence mechanism, not intrinsic memory lengths of text or parameter artifacts.

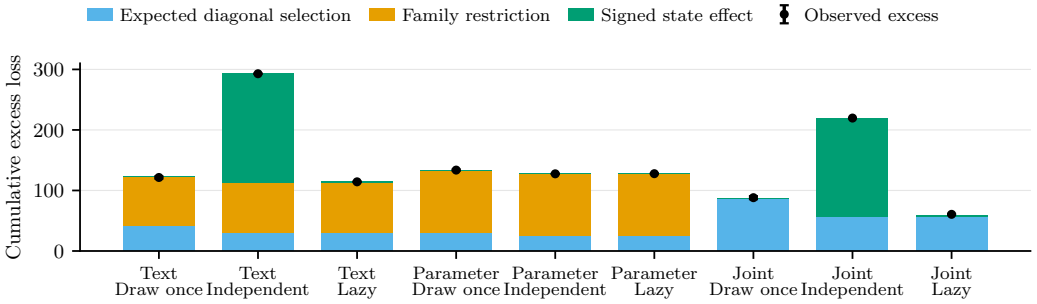


Figure 9: Decomposition of cumulative excess loss on fixed-advantage tables at $T = 1{,}024$ and $\lambda = 0.2$. Black marks show observed totals, including the diagonal sampling residual.

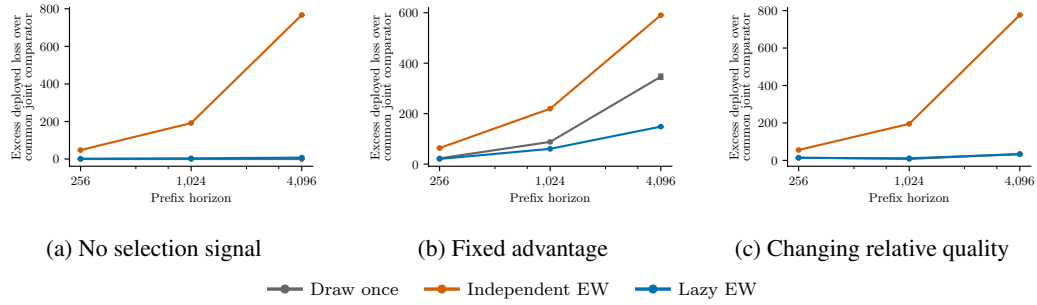


Figure 10: Horizon dependence and the draw-once control under harmful state with $\lambda = 0.2$.

Learning was not uniformly beneficial: without a selection signal in the same harmful-state condition at $T = 4{,}096$, draw-once had zero excess loss, versus 767.456 for independent EW and 7.470 for lazy EW; see Figure 10. For changing quality, lazy EW's 32.835 ± 1.959 and draw-once's 34.386 ± 0.996 differed by less than the SE of the lazy estimate. These mechanisms establish no unconditional advantage for learning, a coupling, or an adaptation mode.

C.4 Feedback Accuracy, Switching, and Evaluation Cost

The approximate-feedback study reuses the fixed-advantage and changing-quality tables from Section C.3, with $T \in \{1{,}024, 4{,}096\}$, $(m_Z, m_W) \in \{(0,0),(4,0)\}$, all three state coefficients, and the same four-configuration joint family.

The eight feedback arms comprise exact losses, a common additive offset of 0.1, two reference-opposing biases of magnitude 0.04 and 0.08, and Monte Carlo estimates with $n \in \{1, 16, 64, 256\}$ samples per candidate-round. The biases raise the best fixed reference's loss and lower its alternatives. These synthetic errors are fixed from the true table before learning and make no assumptions about operational estimators. The offset arm remains in $[0, 1]$ without clipping. Each estimate $\widehat{d}_t(a)$ averages n independent Bernoulli samples of mean $d_t(a)$, independently of learner tapes. We charge $4Tn$ samples per table and compare each table's maximum candidate–round error $|\widehat{d}_t(a) - d_t(a)|$ with Corollary 8's envelope. The Hoeffding envelope uses $\delta = 0.05$ separately for each feedback table and is capped at one. This envelope bounds within-table accuracy separately from empirical error bars and is not a simultaneous confidence statement over the experimental grid.

Each arm uses independent EW, lazy EW, and a centered lazy control with the same prior and rate (167). The centered control diagnoses state effects by applying retention to $\widehat{d}_t(a) - \min_b \widehat{d}_t(b)$, preserving EW marginals while removing sensitivity to action-independent offsets. Across both loss-table families, the grid has 576 algorithm conditions per environment seed and 64 paired learner tapes per condition, totaling 368,640 trajectories over ten seeds.

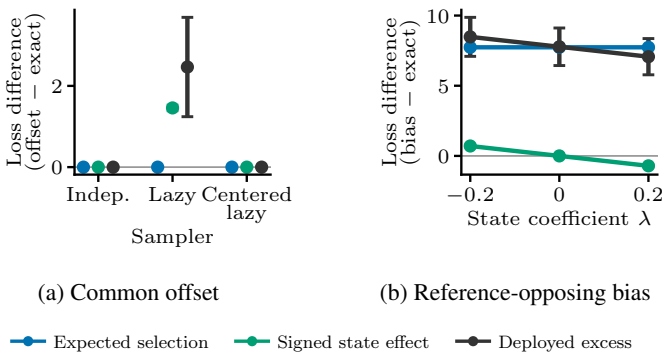


Figure 11: Approximate-minus-exact loss components on changing-quality tables under a common offset in Panel (a) and reference-opposing bias for lazy EW in Panel (b). The offset panel fixes $\lambda = 0.2$, and the panels use different vertical scales.

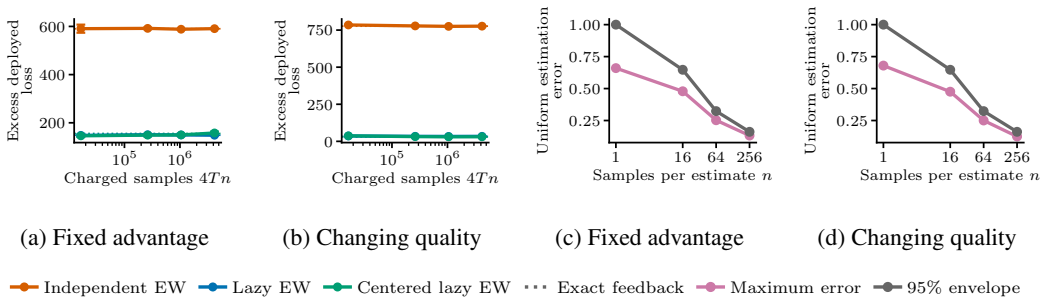


Figure 12: Feedback accuracy and excess deployed loss as the sampling budget grows. Dotted lines mark exact feedback, and the gray curve is the capped Hoeffding envelope. The loss panels use different vertical scales.

Approximate-minus-exact contrasts separate changes in true diagonal selection and signed state, while the shared configuration family fixes restriction. Same-policy comparisons under true and estimated losses check the $2T\epsilon$ bridge in Corollary 7, which does not bound contrasts between differently trained learners. Fresh Monte Carlo estimates satisfy Corollary 9 even at $n = 1$: budget affects accuracy and observed performance, but its coarse expected-regret guarantee holds at every budget. As in Section C.3, contrasts are paired at the table level. The plotted offset and budget comparisons use $T = 4{,}096$, $(m_Z, m_W) = (4, 0)$, and $\lambda = 0.2$. The plotted reference-opposing bias comparison retains this horizon and influence pair, uses lazy EW with bias 0.08, and varies λ over the three state coefficients.

All trajectories passed the decomposition and same-policy bridge checks to floating-point tolerance (maximum accounting residual 2.6×10^{-11}). The common offset preserved EW marginals and the trajectories of independent EW and centered lazy EW under shared tapes, but changed ordinary lazy EW's switches and state effects. At $T = 4{,}096$, $(m_Z, m_W) = (4, 0)$, and $\lambda = 0.2$, ordinary lazy EW's mean state effect increased by 0.779 for fixed advantage and 1.458 for changing quality. The total-loss contrasts were 1.160 ± 0.712 and 2.464 ± 1.223; see Figure 3, Panel (a), and Figure 11, Panel (a). Their means are small relative to table-level variability, despite changed deployed behavior under identical selection distributions.

With reference-opposing bias 0.08 in the displayed fixed-advantage condition, expected true diagonal selection increased by 208.607 and mean harmful state cost by 2.690. For changing quality, selection increased by 7.738 while the state term decreased by 0.707, giving an observed total-loss increase of 7.069 ± 1.291; see Figure 11, Panel (b).

Table 1: Monte Carlo feedback for lazy EW on fixed-advantage tables at $T = 4{,}096$, $(m_Z, m_W) = (4, 0)$, and $\lambda = 0.2$. Max. error is the table mean of the maximum candidate–round error, and observations count represented Bernoulli samples.

Samples n	Observations/table	Max. error	Excess loss $\pm$ SE
1	16,384	0.660	147.835 ± 4.906
16	262,144	0.479	150.527 ± 1.305
64	1,048,576	0.252	149.555 ± 2.290
256	4,194,304	0.132	148.311 ± 1.913

At $\lambda = 0.2$, increasing Bernoulli samples per candidate-round from 1 to 256 improved uniform accuracy without a monotone deployed-loss gain in the displayed conditions; see Table 1 and Figure 12. This is consistent with Corollary 9's common coarse guarantee, without showing that additional samples never help.

For comparison, exact feedback gave excess loss 148.543 ± 1.675. The reported sample counts represent Bernoulli observations rather than measured language-model calls.

A binomial draw divided by n exactly implements the Bernoulli sample average. The study represents 138,035,200 observations, charging each feedback table once despite reuse across learner and state conditions. This known synthetic law tests estimation, without measuring external-simulator runtime or establishing off-policy identification from logged language-model interactions.

C.5 Restricted Feedback on Controlled Losses

We compare the algorithms in Section 5 on the same four-configuration synthetic loss law. The design crosses three diagonal-table families, horizons $T \in \{256, 1{,}024\}$, equal coordinate influence windows $m_Z = m_W \in \{0, 4, 16\}$, and state coefficients $\lambda \in \{-0.2, 0, 0.2\}$. Four independent environment seeds generate length-1,024 master tables, whose prefixes supply both horizons. For each condition, we use 64 selector repetitions per environment. Every selector uses the full joint family and the uniform prior, so restriction is zero relative to the best fixed configuration in that environment. The zero-state control separates learning from history effects because changing λ at fixed influence leaves the diagonal table and prescribed learning rates unchanged.

Full diagonal EW observes all K current diagonal losses after deployment. Sparse diagonal EW, called one-probe lazy EW in the main text, observes one independently and uniformly sampled candidate, using the unbiased importance-weighted vector in (91). Neither learner updates from its deployed loss, and probe randomness is independent of deployment randomness. Both use the lazy coupling, while independent deployment with the same diagonal updates provides a coupling control. Deployment-only EW receives only the actual loss of its selected configuration, holds that configuration for blocks of length $b \in \{1, 2, 4, 8, 16\}$, and updates from the importance-weighted complete-block average in (105). An incomplete tail retains the last complete-block action without another draw or update. Draw-once retains a uniform initial configuration throughout and supplies an exact expected-loss baseline without learning feedback.

Writing $m = \max\{m_Z, m_W\}$, the rates are

$$\eta_{\text{full}} = \sqrt{\frac{\log K}{T/8 + m(T-1)}}, \tag{168}$$

$$\eta_{\text{sparse}} = \sqrt{\frac{\log K}{KT/2 + m(T-1)}}, \tag{169}$$

$$\eta_{\text{block}} = \sqrt{\frac{2 \log K}{K\lfloor T/b \rfloor}}. \tag{170}$$

Comparisons therefore evaluate each feedback regime with its prescribed rate, without holding learning rate fixed while changing information. The $\lambda = 0$ control also reveals how the conservative influence-dependent rates change learning when state itself is neutral.

Table 2: Controlled fixed-advantage excess loss at $T = 1{,}024$ and $m_Z = m_W = 4$, using 64 selector repetitions in each of four fixed environments. Draw-once expectations are exact for each table.

Feedback and Selection	$\boldsymbol{\lambda = -0.2}$	$\boldsymbol{\lambda = 0}$	$\boldsymbol{\lambda = 0.2}$
Full diagonal, lazy	51.574 ± 0.126	56.707 ± 0.000	61.840 ± 0.126
One probe, lazy	58.574 ± 0.495	62.983 ± 0.503	67.392 ± 0.534
Deployment only, $b = 1$	-138.617 ± 0.232	52.135 ± 0.565	232.257 ± 2.160
Deployment only, $b = 16$	41.820 ± 0.436	78.855 ± 0.459	115.524 ± 0.565
Draw once	87.241 ± 0.000	87.241 ± 0.000	87.241 ± 0.000

The evaluation layer records true diagonal losses for accounting, without revealing them to learners that lack diagonal feedback. For full and sparse diagonal algorithms, estimated expected deployed loss adds a sampled signed state effect to $\sum_t \langle q_t, d_t \rangle$, integrating the diagonal term over deployment randomness conditional on autonomous probes. Deployment-only estimates average actual trajectory losses because their update histories depend on deployment. Conditional selector MCSEs follow Appendix C.1 and exclude uncertainty over new environments. This experiment evaluates information access and causal synthetic state. Its measurements do not cover model-generation cost or reactive user behavior.

The fixed-advantage condition exposes the tradeoff between learning and state suppression. With harmful state, deployment-only blocks of length 16 reduce excess relative to $b = 1$, but remain above full-diagonal and sparse lazy EW; see Table 2. The block comparison decomposes into selection plus state as $53.519 + 178.738$ versus $78.924 + 36.600$: less harmful state offsets worse diagonal selection. The ordering reverses under beneficial state, where batching removes the advantage of changing histories over fixed configurations.

Figure 5 adds the intermediate block lengths and influence sweep, using different vertical ranges for harmful and beneficial state.

The independent-coupling diagnostics isolate state at fixed diagonal marginals. At $T = 1{,}024$, $m_Z = m_W = 4$, and $\lambda = 0.2$, sparse independent and sparse lazy EW have the same diagonal loss, but signed state effects 197.545 and 4.409, giving excess losses 260.527 and 67.392. With no diagonal signal, selection regret is zero and excess is entirely a state effect. Under changing quality at the same harmful-state setting, draw-once excess 8.680 is below full lazy 14.303 and sparse lazy 12.904. Even with $\lambda = 0$, the conservative rates can change the ranking: fixed-advantage excess loss at $T = 1{,}024$ and $m_Z = m_W = 16$ is 71.190 for full diagonal EW and 52.050 for deployment-only $b = 1$. Thus the theoretical guarantees do not imply a universal finite-horizon empirical ranking across these loss laws.

C.6 Candidate Construction and SQL Evaluation

The SQL experiments use Qwen3.5-9B (Qwen Team, 2026) on Spider (Yu et al., 2018) and SParC (Yu et al., 2019). Construction, evaluation, and replay protocols are fixed before comparison. Selection constructs no new prompts or adapters. Spider and SParC were jointly partitioned by database, keeping identical schemas together: their official development databases formed the final evaluation set, twenty training-side databases supplied development, and 120 remained for construction. Construction contained 5,876 Spider questions and 7,592 SParC turns, while development contained 1,124 Spider questions and 1,433 SParC turns. Neither construction nor development used final evaluation databases, although checkpoint pretraining exposure to these public benchmarks is unknown. Development scored execution on each original database, whereas final evaluation used released test suites (Zhong et al., 2020), covering the final evaluation databases but not the training databases. These metrics differ: an accidentally correct result on one database can fail elsewhere in its suite.

The six text candidates used different instructions. Candidate z_0 used direct SQL instructions, while z_1 used verification of schema references, joins, filters, aggregation, ordering, and limits. Candidate z_2 used two construction-only SQL demonstrations, z_3 used explicit reference resolution and schema linking, and z_4 used decomposition into SQL clauses. Finally, z_5 combined demonstrations with

Table 3: Spider candidate test-suite accuracy (%). Columns identify the base model and adapter learning rates.

Prompt	**Base**	$\mathbf{5 \times 10^{-5}}$	$\mathbf{10^{-4}}$	$\mathbf{2 \times 10^{-4}}$
Direct z_0	66.05	69.21 ± 0.28	69.34 ± 0.45	69.50 ± 1.01
Verification z_1	65.28	68.73 ± 0.45	69.25 ± 0.19	69.54 ± 0.88
Demonstrations z_2	63.73	69.86 ± 0.70	70.15 ± 0.87	69.21 ± 0.71
Schema linking z_3	65.09	69.54 ± 0.51	69.60 ± 0.35	70.08 ± 1.10
Clause planning z_4	65.86	68.99 ± 0.09	69.60 ± 0.21	70.02 ± 0.81
Demonstrations + verification z_5	64.31	70.08 ± 0.78	70.25 ± 0.86	69.34 ± 0.71

verification. Each was paired with the base weights w_0 or one of three low-rank adaptation (LoRA) adapters (Hu et al., 2022) trained at learning rates 5×10^{-5}, 10^{-4}, and 2×10^{-4}. We repeated this construction for seeds 20260910, 20260911, and 20260912, giving nine trained adapters and a separate 6×4 joint family per seed. The three recipes are candidate choices within a replication, while the training seeds are replications that all share the six base-model prompt configurations. Each adapter used rank 16, scaling 32, query and value projections, one epoch, microbatch size one, and eight-example gradient accumulation. AdamW used $(\beta_1, \beta_2) = (0.9, 0.999)$, $\epsilon = 10^{-8}$, zero weight decay, unit gradient clipping, and fixed unshuffled examples. Completion-only supervision trained 2,228,224 parameters on 11,752 construction examples, equally split between Spider and SParC and assigned cyclically to z_0, z_1, z_2. The other three prompt templates served as evaluation candidates. SParC construction prompts included earlier gold SQL from that dialogue, whereas evaluation excluded target SQL and future utterances. The SParC study reuses each seed's complete candidate family.

Generation used pinned model/tokenizer revisions, bfloat16 weights, greedy decoding without thinking, and limits of 8,192 prompt and 2,048 completion tokens. The local executor required uncapped predictions to match reference results on every test-suite database, preserving SQL literals, `DISTINCT`, duplicate rows, and outer-query ordering. Column-permutation comparison followed the pinned test-suite evaluator, but the executor was not the unmodified leaderboard script. Invalid predictions, 30-second query timeouts, results exceeding 100,000 rows, and completion caps received unit loss, while reference or infrastructure failures stopped evaluation. Whitespace-split comparison operators were normalized outside strings, identifiers, and comments, and SQLite decoding followed upstream malformed-UTF-8 handling. All final evaluation references passed an audit before model evaluation.

C.7 Selection and Family Restriction on Spider

Within each training replication, the text family contains all six prompts with base weights, and the parameter family contains the direct prompt with base weights or any of that seed's three adapters. Their union has nine configurations, and the joint family contains all 24 combinations. All selectors in that replication share the best fixed joint-family comparator on the final evaluation questions, with restriction defined in Section 2. These finite-family comparisons concern unequal candidate availability and do not isolate unrestricted adaptation modes.

We evaluated the full question–configuration matrix, then replayed selectors in frozen question order using only preceding rows. EW used a uniform prior within its family and $\eta = \sqrt{8 \log K/T}$ for family size K and $T = 1{,}034$. Reported loss is the exact expectation $\sum_t \langle q_t, d_t \rangle$, without selector Monte Carlo error. These full-diagonal expectations are also the lazy EW deployment losses: lazy and independent coupling share the same marginals, and Spider has no state effect. Baselines were uniform selection, uniform draw-once, follow-the-leader with uniform tie breaking, and a development-selected candidate held fixed through final evaluation. Uniform and draw-once have equal expected loss here. The SQL feedback replay separately restricts access to this matrix.

The best fixed joint candidates averaged 299.667 ± 7.265 failures across training seeds, whereas the text-family minimum was the shared fixed value 351; see Table 3. Text EW had mean selection regret 8.903, compared with joint EW's 15.358 ± 1.893, but its mean restriction was 51.333 ± 7.265 failures. Averaging within-seed accounting gives text selection plus restriction $8.903 + 51.333 = 60.236$, versus $15.358 + 0 = 15.358$ for the joint family; see Figure 6, Panel (a). Joint EW also

Table 4: Spider expected cumulative failures by family and selector. Uniform also gives the draw-once expectation.

Family	EW	Follow-the-leader	Uniform	Development-fixed
Text	359.903 ± 0.000	358.517 ± 0.000	361.333 ± 0.000	351.000 ± 0.000
Parameter	319.354 ± 5.750	316.222 ± 8.814	325.417 ± 3.495	313.667 ± 9.404
Union	325.435 ± 5.962	319.139 ± 8.585	346.519 ± 1.553	313.667 ± 9.404
Joint	315.024 ± 6.332	310.990 ± 9.123	326.306 ± 3.513	309.333 ± 7.311

Table 5: Spider excess failures after averaging all base-anchored subsets of each size within each seed, relative to the full joint-family comparator.

K	Text	Parameter	Union	Joint
2	56.437 ± 7.265	26.627 ± 5.059	45.258 ± 6.154	31.479 ± 4.488
3	58.223 ± 7.265	21.693 ± 4.498	39.184 ± 5.430	23.931 ± 3.665
4	59.195 ± 7.265	19.687 ± 4.162	34.503 ± 4.988	20.851 ± 3.293

had lower mean expected loss than parameter and union EW, with a paired joint-minus-parameter contrast of -4.329 ± 2.270 failures. Its SE is substantial relative to the mean difference, so the observed ordering should not be treated as a general advantage.

EW was not the lowest-loss rule on these orders: development-fixed and follow-the-leader had lower mean loss in every family; see Table 4.

To control candidate count, we also evaluated every subset of size $K \in \{2, 3, 4\}$ containing (z_0, w_0) within each family. Each subset used the same uniform-prior EW rate for its size and the same full joint-family comparator for its seed. We summarize the complete subset distribution without choosing a subset using final evaluation performance. At $K = 4$, there are 10 text, one parameter, 56 union, and 1,771 joint subsets per seed. Overlapping subsets are not independent replications.

Parameter subsets have lower average excess loss than joint subsets at every tested size; see Table 5. At $K = 4$, the paired parameter-minus-joint contrast is -1.164 ± 1.075, with magnitude comparable to its SE; see Figure 6(b).

Candidate complementarity is a separate question from family size. Relative to direct prompt z_0 and base weights w_0, define the interaction

$$\Delta(z, w) = \frac{1}{T} \sum_{t=1}^{T} \big[d_t(z, w) - d_t(z, w_0) - d_t(z_0, w) + d_t(z_0, w_0) \big]. \tag{171}$$

Negative values mean that joint use improves loss beyond additive effects. Table 6 shows larger negative mean interactions for demonstration prompts than for verification alone. Demonstrations with base weights nevertheless have lower accuracy than direct prompting.

Each interaction is paired within a training bank before averaging across the three seeds. Fourteen of the 15 mean interactions are negative, with clause planning (z_4) at the smallest training rate as the exception. This descriptive pattern does not supply a multiple-comparison significance claim. For the middle-rate comparison, see Figure 7(a). The three Spider matrices each contain 24,816 logical candidate–question evaluations, with 34 completion-cap outcomes receiving unit loss across the three banks. These logical counts include shared base configurations and do not measure a selector's physical deployment cost.

C.8 State Effects on SParC

SParC (Yu et al., 2019) tests history effects under equal per-round configuration distributions. We evaluated 1,203 turns in 422 complete dialogues on the same 20 final evaluation databases, using the complete 24-configuration family for each of the three training seeds. Each seed and history condition uses its own best fixed comparator, evaluated on SParC trajectories rather than Spider questions. Both conditions retained the schema and all raw user utterances through the current turn,

Table 6: Spider prompt–adapter failure interactions in percentage points. Negative values indicate improvement beyond additive effects.

Prompt	$\mathbf{5 \times 10^{-5}}$	$\mathbf{10^{-4}}$	$\mathbf{2 \times 10^{-4}}$
Verification z_1	-0.29 ± 0.17	-0.68 ± 0.30	-0.81 ± 0.17
Demonstrations z_2	-2.97 ± 0.47	-3.13 ± 0.59	-2.03 ± 0.31
Schema linking z_3	-1.29 ± 0.23	-1.23 ± 0.23	-1.55 ± 0.15
Clause planning z_4	0.03 ± 0.23	-0.45 ± 0.31	-0.71 ± 0.21
Demonstrations + verification z_5	-2.61 ± 0.53	-2.64 ± 0.49	-1.58 ± 0.31

Table 7: Recursive SParC accounting over 1,203 turns. Both couplings share exact selection regret, and restriction is zero.

Coupling	**Selection Regret**	**State Effect**	**Deployed Loss**
Independent	26.108	-1.979 ± 1.678	489.795 ± 1.678
Lazy	26.108	0.396 ± 0.175	492.170 ± 0.175

excluding gold SQL and future utterances. Recursive history additionally retained earlier generated SQL, allowing switches to alter inherited context, whereas cleared history removed it.

History reset at dialogue boundaries while EW statistics continued. The longest dialogue had five turns, bounding all recursive influence by four rounds without assuming that deleting an old answer removes its indirect effects. The prompt policy would drop the oldest generated answers at the token limit, but no history was dropped. Model, decoding, token limits, test-suite scoring, and failure rules follow Appendix C.6. Before final model evaluation, we added a missing table alias to one invalid released reference. The frozen protocol recorded this correction, and no final evaluation records were excluded.

Each candidate's complete constant-configuration trajectory started every dialogue with empty generated history and supplied $d_t(a)$ in (1). Independent and lazy EW used preceding rows, the uniform prior, and (12) with $T = 1{,}203$:

$$\eta = \sqrt{\frac{\log 24}{T/8 + 4(T-1)}}. \tag{172}$$

The cleared control retained this rate. Mixed-history losses measured deployment effects without updating EW, preserving matched diagonal marginals. This full diagonal oracle requires own-history losses unavailable from deployment alone. The restricted-feedback replay instead uses only the cleared condition, with its own zero-influence rate.

To estimate signed state effects, we generated independent and lazy sequences in sixteen paired repetitions on the frozen dialogues. Summing each selected turn's mixed-history minus own-trajectory loss and averaging over repetitions estimates signed state effects. Adding this mean to exact expected diagonal loss estimates expected deployment loss without sampled-diagonal noise. For the deployed-loss contrast, paired state sums provide a direct estimate by Proposition 7. Bank averaging and paired conditional MCSEs follow Appendix C.1, conditioning on the banks, benchmark, and model responses.

Under recursive history, the best fixed own-trajectory loss averaged 465.667 across the three banks, and the exact expected diagonal loss averaged 491.774; see Table 7. The paired lazy-minus-independent difference was 2.375 ± 1.626 failures, or 0.197 ± 0.135 percentage points; see Figure 7(b). Here 1.626 is the MCSE of the paired mean, conditional on the evaluated banks, benchmark, and model responses. The estimated difference is about 1.5 MCSEs, providing limited evidence for an expected-loss advantage of independent over lazy EW. The exact expected loss of uniform draw-once averaged 500.500 failures across the banks, compared with estimated losses of 489.795 for independent EW and 492.170 for lazy EW. This draw-once expectation is computed directly over the possible initial configurations on the evaluated banks.

Conditional on the evaluated banks, exact expected switches averaged $1{,}144.669$ under independent coupling and 11.718 under lazy coupling. Within-dialogue expectations averaged 743.637 and

Table 8: SQL restricted-feedback excess failures above each bank's best fixed joint candidate, with separate conditional MCSEs. SParC uses cleared history and the zero-influence rate.

Feedback and Selection	Spider		Cleared SParC	
	Mean ± SE	MCSE	Mean ± SE	MCSE
Full diagonal, lazy	15.358 ± 1.893	0.000	17.674 ± 0.443	0.000
One probe, lazy	24.165 ± 4.312	0.130	34.304 ± 1.979	0.241
Deployment only, $b = 1$	24.340 ± 4.536	0.300	34.182 ± 2.254	0.336
Deployment only, $b = 2$	25.232 ± 4.541	0.327	36.094 ± 2.696	0.342
Deployment only, $b = 4$	25.431 ± 4.694	0.366	38.145 ± 2.836	0.368
Deployment only, $b = 8$	25.382 ± 3.865	0.379	38.908 ± 2.494	0.388
Deployment only, $b = 16$	25.461 ± 4.276	0.378	38.772 ± 3.037	0.415
Draw once	26.639 ± 4.549	0.000	40.514 ± 2.932	0.000

7.637, respectively. Switches at reset boundaries, by contrast, carry no SQL forward. Both couplings produced helpful and harmful changed-history turns. Harmful-minus-helpful counts recover the signed effects for binary loss. Thus fewer mixed histories remove beneficial and harmful effects together, so the large switch reduction does not imply lower deployment loss.

Cleared history supplies a negative control: every deployed turn has its selected diagonal loss. Its best fixed loss averaged 479.667, and both couplings had the same exact expected deployed loss of 508.846, averaged over banks. Clearing changes the diagonal table and its comparator, so differences in total loss across history conditions do not identify signed state effects. The matched-coupling contrast is therefore defined within each condition.

Each seed and condition contains 28,872 logical oracle turns and 38,496 logical selector turns. Because responses are reused, these counts do not represent physical generations. One repeated base-model oracle outcome differs across recursive-history replications: the same generated SQL and token sequence for a `poker_player` turn passed in two banks but incurred a prediction timeout in the third. The original unit-loss timeout outcome is retained, so the bank-averaged result also includes this recorded scoring-runtime exception. Spider and cleared-history SParC base columns agree exactly across all three banks. The evaluated receipts and conditional selector errors do not measure deployment latency or adapter-switching savings.

Appendix F collects the limits of these finite-family, fixed-benchmark studies.

C.9 Restricted Feedback on Precomputed SQL Losses

We specify the feedback rules, learning-rate formulas, and block lengths without using losses from the final evaluation set.

We apply the feedback protocols from the controlled study to the six completed 24-candidate loss matrices: Spider and cleared-history SParC for each training seed. The replay preserves the original questions, candidate outcomes, fixed order, and uniform prior, and compares all regimes with the same best fixed joint candidate within each matrix. Both tasks have zero state effect in this replay, so deployed excess loss equals diagonal selection regret and family restriction is zero. Recursive SParC is excluded: its mixed-history deployment loss cannot be reconstructed from a fixed diagonal matrix.

For each matrix and feedback regime, we run 256 selector repetitions. Full feedback exposes all 24 current diagonal entries only after deployment, while sparse feedback exposes one independent uniform probe. Deployment-only feedback exposes the selected entry and uses $b \in \{1, 2, 4, 8, 16\}$. Draw-once and the independent-coupling diagnostics match the controlled study. The zero-influence setting gives full EW rate $\sqrt{8 \log(24)/T}$, which differs from the rate (172) retained in SParC's recursive-versus-cleared comparison. The two cleared-history experiments therefore test different selector protocols.

Full-feedback and draw-once expected losses are exact. For sparse feedback, errors reflect probe randomness after integrating the deployed diagonal term, whereas deployment-only errors reflect actual trajectories. Conditional MCSEs follow Appendix C.1. Logical diagonal access is $24T$ entries

for full feedback and T for one-probe feedback, while deployment-only learners receive no diagonal oracle and observe T deployed outcomes. All matrices had already been generated.

Full diagonal feedback has the lowest mean excess among these regimes on both tasks; see Table 8. Sparse feedback uses one twenty-fourth as many logical diagonal entries, with paired sparse-minus-full mean differences of 8.807 ± 2.752 expected failures on Spider and 16.630 ± 1.872 on cleared SParC. The increase is larger and has a smaller training-seed SE on cleared SParC. These differences compare the prescribed algorithms and their learning rates, not information alone.

The paired sparse-minus-deployment-only mean differences, with $b = 1$ for deployment-only feedback, are -0.175 ± 0.320 failures on Spider and 0.121 ± 0.440 on cleared SParC. Both mean differences are smaller in magnitude than their training-seed SEs, so their observed ordering does not establish an advantage of either feedback type. Increasing the deployment block length to 16 reduces mean switches from 972.034 to 60.779 on Spider and from $1{,}125.211$ to 71.332 on cleared SParC, while increasing mean excess. The paired $b = 16$ minus $b = 1$ excess-loss contrasts are 1.121 ± 0.315 and 4.590 ± 0.787, respectively, with a larger observed loss increase on cleared SParC. Here switching cannot create beneficial or harmful state, so batching trades update opportunities for fewer configuration changes. Figures 7(c) and 7(d) show the training-seed averages, while Table 8 separately reports conditional selector uncertainty.

D Adaptation Costs and Amortization

After validating admissibility and target satisfaction as specified in Definitions 2 and 3, compare preparation and operating costs under a common evaluator and scalarization. Small regret alone does not establish target satisfaction. A break-even comparison between two modes requires both to be feasible. For $e \in \{Z, W\}$, let κ_e denote one-time preparation and installation costs, and let $c_t(a_{t-1}, a_t)$ denote transition costs, accounted for separately from the loss contamination in (67).

For validated per-use realization losses $\varepsilon_Z, \varepsilon_W \geq 0$ (zero for exact realization), expected scalarized costs over N uses are

$$J_Z(N) = \kappa_Z + N(r_Z + \varepsilon_Z) + I_Z, \tag{173}$$

$$J_W(N) = \kappa_W + N(r_W + \varepsilon_W) + I_W. \tag{174}$$

Here r_e excludes realization loss, and I_e charges interference, irreversibility, validation, or recovery once. For existing candidates, the preparation component of κ_e is a sunk cost. The model excludes nonaffine and path-dependent costs.

Proposition 14 (Feasibility-aware break-even threshold)**.** *Assume both realizations are admissible and satisfy the declared target tolerance, and that*

$$r_W + \varepsilon_W < r_Z + \varepsilon_Z. \tag{175}$$

A parameter-configuration update has no larger total cost than a text-configuration update if and only if

$$N \geq \frac{(\kappa_W - \kappa_Z) + (I_W - I_Z)}{(r_Z + \varepsilon_Z) - (r_W + \varepsilon_W)}. \tag{176}$$

When the right-hand side is nonpositive, the parameter update is weakly preferred for every $N \geq 0$.

The proof is given in Appendix B.5. If $r_W + \varepsilon_W \geq r_Z + \varepsilon_Z$, reuse alone creates no asymptotic parameter advantage. Costs can enter episodic or bounded-influence losses only with the corresponding range and memory assumptions, which exclude arbitrary path dependence.

Illustrative Adaptation Settings. For revised instructions or retrieval, charge preparation per revision and retrieval per use, with N ending at the next preparation or validation event. For target-matched structured output, an adapter with higher fixed cost but lower per-use cost breaks even at (176). Per-use cost includes input processing, generation, and verification. A model, task, or format change can invalidate this reuse horizon. A retrieval collection and a format adapter may jointly satisfy a target that neither restricted candidate family meets, subject to the common resource budget. This finite-family restriction does not establish nonreachability in the full protocol family.

Complexity-Adjusted Cost Certificates. For continued online selection, take validated, target-matched distributions Q_Z, Q_W on the text and parameter slices in Section 2, and let

$$D_Z = \mathrm{KL}(Q_Z \,\|\, \pi), \qquad D_W = \mathrm{KL}(Q_W \,\|\, \rho). \tag{177}$$

Use the slice priors (41)–(42) and comparator-average costs, counting localization once in ε_e. Assume scalarized losses in $[0, 1]$, one mode-level setup charge, and no subsequent installation or transition costs. For a general range of width G, multiply regret by G. The stateless bound (5) gives

$$U_Z(N) = \kappa_Z + N(r_Z + \varepsilon_Z) + I_Z + \sqrt{\frac{N D_Z}{2}}, \tag{178}$$

$$U_W(N) = \kappa_W + N(r_W + \varepsilon_W) + I_W + \sqrt{\frac{N D_W}{2}}. \tag{179}$$

These selection-regret certificates satisfy $U_W(N) \leq U_Z(N)$ exactly when

$$N\big((r_Z + \varepsilon_Z) - (r_W + \varepsilon_W)\big) \geq (\kappa_W - \kappa_Z) + (I_W - I_Z) + \sqrt{\frac{N}{2}}\big(\sqrt{D_W} - \sqrt{D_Z}\big). \tag{180}$$

Equation (180) orders ex-post, comparator-tuned upper bounds, without establishing an ordering of realized costs. Stateless switching adds $C_e^{\mathrm{sw}} = \mathbb{E}\sum_{t=2}^{N} c_t(A_{t-1}, A_t)$ to U_e, including per-configuration installation and separately charging any unpaid initial transition. Corollaries 3 and 4 give description-length and localized versions, with approximation counted in ε_e.

Count resource charges and signed state effects once each. Persistent state additionally requires loss control, such as Theorem 2, because transition charges alone do not justify these stateless certificates.

E Additional Related Work

Automatic Optimization of Discrete Text. Prompt optimization uses token gradients (Shin et al., 2020), generated or reinforcement-learned instructions (Zhou et al., 2023; Deng et al., 2022), critiques (Pryzant et al., 2023), optimizer histories (Yang et al., 2024a), evolution (Guo et al., 2024; Fernando et al., 2024), and textual gradients or pairwise feedback (Yüksekgönül et al., 2025; Lee et al., 2026a). MIPRO searches instructions and demonstrations (Opsahl-Ong et al., 2024), GEPA evolves textual system components (Agrawal et al., 2026), and SPRIG evolves reusable prompts and evaluates transfer (Zhang et al., 2026). Appendix A.2.5 separates candidate discovery from selection.

Compound Systems, Context, and External Memory. Demonstrations and retrieval configure inference (Brown et al., 2020; Min et al., 2022; Rubin et al., 2022; Lewis et al., 2020). Text also governs tools (Yao et al., 2023; Schick et al., 2023), retained feedback and memory (Shinn et al., 2023; Packer et al., 2023), and optimized pipelines or harnesses (Khattab et al., 2024; Lee et al., 2026b). A-Mem revises linked memories (Xu et al., 2025), LightMem separates online use from offline consolidation (Fang et al., 2026), and LongMemEval evaluates sustained retention and updates (Wu et al., 2025). These systems motivate reusable text with explicit state and resource accounting under protocol Ξ.

Prior Complexity, Description Length, and Language-Model Priors. Probably approximately correct (PAC)-Bayesian bounds charge prior KL (McAllester, 1999), while minimum description length connects coding and complexity (Grünwald, 2007). Language-model priors can also certify discrete prompts (Akinwande et al., 2024). Our online bounds use prior complexity without requiring independent and identically distributed samples.

Stateful and Adaptive Environments. Policy regret compares counterfactual trajectories (Arora et al., 2012). Related approaches include learning with memory, online Markov decision processes, reactive experts, and stateful reference policies (Anava et al., 2015; Even-Dar et al., 2009; Neu et al., 2010; de Farias & Megiddo, 2006; Feige et al., 2017). Anava et al. (2015) obtain square-root rates with oblivious memory losses, counterfactual feedback, and low switching, the mechanism underlying our full-diagonal guarantee. Decoupled exploration separates evaluation from deployment (Avner et al., 2012), including under limited switching (Cheng et al., 2023). Deployment-only feedback instead motivates policy-regret mini-batching (Arora et al., 2012).

F Limitations

Our analysis assumes a fixed, known candidate family constructed before evaluation and does not address how to construct or expand that family. A practical extension is to expand the prompt and adapter banks systematically and repeat the comparisons, examining how candidate diversity affects selection performance and family restriction. We leave this extension to future work.

Our language-model experiments focus on Spider and SParC with one base model. Repeating the evaluation with additional models and task orders would help assess how consistently the observed patterns appear. Longer dialogue sequences would also allow closer examination of how inherited state affects later decisions.