OPTIMISTIC PREFERENCE EXPLORATION WITH FINITE REWARD ENSEMBLES

Anonymous authors
Paper under double-blind review

ABSTRACT

Learning systems increasingly choose the feedback and experience used to improve them. In preference learning, a comparison can favor a strong current response or reveal information that helps future decisions. We develop Optimistic Learning with Ensembles (OLE), a framework connecting finite reward sampling, comparison selection and feedback updates. Its central idea is to assess reward ensembles through the learning decisions they induce. Our analysis establishes a query-regret guarantee for optimistic posterior sampling in a bounded linear preference model. An exact construction explains when information acquired by exploration repays its immediate cost and how dependence between reward samples changes this tradeoff. For implementation, we identify the relative score errors that govern the precision of a chosen selection rule. Controlled learning studies examine the interaction between sampling and selection, while neural experiments demonstrate more precise implementation at a fixed reward-evaluation budget. Agentic OLE applies reward-sensitive selection to generated training trajectories, with evidence from question answering, research and interactive tasks. Together, the results connect preference exploration to the broader design of learning systems that choose their own experience.

1 INTRODUCTION

Learning systems increasingly help choose the data used to improve them. A preference learner decides which responses to compare, while a language-model agent can generate many trajectories from which to train. Iterative training makes this selection problem part of the learning algorithm itself (Xiong et al., 2024). Generating more alternatives does not determine which deserve feedback or should influence an update. Feedback and computation remain limited, and the selected experience changes the model that makes the next decision. Selecting useful experience is therefore central to learning from an expanding supply of model-generated alternatives.

Preference exploration makes this problem precise. A learner selects a response to compare with a reference, observes a preference label and updates a reward model. Active exploration has been studied for preference alignment (Mehta et al., 2025), while ActiveUltraFeedback uses active learning to generate preference data (Melikidze et al., 2026). A response that appears strong can offer high reward now. An uncertain challenger can instead supply a label that helps distinguish future alternatives. A finite ensemble of plausible reward models represents these possibilities through several evaluations of the same candidates. The question is *how should limited reward evaluations guide preference queries that account for current reward and future learning?*

Optimistic Learning with Ensembles (OLE) treats reward sampling and comparison selection as **joint design choices within a learning procedure**. Each reward draw evaluates the same candidates, but the way these draws are arranged determines which predictions compete to select a query. Ensembles with identical individual sampling laws can therefore acquire different labels and follow different learning histories. OLE assesses these choices through the reward value of selected comparisons and the information they provide for subsequent learning. Mean and optimistic selection are variants of this framework. Building on the role of joint predictions in decision making (Wen et al., 2021), OLE follows their consequences through the preference-feedback loop. Its contribution is to connect finite reward sampling, reliable selection and the value of acquired information in one account of

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.

preference exploration. The framework distinguishes the learning value of a selection rule from the numerical accuracy of computing it with a finite reward ensemble.

Analyzing this loop requires more than controlling reward-estimation error. The same samples construct candidate scores and choose their maximizer, so accuracy for one fixed candidate does not directly bound the selected response. A small score perturbation can change the maximizing response while changing the maximum score only slightly. Controlling the selected score loss allows the analysis to accommodate such changes. The selected comparison also determines what the learner discovers. Sacrificing reward now is useful only when learning from the feedback improves later choices enough to repay the cost. Finally, a small neural ensemble introduces numerical errors that can alter the intended selection rule. These challenges link the reliability of a query to its learning value and the computation used to choose it.

We analyze an optimistic OLE instance that selects the larger sampled reward gap from two fresh draws, using established maximum-over-sample optimism (Janz et al., 2024, Remark 5). We establish **a Bayesian query-regret guarantee for approximate, dependent reward draws** under bounded linear logistic preferences. The guarantee charges the error of their common posterior approximation and permits arbitrary dependence within each fresh pair. Its proof controls approximation error at the response selected by those same samples and relates posterior uncertainty to information in the comparison actually queried. This separates the accuracy required of individual draws from the freedom to arrange them together. An exact two-round construction then identifies when an additional informative query repays its immediate reward cost. It shows how a pairing that preserves each sampling law can improve both the acquisition and the later use of information.

For implementation, we derive **a sampling criterion based on relative score errors**. Once the reward law and selection rule are fixed, the task is to compute that rule accurately with finite reward evaluations. The leading asymptotic error depends on the variance of errors in candidate-score differences, weighted by the selection probabilities. Common shifts in all scores cancel from the rule. This identifies which sampling errors matter for a mean-plus-dispersion selector and motivates balanced neural sampling. The accompanying analysis specifies which mask contributions this construction cancels or amplifies, linking the target selection rule to its numerical implementation.

The experiments test these ideas in progressively broader settings. Controlled preference learning reveals that the same pairing can improve optimistic selection while worsening mean selection. A retained-ensemble study also finds lower regret for its optimistic variant than for its matched mean variant. Neural reward experiments show more precise implementation of a fixed selection rule at matched projection work. Agentic OLE uses realized trajectory rewards to select policy-training experience. QA comparisons show higher mean F1 than the strongest listed baseline at each of three model sizes, while rejection configurations assess the role of selecting training experience. Research and interactive tasks extend the application evidence. The resulting progression connects a precise preference-exploration analysis to practical systems that learn from selected experience.

2 OLE FOR PREFERENCE EXPLORATION

OLE specifies how a finite reward ensemble chooses the next comparison and learns from its label. Sampling and selection are joint design choices: their combination determines which feedback enters the learning history. The procedure draws reward functions, selects a query from their predictions and updates reward inference (Figure 1).

At round t, prompt x_t has candidate responses $\mathcal{Y}_t$ and a reference y_t^0 chosen before the current samples. Write $\theta_t^{(1:N)}$ for N sampled reward parameters and $g_{\theta_t^{(i)}}(x_t, y)$ for sample i's reward gap between candidate y and the reference. Each sampled reward function scores the same candidate set. Their joint distribution is called a coupling. Specifying each sample's marginal law leaves this dependence open. The reference may depend on past feedback.

The empirical mean and population standard deviation, with denominator N, are

$$\widehat{\mu}_t(y) = \frac{1}{N}\sum_{i=1}^{N} g_{\theta_t^{(i)}}(x_t, y), \qquad \widehat{s}_t(y)^2 = \frac{1}{N}\sum_{i=1}^{N}(g_{\theta_t^{(i)}}(x_t, y) - \widehat{\mu}_t(y))^2. \tag{1}$$

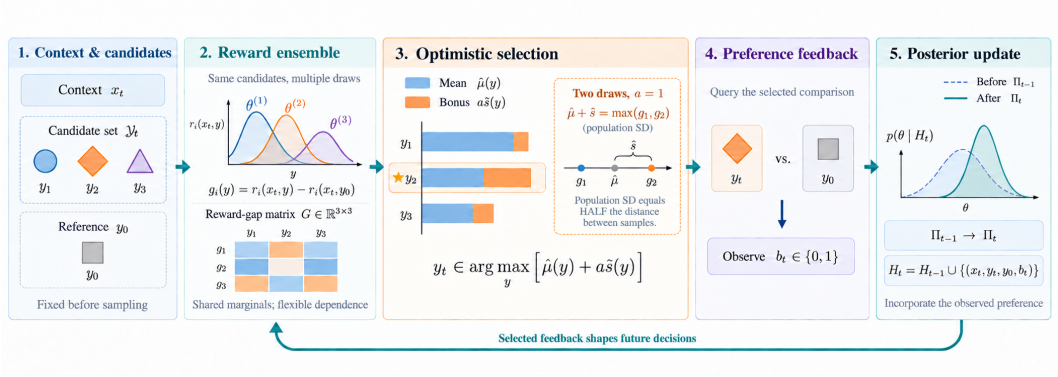

Figure 1: **OLE learns from the comparisons its reward samples select.** Sampled reward functions score the same candidate responses against a reference. The sampling arrangement and selection rule jointly determine which comparison is queried. Its preference label updates the reward model before the next selection. Candidates and the reference are set before sampling in each round.

For $a_t \geq 0$, the learner selects

$$y_t \in \arg\max_{y \in \mathcal{Y}_t} \{\widehat{\mu}_t(y) + a_t \widehat{s}_t(y)\}, \tag{2}$$

with parameter-independent tie breaking. Setting $a_t = 0$ gives mean selection, while a positive coefficient favors strong plausible reward evaluations. Both are OLE variants. The analyzed optimistic instance uses $N = 2$ and $a_t = 1$. Its matched mean variant, denoted Mean2, uses the same pair with $a_t = 0$ and remains stochastic and adaptive. The learner queries whether y_t is preferred to y_t^0 and updates the reward posterior with this comparison and its label before the next draw.

2.1 Statistical model and objective

At round t, $\mathcal{G}_t$ contains the history H_{t-1}, prompt, shortlist of at most K responses and reference y_t^0, also called the anchor. These inputs are fixed before sampling and have no additional conditional dependence on the unknown reward parameter beyond the history.

Assumption 2.1 (Bounded linear preference model)**.** Draw $\theta^\star$ once from a log-concave prior density π_0 on a compact convex set $\Theta \subseteq B_2(B) \subset \mathbb{R}^d$ with nonempty interior. For every feasible comparison and every $\theta \in \Theta$, rewards lie in $[0, 1]$ and the anchor-centered gap satisfies

$$g_\theta(x_t, y) = r_\theta(x_t, y) - r_\theta(x_t, y_t^0) = \psi(x_t, y)^\top \theta, \qquad \|\psi(x_t, y)\|_2 \leq S. \tag{3}$$

The only parameter-dependent feedback kernel is the Bradley–Terry model

$$b_t \mid \theta^\star, \mathcal{G}_t, y_t \sim \mathrm{Ber}(\sigma(g_{\theta^\star}(x_t, y_t))), \qquad \sigma(u) = (1 + e^{-u})^{-1}. \tag{4}$$

Uniform and truncated Gaussian priors satisfy the assumption. The cumulative query regret is

$$R_T = \sum_{t=1}^{T} \left[\max_{y \in \mathcal{Y}_t} g_{\theta^\star}(x_t, y) - g_{\theta^\star}(x_t, y_t)\right], \qquad 0 \leq R_T \leq T.$$

The anchor cancels from the regret comparator but determines the feedback. This is a contextual logistic bandit with a predictable reference. Query regret measures the quality of queried responses. Prediction and subsequent policy performance are separate outcomes.

For $\ell_s(\theta) = \log(1 + e^{g_\theta(x_s, y_s)}) - b_s g_\theta(x_s, y_s)$, the posterior after the completed comparisons is

$$\Pi_{t-1}(d\theta) \propto \pi_0(\theta) \exp\left(-\sum_{s<t} \ell_s(\theta)\right) d\theta, \tag{5}$$

Let $\mu_t(y), s_t(y)$ denote the target posterior gap mean and SD.

3 LEARNING GUARANTEES AND INFORMATION VALUE

Reliable preference exploration requires controlling the value of the selected response and understanding what its feedback contributes. Under the linear logistic model, optimistic OLE with two fresh draws admits a Bayesian query-regret bound despite dependence and posterior approximation. An exact two-round construction then shows when acquired information repays the immediate cost of exploration.

3.1 SELECTING WITH TWO REWARD SAMPLES

The optimistic instance uses two fresh complete reward-function draws with marginal law ν_t, approximating the current posterior. For $N = 2$ and $a_t = 1$, its score is

$$\widehat{\mu}_t(y) + \widehat{s}_t(y) = \max\{g_{\theta_t^{(1)}}(x_t, y), g_{\theta_t^{(2)}}(x_t, y)\}. \tag{6}$$

This identity uses population SD. With sample SD, the equivalent coefficient is $1/\sqrt{2}$. It connects Algorithm 1 to established maximum-over-sample selection (Janz et al., 2024, Remark 5).

Algorithm 1 OLE's optimistic instance with two fresh reward samples

Require: Approximate posterior law ν_t, joint sampling law with both marginals ν_t.
1: **for** $t = 1, \ldots, T$ **do**
2: Observe prompt x_t, candidates $\mathcal{Y}_t$ and reference response y_t^0.
3: Draw a fresh pair $(\theta_t^{(1)}, \theta_t^{(2)})$ with marginals ν_t conditional on $\mathcal{G}_t$.
4: Select $y_t \in \arg\max_y \max_{i=1,2} g_{\theta_t^{(i)}}(x_t, y)$.
5: Observe preference b_t and append (x_t, y_t, y_t^0, b_t) to the history.
6: Update reward inference before the next selection.
7: **end for**

Different joint arrangements of the same reward marginals can therefore select different queries. The argument draws on posterior-sampling analysis (Russo & Van Roy, 2014) and information-based regret bounds (Russo & Van Roy, 2016). We combine these tools with uniform score-error control for the selected response.

For $S > 0$, put $\Gamma_T = d\log(1 + 2BS\sqrt{T}) + 1/8$, $\lambda = \sigma(1)(1 - \sigma(1))/2$, $L_K = 1 + \log(2K)$, and $C_K^2 = \min\{2d, 2K, L_K^2 + 2L_K + 2\}$. When $S = 0$, all regrets vanish.

Theorem 3.1 (Optimistic OLE with two fresh approximate posterior draws)**.** *Under Assumption 2.1, use $N = 2$, $a_t = 1$ and the hard maximizer in* (2)*. Conditional on $\mathcal{G}_t$, generate a fresh pair independently of $\theta^\star$, with both marginal laws equal to ν_t on Θ. Dependence within the pair is unrestricted. With $e_t = W_2(\nu_t, \Pi_{t-1})$,*

$$\mathbb{E}R_T \le \min\left\{T,\, C_K\sqrt{T\Gamma_T/\lambda} + 2\sqrt{2}S\sum_{t=1}^{T}\mathbb{E}e_t\right\}. \tag{7}$$

For an arbitrary prior on Θ, the same result holds with $C_K^2 = \min\{2d, 2K\}$. All expectations are Bayesian.

Here W_2 is the smallest root-mean-square parameter discrepancy over pairings of the two laws. Under log-concavity, at fixed geometry and fixed or polynomial-sized shortlists, the first term is $\widetilde{O}(\sqrt{dT})$. The second charges posterior approximation. The bound does not require accurate two-sample moments. Identical exact-posterior draws recover Thompson sampling (Russo & Van Roy, 2014).

The proof transports each actual draw to a posterior marginal. The expected uniform score change is at most $\sqrt{2}Se_t$, and maximization pays this change twice even if the selected response changes. For round regret r_t, posterior optimism and a second-moment bound on standardized gaps give $\mathbb{E}_t r_t \le C_K\sqrt{\mathbb{E}_t s_t(y_t)^2} + 2\sqrt{2}Se_t$. The acquired labels satisfy $\lambda\mathbb{E}\sum_t s_t(y_t)^2 \le I(\theta^\star; H_T) \le \Gamma_T$ by Lemma B.1. Summing and applying Cauchy–Schwarz proves the theorem. Appendix B.1 proves the information lemma, and Appendix B.2 completes the regret argument.

3.2 When information repays exploration

Consider a learner that makes two decisions. In the first round, a safe response has higher average reward, while an uncertain response provides feedback that helps distinguish the next round's alternatives. Querying the uncertain response sacrifices reward now but can improve the later choice. The total effect is the extra query probability multiplied by the difference between the immediate cost and the later benefit.

An exact instance has $\theta = (c, z)$ uniform on $[1/2, 3/2] \times [-1, 1]$ and reference reward $1/2$. The first round offers a safe gap δc and an informative gap $z/2$. After one preference label, the second round offers gaps $z/2$ and $-z/2$. For $0 < \delta < 1/8$, optimistic selection with independent reward draws chooses the informative query more often than mean selection. Writing their query probabilities as $q_{\mathrm{opt}} > q_{\mathrm{mean}}$, Proposition B.3 gives

$$\mathbb{E}R_2^{\mathrm{opt}} - \mathbb{E}R_2^{\mathrm{mean}} = (q_{\mathrm{opt}} - q_{\mathrm{mean}})(\delta - m^2), \qquad m = \int_0^1 z \tanh(z/4)\, dz. \tag{8}$$

Here δ is the immediate expected cost of an additional informative query and m^2 is its expected benefit for the next decision. Optimism helps when the benefit exceeds the cost and hurts when the cost exceeds the benefit.

Pairing opposite posterior quantiles preserves each reward function's sampling law and improves the use of acquired information in this example. The mean variant never probes, while the optimistic variant probes with probability $1 - 5\delta/2$. Its later benefit per probe rises from m^2 to $m/2$, so the corresponding regret difference becomes $(1-5\delta/2)(\delta-m/2)$. A larger immediate cost can therefore be repaid. The paired optimistic rule also has lower total regret than the independent optimistic rule throughout the stated range (Corollary B.5). Appendix Figure 3 displays these comparisons. The exact pairing differs from the rejection sampler used in our controlled experiment.

4 Sampling design and implementation

Sampling design affects both which feedback is acquired and how accurately a specified selection rule is computed. Coupled posterior draws can change the learning history while preserving each draw's law. For a fixed rule, the leading asymptotic approximation error depends on relative score errors between candidates. These two roles guide the acquisition sampler and neural implementation developed below.

4.1 Sampling for preference acquisition

The controlled study pairs rejection proposals using symmetric innovations $z, -z$ and acceptance uniforms $u, 1-u$. Each stream stops at its own first acceptance. In exact arithmetic, this preserves both posterior marginals without requiring a symmetric posterior, satisfying Theorem 3.1. Appendix D.2 gives the sampler, and Appendix B.7 gives a projected-Langevin alternative and larger-ensemble extension under additional assumptions.

4.2 Approximating a fixed selection rule

Fix a bounded reward-vector law on a finite candidate set. For population reward means μ_y, standard deviations s_y, $a \geq 0$, temperature $\beta > 0$ and a full-support base distribution p, the target rule is

$$q(y) = \frac{p(y)\exp((\mu_y + as_y)/\beta)}{\sum_z p(z)\exp((\mu_z + as_z)/\beta)}. \tag{9}$$

Writing $v_y = \mu_y + as_y$, this rule maximizes, over candidate distributions π, the regularized score $J_v(\pi) = \sum_y \pi(y)v_y - \beta\,\mathrm{KL}(\pi\|p)$. Replacing q by an approximation $\widehat{q}$ loses exactly $\beta\,\mathrm{KL}(\widehat{q}\|q)$ in this objective (Appendix C.1). This gives numerical selection error a direct decision-objective interpretation.

Empirical means and SDs replace the population moments to give $\widehat{q}_N$, using variance denominator N or $N-1$. Write $Z_y = g_y - \mu_y$ and define the centered score influence

$$\xi_y = Z_y + \frac{a}{2s_y}(Z_y^2 - s_y^2), \qquad \bar{\xi}_y = \frac{1}{b}\sum_{j=1}^{b} \xi_y^{(j)}.$$

Set $\xi_y = 0$ when $s_y = 0$, since that candidate's reward is constant.

Proposition 4.1 (Selection error under dependent sampling)**.** *Pool m independent, identically distributed blocks of fixed size $b \geq 2$, each with joint law γ and the prescribed reward law for every member. For $N = bm$, as $m \to \infty$,*

$$\mathbb{E}\,\mathrm{KL}(q\|\widehat{q}_N) = \frac{\mathcal{A}_\gamma}{N} + O(N^{-3/2}), \qquad \mathcal{A}_\gamma = \frac{b}{4\beta^2}\sum_{y,z} q(y)q(z)\,\mathrm{Var}_\gamma(\bar{\xi}_y - \bar{\xi}_z). \tag{10}$$

The reward law, candidate set, score, temperature and block law remain fixed.

The same leading term holds with the KL direction reversed. Appendix C.1 proves the expansion, whose remainder is not uniform as positive standard deviations or temperature vanish. Common score errors cancel from the softmax. The leading selection error therefore depends on score differences, weighted by the target selection probabilities.

Balanced dropout makes each mask coordinate drop in exactly one of b members, chosen uniformly and independently across coordinates. This preserves each member's product-dropout law with drop probability $1/b$. It cancels first-order mask contributions to the leading score influence and amplifies second-order contributions by $b/(b-1)$ (Theorem C.2). The neural study assesses this tradeoff through finite-reference selection error at matched projection work. Its independent streams across candidates preserve candidate-specific dropout laws.

5 Experiments

The experiments examine whether OLE's selection principle improves learning and its numerical implementation. Controlled studies show that optimism can lower regret, with gains depending on the query setting and sampling design. Four neural reward fits show lower approximation error at matched work. Agentic OLE improves mean QA F1 over the listed baselines, and rejection ablations identify a contribution from selecting training experience (Table 1 and Figure 2).

5.1 Learning from selected comparisons

We cross mean and optimistic selection with independent (IID) or coupled posterior draws. Candidate programs from PPE (Frick et al., 2024) are encoded by frozen Qwen2.5 models at 1.5B, 3B and 7B. A two-dimensional reward head learns from synthetic logistic feedback. A development-selected informative-challenger rule compares a high-mean candidate with a slightly lower-mean candidate in a less certain direction. Each encoder/orientation cell uses 512 paired episodes of 512 queries. We weight cells equally, using disjoint evaluation tasks and conditioning on the selected feature geometry. Coupled rejection sampling preserves each posterior marginal (Section 4). Appendix D.2 gives the protocol.

Figure 2(a) shows that **the effect of coupling depends on the selection rule**. Relative to independent sampling, coupling lowers optimistic regret by 0.79 and raises mean regret by 1.44. Within coupled sampling, optimistic selection achieves **12.1% lower query regret**, reducing it from 19.82 to 17.43. Its held-out log loss also improves, by 0.04% of total loss. Both primary contrasts pass the 1% familywise level after Bonferroni correction over four tests, while the IID regret contrast is inconclusive. The learning dynamics show the cost before the gain: the coupled optimistic-minus-mean contrast is positive at 64 queries and negative by 256 queries (Appendix Figure 4). These outcomes support evaluating a sampling design through the queries and updates produced by its selection rule.

Table 1A compares the variants with established acquisition rules on query regret, predictive excess loss and the utility shortfall of a later decision. Coupled optimistic OLE lowers regret by 2.78 relative

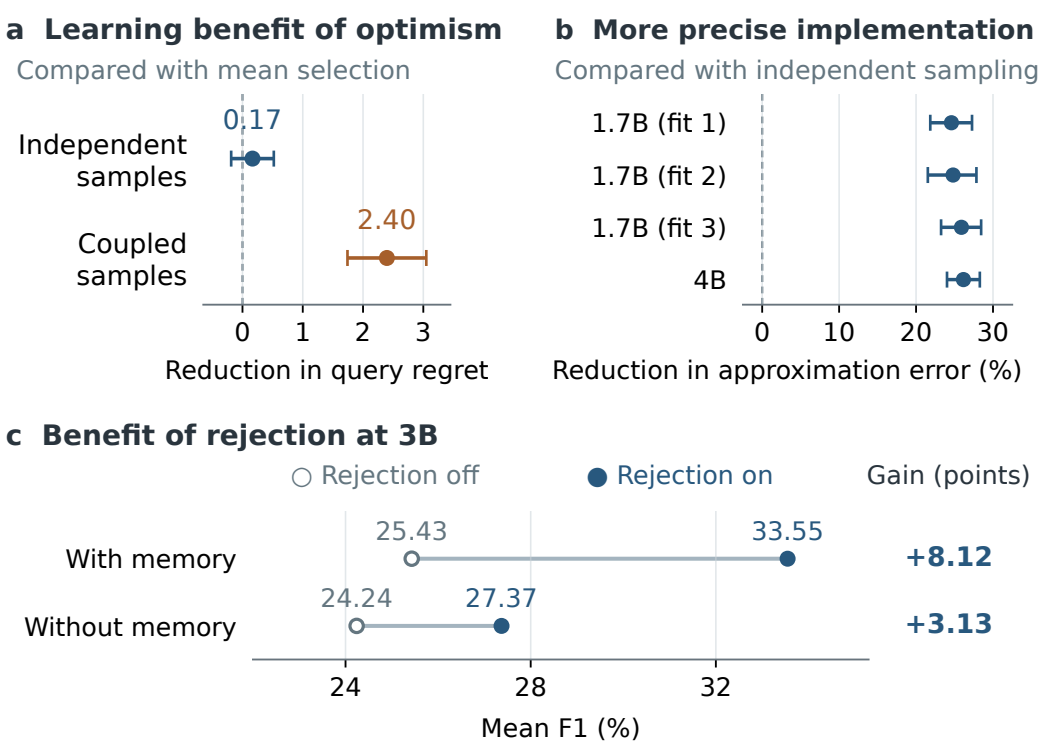

Figure 2: **Learning, implementation precision and trajectory selection.** (a) Regret reduction from optimistic versus mean selection, with descriptive paired 95% intervals. Each encoder/orientation cell uses 512 episodes. (b) Forward-KL error reduction versus independent sampling at matched projection work. Intervals are 98.33% for the three 1.7B fits and 95% for 4B, conditional on fitted models, the evaluation cohort and the finite reference bank (Appendix E.1). (c) Mean F1 (%) with rejection on and off at 3B, averaged equally over seven QA datasets. These descriptive comparisons assess reward aggregation and retention jointly (Appendix G.1).

to exact Thompson sampling (Russo & Van Roy, 2014). Local Fisher-UCB (Faury et al., 2020) achieves lower query regret, while OLE has lower aggregate utility shortfall. Their predictive-loss contrast is inconclusive. A separate implementation retains 20 reward particles across updates and uses mean plus 0.3 times the sample standard deviation. Across 1,536 environments, its regret is 228.89, compared with 247.16 for matched mean selection and 214.12 for Fisher-UCB. This is a **7.39% regret reduction**, or 18.27 cumulative-regret units (95% confidence interval: 10.49 to 26.05), averaging two repetitions within each environment (Appendix F.1).

The preference setting matters. On OpenAssistant human ranks, both pair laws increase query mistakes relative to mean selection and fail the joint query-quality and prediction criterion. Correctly specified synthetic feedback on the same geometry also favors mean selection in regret. A prospective geometry–horizon study does not establish joint regret and prediction gains (Appendices D.3 and D.4). The positive result therefore establishes an exploration benefit within the informative-challenger setting.

5.2 Implementing selection with neural reward models

We fix the fitted reward law and selection rule in (9), isolating *implementation precision* from learning. Four Qwen3 reward fits at 1.7B and 4B use 144 held-out RewardBench 2 groups and 32 paired sampling repetitions. Both sampling laws use 20 reward evaluations. Balanced dropout arranges them in two ten-member blocks while preserving each mask's Bernoulli marginals. Scores use mean plus twice the sample standard deviation and a temperature-one softmax. An independent 2,048-draw bank defines the reference selector. Candidate-specific streams make this a fixed neural-law study, distinct from the complete posterior reward-function draws of the regret theorem.

Table 1: **Learning, implementation precision and agentic performance.** A: preference learning with frozen encoders and a learned 2D reward head. B: precision using 20 reward evaluations per candidate. C: mean F1 (%) over seven QA datasets, with equal task weights and method-specific settings. Bold marks the best point estimates.

A. Learning from selected comparisons (Section 5.1)

Method and sampling	Query regret ↓ ($T = 512$)	Excess prediction loss (10^{-4}) ↓	Utility shortfall (10^{-2}) ↓
OLE (mean), independent	18.39	6.25	1.38
OLE (optimistic), independent	18.22	5.16	1.08
OLE (mean), coupled	19.82	7.48	1.79
OLE (optimistic), coupled	17.43	**4.90**	**1.02**
Exact TS (Russo & Van Roy, 2014)	20.21	6.07	1.32
Local Fisher-UCB (Faury et al., 2020)	**14.30**	5.05	1.09

B. Precision at matched reward-evaluation work (Section 5.2)

Reward model	Independent sampling error ↓	Balanced sampling error ↓	Error reduction (%) ↑
Qwen3-1.7B (fit 1)	8.35	**6.29**	24.60
Qwen3-1.7B (fit 2)	7.37	**5.54**	24.81
Qwen3-1.7B (fit 3)	7.34	**5.44**	25.90
Qwen3-4B	4.71	**3.48**	26.16

C. Agentic question answering (Section 5.3)

Method	Qwen2.5-1.5B	Qwen2.5-3B	Qwen2.5-7B
Base (Qwen et al., 2025)	15.70	18.33	23.34
CoT (Wei et al., 2022)	15.00	19.61	27.71
FS-RAG†	20.94	19.30	19.84
FL-RAG†	23.21	20.22	23.61
ReAct (Yao et al., 2023)	17.15	26.00	31.02
IRCoT (Trivedi et al., 2023)	21.59	23.13	24.12
TC-RAG (Jiang et al., 2025)	17.94	22.27	30.50
ReSearch (Chen et al., 2025)	24.38	26.33	27.28
Search-R1 (Jin et al., 2025)	16.67	30.16	34.08
AEPO (Dong et al., 2025a)	13.19	24.62	12.50
ARPO (Dong et al., 2025b)	13.76	29.60	24.53
Mem1 (Zhou et al., 2025)	13.65	15.02	28.21
OLE	**27.65**	**33.55**	**36.60**

A: means weight encoders and prior orientations equally. Prediction loss subtracts teacher entropy. Utility shortfall measures the true-utility gap of the final posterior-mean choice. B: errors are forward KL in units of 10^{-4}. Figure 2 gives paired intervals for regret reductions in A and relative error reductions in B. Fit details are in Appendix E.1. C: comparisons are descriptive. † FS-RAG and FL-RAG are sentence-based and fixed-token retrieval controls. Appendix G.1 gives complete task scores and settings.

Balanced sampling reduces forward-KL error by **about 25%** across all four fits (Table 1B and Figure 2(b)). At 4B, error falls from 4.71×10^{-4} to 3.48×10^{-4}, and top-response agreement improves by 0.79 percentage points. Each sampling law projects 290.16 million token–member rows. Stochastic projections take 1.78 GPU-hours with independent sampling and 1.77 with balanced sampling. This finite-budget construction improves numerical precision at matched work, without optimizing the asymptotic coefficient in Proposition 4.1. Lower error persists against two independent reference banks (Appendix E.1). Additional RewardBench 2 quality contrasts span zero (Appendix F.1), so the measured precision gain does not establish improved downstream quality.

5.3 Agentic learning from selected trajectories

Agentic OLE uses reward-sensitive selection to retain generated trajectories containing reasoning, retrieval and memory actions. Answer-quality and intermediate-action rewards guide rejection before a group relative policy optimization (GRPO) update (Shao et al., 2024). Reward variation here comes from realized trajectories, and selection determines which experience trains the policy. Qwen2.5

models at 1.5B, 3B and 7B (Qwen et al., 2025) train on curated HotpotQA and 2Wiki questions and are evaluated on seven QA datasets, five beyond the training sources. Appendix G.1 specifies the selection implementations and method-specific settings.

OLE improves on the strongest listed baseline by **3.27, 3.39 and 2.52 F1 points** at 1.5B, 3B and 7B, respectively (Table 1C). It has the highest point estimate on six, seven and six tasks at the respective model sizes. The exceptions are Bamboogle at 1.5B and Natural Questions at 7B. At 3B, enabling rejection raises mean F1 by **8.12 points with memory and 3.13 without memory**, with gains on seven and five tasks, respectively (Figure 2(c)). This comparison evaluates reward aggregation and retention jointly. The ablation does not identify whether selection acted on trajectories or prompt groups. It also does not isolate dispersion from mean ranking and retained volume. Per-seed uncertainty and matched-resource comparisons are unavailable.

Research and interactive tasks broaden the application evidence. On DeepResearch Bench, OLE reaches 43.60 in report quality, 1.39 points above the strongest listed baseline. ALFWorld success increases from 36.5% to 51.1%, with improvements in four of six task categories. WebShop score increases from 14.7 to 27.8, while success reaches 2.3%. These results evaluate the complete agentic system. Appendix G.1.4 gives all application results, including citation measures, task-category comparisons and medical QA.

6 Related work

OLE connects optimistic sampling, active preference learning and reward ensembles through the feedback selected for learning. Its contribution combines regret control for dependent approximate draws, an exact two-round analysis of information value and a criterion for numerical selection error.

Optimistic Bayesian sampling is established in contextual bandits (May et al., 2012). Maximum-over-sample rules appear in linear-bandit ensembles (Janz et al., 2024) and in reinforcement learning with posterior transition models (Tiapkin et al., 2022). Related randomized exploration uses absolute perturbations (Kitamura et al., 2026). Our fresh-draw analysis complements guarantees for maintained ensembles (Qin et al., 2022). More recent work sharpens the analysis of linear ensemble sampling (Janz et al., 2026) and studies action-distribution mismatch along adaptive histories (Qin et al., 2022).

Active reward modeling uses Fisher-information design (Shen et al., 2025). ActiveDPO (Lin et al., 2026) and ActiveUltraFeedback (Melikidze et al., 2026) select data for preference fine-tuning, while XPO studies online policy exploration (Xie et al., 2025). OLE's reference is fixed before the current reward draws, giving a contextual logistic feedback model (Faury et al., 2020). Methods that choose both comparison members include Double Thompson Sampling (Wu & Liu, 2016), FGTS.CDB (Li et al., 2024a) and PF-TS (Lazzaro et al., 2026), with different sampling laws and regret objectives. Appendix A.2 compares these guarantees.

Reward ensembles (Coste et al., 2024) and RewardUQ (Yang et al., 2026) estimate neural reward uncertainty. Joint-prediction analysis relates dependence to decisions (Wen et al., 2021), while the Neural Testbed evaluates joint predictions empirically (Osband et al., 2022). Stratified integration includes Latin hypercube sampling (Stein, 1987) and orthogonal-array methods (Owen, 1992). At the application level, ReSearch (Chen et al., 2025), Search-R1 (Jin et al., 2025) and Mem1 (Zhou et al., 2025) learn reasoning, search and memory behavior from generated trajectories. Agentic OLE studies reward-sensitive rejection within such a learning system.

7 Discussion and conclusion

OLE treats finite reward sampling as part of an adaptive learning procedure. The ensemble selects a comparison, and its feedback changes later decisions. The analysis connects three requirements: control the reward value of the selected response, account for the information gained through exploration, and compute the intended selection rule accurately.

These findings motivate evaluating sampling choices through the selection rule and learning update they support. A general criterion for predicting which feedback improves learning under a limited computational budget remains open.

AI USE STATEMENT

Generative AI tools were used in a limited capacity to assist with software build configuration, debugging of experimental code, and language polishing for improved clarity and organization to mitigate potential misunderstanding. The authors reviewed and verified all code changes and manuscript revisions. All numerical results were obtained by executing the experimental code. Responsibility for the submitted claims, proofs, experiments, and results remains with the authors.

ETHICS STATEMENT

This work studies preference exploration through theoretical analysis, controlled learning experiments, and evaluations using existing models and datasets. Human preference information used in the replay experiments comes from existing benchmark annotations rather than newly elicited judgments. Protected human-ranking data are excluded from the supplementary implementation. Preference-based selection can inherit biases in the underlying feedback and reward models, and optimistic exploration can temporarily select lower-quality alternatives to acquire information. Our theoretical guarantees concern query regret under explicit modeling assumptions; they do not establish fairness, safety, or reliability outside those assumptions. The agentic evaluations assess benchmark performance and should not be interpreted as evidence of readiness for unsupervised deployment in consequential settings.

REPRODUCIBILITY STATEMENT

The preference-learning setup, selection rules, and theoretical guarantees are specified in Sections 2–4, with complete assumptions and proofs in Appendices B–C. Experimental protocols and implementation details are provided in Appendices D–G.1. The supplementary implementation supports controlled preference acquisition, neural selector evaluation, and retained-ensemble analysis. Reproducing the neural experiments additionally requires the corresponding external model weights and datasets; protected human-ranking data are not redistributed. The agentic experiments are accompanied by reported results and methodological details, but a complete executable reproduction of the training runs is not included. Appendix G.1 documents the available settings and unresolved implementation details that limit exact replication.

REFERENCES

Emile Anand and Sarah Liaw. Feel-good Thompson sampling for contextual bandits: a Markov Chain Monte Carlo showdown. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, volume 38, 2025.

Sébastien Bubeck, Ronen Eldan, and Joseph Lehec. Sampling from a log-concave distribution with projected Langevin Monte Carlo. *Discrete & Computational Geometry*, 59(4):757–783, 2018.

Mingyang Chen, Linzhuang Sun, Tianpeng Li, Haoze Sun, Yijie Zhou, Chenzheng Zhu, Haofen Wang, Jeff Z. Pan, Wen Zhang, Huajun Chen, Fan Yang, Zenan Zhou, and Weipeng Chen. ReSearch: Learning to reason with search for LLMs via reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 38, pp. 85287–85307. Curran Associates, Inc., 2025.

Thomas Coste, Usman Anwar, Robert Kirk, and David Krueger. Reward model ensembles help mitigate overoptimization. In *International Conference on Learning Representations*, 2024.

Guanting Dong, Licheng Bao, Zhongyuan Wang, Kangzhi Zhao, Xiaoxi Li, Jiajie Jin, Jinghan Yang, Hangyu Mao, Fuzheng Zhang, Kun Gai, Guorui Zhou, Yutao Zhu, Ji-Rong Wen, and Zhicheng Dou. Agentic entropy-balanced policy optimization, 2025a.

Guanting Dong, Hangyu Mao, Kai Ma, Licheng Bao, Yifei Chen, Zhongyuan Wang, Zhongxia Chen, Jiazhen Du, Huiyang Wang, Fuzheng Zhang, Guorui Zhou, Yutao Zhu, Ji-Rong Wen, and Zhicheng Dou. Agentic reinforced policy optimization, 2025b.

Mingxuan Du, Benfeng Xu, Chiwei Zhu, Xiaorui Wang, and Zhendong Mao. DeepResearch Bench: A comprehensive benchmark for deep research agents. *arXiv preprint arXiv:2506.11763*, 2025.

Louis Faury, Marc Abeille, Clément Calauzènes, and Olivier Fercoq. Improved optimistic algorithms for logistic bandits. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 3052–3060. PMLR, 2020.

Sarah Filippi, Olivier Cappé, Aurélien Garivier, and Csaba Szepesvári. Parametric bandits: The generalized linear case. In *Advances in Neural Information Processing Systems*, volume 23, 2010.

Evan Frick, Tianle Li, Connor Chen, Wei-Lin Chiang, Anastasios N. Angelopoulos, Jiantao Jiao, Banghua Zhu, Joseph E. Gonzalez, and Ion Stoica. How to evaluate reward models for RLHF, 2024. arXiv:2410.14872.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 6609–6625, Barcelona, Spain (Online), 2020. International Committee on Computational Linguistics.

Haque Ishfaq, Yixin Tan, Yu Yang, Qingfeng Lan, Jianfeng Lu, A. Rupam Mahmood, Doina Precup, and Pan Xu. More efficient randomized exploration for reinforcement learning via approximate sampling. *Reinforcement Learning Journal*, 3:1211–1235, 2024.

Haque Ishfaq, Guangyuan Wang, Sami Nur Islam, and Doina Precup. Langevin soft actor-critic: Efficient exploration through uncertainty-driven critic learning. In *International Conference on Learning Representations*, 2025.

David Janz, Alexander E. Litvak, and Csaba Szepesvári. Ensemble sampling for linear bandits: Small ensembles suffice. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

David Janz, Arya Akhavan, and Csaba Szepesvári. Sharp analysis of linear ensemble sampling. In *Proceedings of Thirty Ninth Conference on Learning Theory*, volume 336 of *Proceedings of Machine Learning Research*, pp. 3716–3750. PMLR, 2026.

Xinke Jiang, Yue Fang, Rihong Qiu, Haoyu Zhang, Yongxin Xu, Hao Chen, Wentao Zhang, Ruizhe Zhang, Yuchen Fang, Xinyu Ma, Xu Chu, Junfeng Zhao, and Yasha Wang. TC–RAG: Turing–complete RAG's case study on medical LLM systems. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11400–11426. Association for Computational Linguistics, 2025.

Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan Arik, Dong Wang, Hamed Zamani, and Jiawei Han. Search-R1: Training LLMs to reason and leverage search engines with reinforcement learning, 2025.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? A large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421, 2021.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1601–1611, Vancouver, Canada, 2017. Association for Computational Linguistics.

Toshinori Kitamura, Shuai Liu, and Csaba Szepesvári. Randomized exploration for linear bandits via absolute perturbations. *Reinforcement Learning Journal*, 7, 2026.

Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi Rui Tam, Keith Stevens, Abdullah Barhoum, Duc Nguyen, Oliver Stanley, Richárd Nagyfi, Shahul ES, Sameer Suri, David Glushkov, Arnav Dantuluri, Andrew Maguire, Christoph Schuhmann, Huu Nguyen, and Alexander Mattick. OpenAssistant Conversations - democratizing large language model alignment. In *Advances in Neural Information Processing Systems*, volume 36, pp. 47669–47681, 2023.

Satyapriya Krishna, Kalpesh Krishna, Anhad Mohananey, Steven Schwarcz, Adam Stambler, Shyam Upadhyay, and Manaal Faruqui. Fact, fetch, and reason: A unified evaluation of retrieval-augmented generation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 4745–4759, Albuquerque, New Mexico, 2025. Association for Computational Linguistics.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466, 2019.

Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, Noah A. Smith, and Hannaneh Hajishirzi. RewardBench: Evaluating reward models for language modeling. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 1755–1797. Association for Computational Linguistics, 2025.

Joseph Lazzaro, Davide Buffelli, Da-shan Shiu, and Sattar Vakili. A finite time analysis of Thompson sampling for Bayesian optimization with preferential feedback. In *Proceedings of the 29th International Conference on Artificial Intelligence and Statistics*, volume 300 of *Proceedings of Machine Learning Research*, pp. 991–999. PMLR, 2026.

Joseph Lehec. The Langevin Monte Carlo algorithm in the non-smooth log-concave case. *The Annals of Applied Probability*, 33(6A), 2023.

Xuheng Li, Heyang Zhao, and Quanquan Gu. Feel-good Thompson sampling for contextual dueling bandits. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 29406–29426. PMLR, 2024a.

Yingru Li, Jiawei Xu, Lei Han, and Zhi-Quan Luo. Q-Star meets scalable posterior sampling: Bridging theory and practice via HyperAgent. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 29022–29062. PMLR, 2024b.

Yingru Li, Jiawei Xu, Baoxiang Wang, and Zhiquan Luo. Scalable exploration via Ensemble++. In *Advances in Neural Information Processing Systems*, volume 38, pp. 55083–55145, 2025.

Xiaoqiang Lin, Arun Verma, Zhongxiang Dai, Daniela Rus, See-Kiong Ng, and Bryan Kian Hsiang Low. ActiveDPO: Active direct preference optimization for sample-efficient alignment. In *International Conference on Learning Representations*, 2026.

László Lovász and Santosh Vempala. The geometry of logconcave functions and sampling algorithms. *Random Structures & Algorithms*, 30(3):307–358, 2007.

Xiuyuan Lu and Benjamin Van Roy. Ensemble sampling. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30, pp. 3258–3266. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/hash/49ad23d1ec9fa4bd8d77d02681df5cfa-Abstract.html.

Saumya Malik, Valentina Pyatkin, Sander Land, Jacob Morrison, Noah A. Smith, Hannaneh Hajishirzi, and Nathan Lambert. RewardBench 2: Advancing reward model evaluation. In *International Conference on Learning Representations*, 2026.

Benedict C. May, Nathan Korda, Anthony Lee, and David S. Leslie. Optimistic Bayesian sampling in contextual-bandit problems. *Journal of Machine Learning Research*, 13(67):2069–2106, 2012.

Eric Mazumdar, Aldo Pacchiano, Yian Ma, Michael Jordan, and Peter Bartlett. On approximate Thompson sampling with Langevin algorithms. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 6797–6807. PMLR, 2020.

Viraj Mehta, Syrine Belakaria, Vikramjeet Das, Ojash Neopane, Yijia Dai, Ilija Bogunovic, Barbara Engelhardt, Stefano Ermon, Jeff Schneider, and Willie Neiswanger. Sample efficient preference alignment in LLMs via active exploration. In *Conference on Language Modeling*, 2025.

Davit Melikidze, Marian Schneider, Jessica Lam, Martin Wertich, Ido Hakimi, Barna Pásztor, and Andreas Krause. ActiveUltraFeedback: Efficient preference data generation using active learning. In *International Conference on Machine Learning*, 2026.

Ian Osband, Zheng Wen, Seyed Mohammad Asghari, Vikranth Dwaracherla, Xiuyuan Lu, Morteza Ibrahimi, Dieterich Lawson, Botao Hao, Brendan O'Donoghue, and Benjamin Van Roy. The neural testbed: Evaluating joint predictions. In *Advances in Neural Information Processing Systems*, volume 35, pp. 12554–12565, 2022.

Art B. Owen. Orthogonal arrays for computer experiments, integration and visualization. *Statistica Sinica*, 2(2):439–452, 1992. URL https://www3.stat.sinica.edu.tw/statistica/j2n2/j2n27/j2n27.htm.

Art B. Owen. Monte Carlo variance of scrambled net quadrature. *SIAM Journal on Numerical Analysis*, 34(5):1884–1910, 1997.

Akshara Prabhakar, Roshan Ram, Zixiang Chen, Silvio Savarese, Frank Wang, Caiming Xiong, Huan Wang, and Weiran Yao. Enterprise deep research: Steerable multi-agent deep research for enterprise analytics, 2025.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A. Smith, and Mike Lewis. Measuring and narrowing the compositionality gap in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 5687–5711, Singapore, 2023. Association for Computational Linguistics.

Chao Qin, Zheng Wen, Xiuyuan Lu, and Benjamin Van Roy. An analysis of ensemble sampling. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2.5 technical report, 2025. arXiv:2412.15115.

Daniel Russo and Benjamin Van Roy. Learning to optimize via posterior sampling. *Mathematics of Operations Research*, 39(4):1221–1243, 2014.

Daniel Russo and Benjamin Van Roy. An information-theoretic analysis of Thompson sampling. *Journal of Machine Learning Research*, 17(68):1–30, 2016.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models, 2024.

Yunyi Shen, Hao Sun, and Jean-François Ton. Active reward modeling: Adaptive preference labeling for large language model alignment. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 54410–54430. PMLR, 2025.

Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht. ALFWorld: Aligning text and embodied environments for interactive learning. In *International Conference on Learning Representations*, 2021.

Michael Stein. Large sample properties of simulations using Latin hypercube sampling. *Technometrics*, 29(2):143–151, 1987.

Daniil Tiapkin, Denis Belomestny, Daniele Calandriello, Eric Moulines, Remi Munos, Alexey Naumov, Mark Rowland, Michal Valko, and Pierre Ménard. Optimistic posterior sampling for reinforcement learning with few samples and tight guarantees. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. MuSiQue: Multihop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554, 2022.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10014–10037. Association for Computational Linguistics, 2023.

Aad W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 1998.

Arun Verma, Zhongxiang Dai, Xiaoqiang Lin, Patrick Jaillet, and Bryan Kian Hsiang Low. Neural dueling bandits: Preference-based optimization with human feedback. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 97145–97182, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/f1e7c90552850afcc2558d78950c519d-Abstract-Conference.html.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 24824–24837. Curran Associates, Inc., 2022.

Zheng Wen, Ian Osband, Chao Qin, Xiuyuan Lu, Morteza Ibrahimi, Vikranth Dwaracherla, Mohammad Asghari, and Benjamin Van Roy. From predictions to decisions: The importance of joint predictive distributions, 2021. arXiv:2107.09224.

Di Wu, Chengshuai Shi, Jing Yang, and Cong Shen. Greedy sampling is provably efficient for RLHF. In *Advances in Neural Information Processing Systems*, volume 38, pp. 108198–108232, 2025.

Huasen Wu and Xin Liu. Double Thompson sampling for dueling bandits. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

Tengyang Xie, Dylan J. Foster, Akshay Krishnamurthy, Corby Rosset, Ahmed Awadallah, and Alexander Rakhlin. Exploratory preference optimization: Harnessing implicit Q*-approximation for sample-efficient RLHF. In *International Conference on Learning Representations*, 2025.

Wei Xiong, Hanze Dong, Chenlu Ye, Ziqi Wang, Han Zhong, Heng Ji, Nan Jiang, and Tong Zhang. Iterative preference learning from human feedback: Bridging theory and practice for RLHF under KL-constraint. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 54715–54754. PMLR, 2024.

Daniel Yang, Samuel Stante, Florian Redhardt, Lena Libon, Parnian Kassraie, Ido Hakimi, Barna Pásztor, and Andreas Krause. RewardUQ: A unified framework for uncertainty-aware reward models, 2026. arXiv:2602.24040.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2369–2380, Brussels, Belgium, 2018. Association for Computational Linguistics.

Shunyu Yao, Howard Chen, John Yang, and Karthik Narasimhan. WebShop: Towards scalable real-world web interaction with grounded language agents. In *Advances in Neural Information Processing Systems*, volume 35, pp. 20744–20757. Curran Associates, Inc., 2022.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*, 2023.

Yao Zhao and Kwang-Sung Jun. Nearly optimal active preference learning and its application to LLM alignment. *arXiv preprint arXiv:2602.01581*, 2026.

Zijian Zhou, Ao Qu, Zhaoxuan Wu, Sunghwan Kim, Alok Prakash, Daniela Rus, Jinhua Zhao, Bryan Kian Hsiang Low, and Paul Pu Liang. MEM1: Learning to synergize memory and reasoning for efficient long-horizon agents, 2025.

GUIDE TO THE APPENDIX

Appendix A relates the contribution to earlier guarantees. Appendices B and C provide complete proofs, sampling constructions and implementation bounds. The remaining appendices follow the experiments: controlled preference learning (Appendix D), neural selection (Appendix E), retained ensembles and additional controls (Appendix F), and agentic learning (Appendix G.1). Study-specific protocols and uncertainty calculations accompany the empirical results.

A POSITIONING AND SCOPE

A.1 IMPLEMENTATION CONVENTIONS

Two-sample contrasts denoted OLE–Mean2 compare optimistic and mean selection. The controlled learner resamples two complete posterior reward functions after each comparison and selects by their mean or maximum. The retained learner updates 20 particles and compares mean selection with mean plus 0.3 times the sample standard deviation. Both update reward inference from the acquired preference labels. The fixed-model neural study uses 20 dropout evaluations, mean plus twice the sample standard deviation, and a temperature-one softmax. Its fitted reward model stays fixed, so the outcome is numerical selection precision. Agentic OLE instead uses realized trajectory rewards to select or weight policy-training experience. Its trajectory and prompt-group implementations are specified in Appendix G.1.

A.2 RELATED GUARANTEES AND THEIR ASSUMPTIONS

The specific contribution. Theorem 3.1 controls the response selected from the same fresh pair that defines its score, under bounded linear-BTL feedback against a predictable anchor. It allows unrestricted pair dependence and charges conditional marginal error through $W_2(\nu_t, \Pi_{t-1})$. The comparisons below distinguish this combination from prior optimistic and posterior-sampling guarantees. The two-value identity and maximum selection are not individually new, and the bound asserts neither optimal ensemble size nor computational superiority. The finite-N extension gives the minimal deterministic envelope normalization and its explicit N-dependent costs.

Optimistic sampling and maintained ensembles. Optimistic Bayesian Sampling (OBS) clips each action's sampled posterior value below at its posterior mean and establishes asymptotic average-reward consistency (May et al., 2012, Alg. 2, Thm. 2). Its contextual experiments include binary logistic rewards. OPSRL uses optimistic maxima over logarithmically many transition-posterior samples per state–action pair in episodic tabular MDPs (Tiapkin et al., 2022). Dirichlet anti-concentration gives its high-probability $\widetilde{O}(\sqrt{H^3SAT})$ regret bound. The small-ensemble analysis of Janz et al. (2024, Remark 5) also covers maximum selection. For retained perturbed linear estimates with $M = \Theta(d \log T)$, Janz et al. (2026) obtain high-probability $\widetilde{O}(d^{3/2}\sqrt{T})$ regret. Their $M \leq d/2$ lower bound concerns the specified retained ensemble-sampling algorithm. Theorem 3.1 analyzes two fresh approximate posterior draws with arbitrary dependence, under bounded linear logistic preference feedback.

Posterior approximation and action distributions. Ensemble sampling chooses a member of an incrementally updated ensemble (Lu & Van Roy, 2017). For linear-Gaussian scalar feedback, Qin et al. (2022, Thms. 1–2) bound Bayesian regret by a Thompson-sampling-order term plus an action-distribution mismatch of order $\widetilde{O}(T\sqrt{K \min\{d, \log K\}/M})$, accounting for adaptive histories. Langevin Thompson sampling instead controls approximation through Wasserstein error for per-arm smooth, strongly log-concave reward families (Mazumdar et al., 2020, Thms. 3–4). Warm-start ULA/SGLD reuses the previous sample, with instance-dependent logarithmic regret and horizon-independent SGLD step and batch counts under the stated assumptions. Our guarantee charges marginal parameter error directly in selected reward value. Our separate projected-chain certificate uses additional geometric conditions and provides a sufficient computational budget.

Preference queries and regret objectives. Double Thompson Sampling draws independent Beta posterior samples and uses confidence filtering to choose both comparison members in finite-arm

Copeland dueling (Wu & Liu, 2016, Alg. 1). Its guarantee is logarithmic, gap-dependent dueling regret. Under linear Bradley–Terry feedback, FGTS.CDB draws two independent models from optimism-modified distributions to select both responses, with $\widetilde{O}(d\sqrt{T})$ expected average two-response regret (Li et al., 2024a, Thm. 5.2). PF-TS selects both points from independent draws of an inflated Gaussian surrogate in preference Bayesian optimization, with a high-probability information-gain bound for pair regret in an RKHS (Lazzaro et al., 2026, Thm. 1). OLE's analyzed instance uses ordinary posterior marginals and a reference chosen before sampling. Its Bayesian query regret charges the selected response, and its approximation term is explicit. These different feedback interfaces, sampling laws and regret objectives determine what each guarantee establishes.

Why the error terms are different. Small parameter Wasserstein error does not in general imply a small discrepancy between maximizing-action distributions. For example, for rewards $r_\theta(\pm) = 1/2 \pm \theta/2$, let $\Pi = \mathrm{Unif}[-\varepsilon, 0]$ and $\nu = \mathrm{Unif}[0, \varepsilon]$, with $0 < \varepsilon \leq 1$. Then $W_2(\nu, \Pi) = \varepsilon$, but their sampled maximizing actions are opposite almost surely. Consequently an action-law approximation bound does not directly yield our marginal-W_2 residual. At the proof level, Janz et al. (2024, Thm. 2) control conditional optimism probabilities within confidence ellipsoids, while Mazumdar et al. (2020, Sec. 4, Thms. 3–4) combine Wasserstein sampler control with concentration and anti-concentration for their Langevin TS rule. Here posterior marginality supplies optimism in expectation after coordinatewise transport of the actual pair. The expected uniform score deviation is at most $\sqrt{2}Se_t$. Comparing the actual selected value with the transported maximum pays this deviation twice. The remaining selected posterior width is charged to information in the comparison actually queried, giving the additive marginal-error term under arbitrary within-pair dependence (Appendix B.2). This supplements the established optimism and posterior-sampling tools rather than replacing them.

What the dependence statement covers. For the main theorem, each parameter particle defines one gap function over all candidates. Arbitrary dependence is allowed between the two particles, while each conditional marginal remains specified. The finite-N extension allows the same for the whole tuple under its stronger normalization. The general bound alone does not prove that dependence improves regret. Appendix B.5 gives an instance-specific strict comparison for an exact quantile coupling. In particular, identical particles recover exact TS (Russo & Van Roy, 2014) when $\nu_t = \Pi_{t-1}$ and approximate posterior sampling under ν_t otherwise. The neural diagnostic instead samples candidate-specific dropout streams. Its balancing preserves each member's per-candidate dropout marginal, not a joint reward-function draw shared across candidates or a calibrated Bayesian posterior.

Imported results and conditions. Appendix B.1 applies the linear-marginal and standardized-tail results of Lovász & Vempala (2007, Thm. 5.1 and Lem. 5.17) to the log-concave logistic posterior. Appendix B.7 states the convex-support, interior-start and Lipschitz conditions of Lehec (2023, Thm. 1) before its substitution $h = 2\eta_t$, $k = J_t$ in (48). Its contraction and minibatch bounds additionally require $\rho > 0$ and $\eta_t \leq 1/M_t$. The information and transport arguments are proved here.

When the regret bound is informative. Table 2 uses natural logarithms and $\lambda = \sigma(1)(1-\sigma(1))/2$. Its approximate-law column assumes $e_t \leq [2\sqrt{2}S\sqrt{T}]^{-1}$ at every round, giving residual at most $\sqrt{T}$. These conditional examples illustrate the sufficient bound, including its vacuous configurations. They are not fitted simulator bounds.

A.3 Preference acquisition and joint reward inference

Acquisition objectives. Active Reward Modeling chooses pairs for reward-model fitting through Fisher-information design (Shen et al., 2025). XPO adds exploration to online DPO and analyzes policy learning (Xie et al., 2025), while Greedy Sampling gives guarantees for empirical-estimate policies in KL-regularized RLHF (Wu et al., 2025). Active preference alignment (Mehta et al., 2025), ActiveDPO (Lin et al., 2026) and neural dueling bandits (Verma et al., 2025) study further feedback-acquisition settings. Zhao & Jun (2026) target preference signs on a fixed query set, and ActiveUltraFeedback selects both responses for dataset construction and subsequent fine-tuning (Melikidze et al., 2026). OLE instead controls the cumulative reward gap of a selected response against a predictable anchor. Annotation efficiency and trained-policy quality are different endpoints.

Table 2: Literal evaluations of Theorem 3.1, with $B = 1$. The exact-law bound is shown before capping it by T.

d	K	S	T	Exact-law bound	Capped bound with $\sqrt{T}$ residual
1	2	0.5	100	71.64	81.64
1	2	0.5	2000	400.78	445.50
8	20	1	2000	3430.96	2000.00
8	20	1	100000	29017.52	29333.75

Implementable posterior exploration. Posterior sampling provides a general framework for sequential optimization (Russo & Van Roy, 2014). Approximate sampling provides one route to efficient randomized exploration (Ishfaq et al., 2024). Langevin Soft Actor-Critic uses uncertainty-driven critic learning (Ishfaq et al., 2025). Scalable posterior sampling is studied in HyperAgent (Li et al., 2024b), while Ensemble++ develops scalable ensemble exploration (Li et al., 2025). Anand & Liaw (2025) investigate interactions between approximation and optimism. Appendix A.2 compares the closest guarantees and their feedback, sampling and regret assumptions.

Which joint law is changed? Joint-prediction analyses (Wen et al., 2021) and the Neural Testbed (Osband et al., 2022) study correlations across predictions. Our distinction between shared parameter draws and candidate-specific dropout streams is stated in Appendix A.2. Reward ensembles (Coste et al., 2024) and RewardUQ (Yang et al., 2026) provide the neural uncertainty context. Agentic learning instead uses realized rewards to select training trajectories. The corresponding reward source, selection unit and update are specified in Appendix G.1.

B Proofs for preference exploration

B.1 Statistical foundations

We use Assumption 2.1 and the posterior in (5). For $S > 0$, $\Gamma_T = d\log(1 + 2BS\sqrt{T}) + 1/8$ and $\lambda = \sigma(1)(1 - \sigma(1))/2$. When $S = 0$, all query regrets vanish. The proof first establishes posterior calibration and information control, then applies them to the comparison selected by OLE.

B.1.1 Conditioning and the posterior

We specify the filtration to account for adaptive actions and auxiliary randomness. At the start of round t, let $\mathcal{G}_t$ contain the entire completed transcript through round $t - 1$ and the newly revealed context, reference, and candidate set. Their conditional generation kernels are parameter-independent. Given $\mathcal{G}_t$, the new sampler randomness is independent of $\theta^\star$. Let $\mathcal{F}_t$ enlarge $\mathcal{G}_t$ by this randomness, its resulting particles, any auxiliary coupled particles used in the analysis, and the selected action. Thus

$$\mathrm{Law}(\theta^\star \mid \mathcal{G}_t) = \mathrm{Law}(\theta^\star \mid \mathcal{F}_t) = \Pi_{t-1}.$$

This identity concerns fresh parameter-independent kernels. It does not justify treating particles already present in the conditioning history as independent posterior draws. Ties in action selection use a measurable parameter-independent rule. All regular conditional laws exist on the standard Borel spaces considered here. Candidate enumeration is measurable and fixed before current sampling.

For a realized history, set $\psi_i = \psi(x_i, y_i)$ and $A(u) = \log(1 + e^u)$. Relative to Lebesgue measure on Θ, Bayes' formula is

$$\pi_{t-1}(\theta) = \frac{1}{Z_{t-1}}\pi_0(\theta)\exp\left\{\sum_{i<t}[b_i\psi_i^\top\theta - A(\psi_i^\top\theta)]\right\}. \tag{11}$$

Parameter-independent action and auxiliary kernels cancel in Bayes' formula. Since A is convex, every likelihood factor is log-concave in θ. Their product with the prior, extended by zero outside the convex support, is log-concave. The positive bounded likelihood ensures a finite, nonzero normalizer.

B.1.2 CALIBRATION AND ADAPTIVE INFORMATION

Lemma B.1 (Posterior information and calibration)**.** *Under Assumption 2.1,*

$$\lambda\,\mathbb{E}\sum_{t=1}^{T} s_t(y_t)^2 \le I(\theta^\star; H_T) \le \Gamma_T. \tag{12}$$

This information bound also holds for arbitrary priors on Θ. For the log-concave prior, positive-variance gaps satisfy, for $u > 1$,

$$\Pr\{|g_{\theta^\star}(x_t, y) - \mu_t(y)| > u s_t(y) \mid \mathcal{G}_t\} \le e^{1-u}. \tag{13}$$

Zero-variance gaps equal their posterior means almost surely.

Proof of Lemma B.1. The marginal log-concavity theorem and the second-moment tail inequality of Lovász & Vempala (2007, Theorem 5.1 and Lemma 5.17) imply the following: a log-concave real random variable X with mean μ and positive standard deviation s satisfies

$$\Pr\{|X - \mu| > us\} \le e^{1-u}, \qquad u > 1.$$

Indeed the centered, rescaled linear marginal $(X - \mu)/s$ is log-concave and has second moment one. Apply this to $X = \psi(x_t, y)^\top \theta$ under (11) and then use the exact posterior identity for $\theta^\star$. If $s = 0$, the nonnegative random variable $(X - \mu)^2$ has expectation zero and is zero almost surely. This proves (13).

For the information upper bound, take a maximal set of points $c_1, \dots, c_M \in \Theta$ with pairwise distance greater than a. Compactness makes it finite, maximality makes it an a-net, and the disjoint open radius-$a/2$ balls centered at these points are contained in the radius-$(B + a/2)$ ball. Comparing volumes gives $M \le (1 + 2B/a)^d$. Assign each θ to its closest net point, breaking ties by index, and write this finite-valued assignment as $Z = c(\theta^\star)$.

Let P_θ denote the law of the full adaptive transcript with true parameter fixed to θ. For the same realized prefix, all environment, action, and auxiliary kernels in P_θ and P_c are identical. The relative-entropy chain rule therefore gives

$$\mathrm{KL}(P_\theta \| P_c) = \sum_{t=1}^{T} \mathbb{E}_{P_\theta}\, \mathrm{kl}\big(\sigma(\psi_t^\top \theta), \sigma(\psi_t^\top c)\big)\,. \tag{14}$$

Here kl is binary relative entropy in nats and ψ_t is evaluated at the action in that prefix, including its sampler randomness. Strictly positive logistic label probabilities guarantee the needed absolute continuity. For logits u, v, the exponential-family identity and $0 \le A'' \le 1/4$ give

$$\mathrm{kl}(\sigma(u), \sigma(v)) = A(v) - A(u) - A'(u)(v - u) \le (u - v)^2/8.$$

Thus (14) is at most $TS^2\|\theta - c\|^2/8$.

For completeness, let Q_z be the transcript law conditional on $Z = z$. For any reference law P_z, integration of log density ratios yields

$$\mathbb{E}[\mathrm{KL}(P_{\theta^\star} \| P_z) \mid Z = z] = I(\theta^\star; H_T \mid Z = z) + \mathrm{KL}(Q_z \| P_z).$$

Choose P_z to be the law at net center z and drop the last, nonnegative term. Since Z is a function of $\theta^\star$,

$$\begin{aligned} I(\theta^\star; H_T) &= I(Z; H_T) + I(\theta^\star; H_T \mid Z) \\ &\le \log M + \mathbb{E}\,\mathrm{KL}(P_{\theta^\star} \| P_{c(\theta^\star)}) \le d \log(1 + 2B/a) + TS^2a^2/8. \end{aligned}$$

Taking $a = 1/(S\sqrt{T})$ gives Γ_T for $S > 0$. No nonadaptive-design restriction entered this argument.

It remains to lower-bound information by posterior variance. Conditional on $\mathcal{F}_t$, let $G = \psi_t^\top \theta^\star$, $\bar{p} = \mathbb{E}[\sigma(G) \mid \mathcal{F}_t]$, and $z = \log(\bar{p}/(1 - \bar{p}))$. Since $G \in [-1, 1]$, also $z \in [-1, 1]$. The predictive label law is $\mathrm{Ber}(\bar{p})$. On this interval $A''(u) = \sigma'(u) \ge c = \sigma(1)(1 - \sigma(1))$, so strong convexity gives

$$\begin{aligned} \mathbb{E}[\mathrm{kl}(\sigma(G), \bar{p}) \mid \mathcal{F}_t] &= \mathbb{E}[A(z) - A(G) - A'(G)(z - G) \mid \mathcal{F}_t] \\ &\ge (c/2)\,\mathbb{E}[(G - z)^2 \mid \mathcal{F}_t] \ge (c/2)\ \mathrm{Var}(G \mid \mathcal{F}_t) = (c/2)\, s_t(y_t)^2. \end{aligned}$$

Finally apply the mutual-information chain rule to the full transcript. The conditional information in all non-label kernels is zero, including current independent sampler randomness. Taking expectations of the label bounds and summing yields (12) with $\lambda = c/2$. This is an inequality between expectations, not a pathwise identity. The information argument requires no log-concavity: for an arbitrary prior on Θ, Bayes' formula is interpreted relative to that prior measure, and all transcript and variance calculations above are unchanged. □

B.2 Two-particle OLE with arbitrary within-pair dependence

This result uses optimism in expectation of a maximizing value, rather than a simultaneous upper confidence bound. The identity in (6) is specific to two particles, denominator-2 empirical variance and bonus 1. It is not asserted for denominator-1 variance or a larger ensemble with the same bonus.

Proof of Theorem 3.1. Fix a pre-round history $\mathcal{G}_t$ and suppress the round index. Let (X_1, X_2) have the actual joint particle law. Both marginals are ν, and the whole pair is independent of $\theta^\star$ conditional on this history. For two scores the empirical SD is $|g_1 - g_2|/2$, so the implemented index is $U^X(y) = \max_i g_{X_i}(x, y)$.

Choose an optimal quadratic coupling of ν and Π, and disintegrate it as $\nu(dx)Q(dy \mid x)$. Given the actual pair, draw $Y_i \sim Q(\cdot \mid X_i)$ using fresh auxiliary randomness. Then each Y_i has posterior marginal Π, and $\mathbb{E}\|X_i - Y_i\|^2 = e^2$. No claim is made about independence of Y_1, Y_2 or agreement with a prescribed posterior pair law. Compactness and standard Borel disintegration justify this construction. Measurable near-optimal couplings and a limit give the same bound. The auxiliary kernel is parameter-independent, including conditional on the actual pair, so the posterior identity and information bound continue to apply along the actual learner's history.

Write $U^Y(y) = \max_i g_{Y_i}(x, y)$ and $\Delta = \max_y |U^X(y) - U^Y(y)|$. Lipschitz continuity implies

$$\Delta \le S \max_i \|X_i - Y_i\|, \qquad \mathbb{E}[\Delta \mid \mathcal{G}_t] \le \sqrt{2} S e. \tag{15}$$

Indeed the squared maximum is at most the sum of the two squared distances. Since the first auxiliary marginal is a posterior draw,

$$\mathbb{E}[\max_y U^Y(y) \mid \mathcal{G}_t] \ge \mathbb{E}[\max_y g_{Y_1}(x, y) \mid \mathcal{G}_t] = \mathbb{E}[\max_y g_{\theta^\star}(x, y) \mid \mathcal{G}_t]. \tag{16}$$

This step does not require independent particles or posterior log-concavity.

Let A denote the action selected by X. Conditional independence from $\theta^\star$ gives $\mathbb{E}[g_{\theta^\star}(x, A) \mid \mathcal{G}_t] = \mathbb{E}[\mu(A) \mid \mathcal{G}_t]$. Since A maximizes U^X, pathwise $\max_y U^Y(y) \le U^Y(A) + 2\Delta$. Consequently

$$\mathbb{E}[r_t \mid \mathcal{G}_t] \le \mathbb{E}[U^Y(A) - \mu(A) \mid \mathcal{G}_t] + 2\sqrt{2} S e. \tag{17}$$

For coordinates with $s(y) > 0$ put $Z_{i,y} = [g_{Y_i}(x, y) - \mu(y)]/s(y)$, and put $Z_{i,y} = 0$ when $s(y) = 0$. In the latter case each exact-marginal sample equals the gap mean almost surely. Define $M = \max_{i \le 2, y \in \mathcal{Y}} (Z_{i,y})_+$. Then $U^Y(A) - \mu(A) \le s(A)M$. We give three conditional second-moment bounds for M.

First, $\mathbb{E}Z_{i,y}^2 \le 1$ and the squared maximum is bounded by the sum, so $\mathbb{E}[M^2 \mid \mathcal{G}_t] \le 2K$. Second, let $m = \mathbb{E}_\Pi \theta$ and $\Sigma = \mathrm{Cov}_\Pi(\theta)$. Write $W_i = \Sigma^{\dagger/2}(Y_i - m)$, using the Moore–Penrose pseudoinverse. Centered samples lie in the range of Σ almost surely: their components along every zero-variance direction vanish. For a positive-variance gap,

$$Z_{i,y} = \left(\frac{\Sigma^{1/2}\psi(x, y)}{s(y)} \right)^\top W_i, \qquad \left\| \frac{\Sigma^{1/2}\psi(x, y)}{s(y)} \right\| = 1.$$

Thus $M^2 \le \|W_1\|^2 + \|W_2\|^2$ and $\mathbb{E}[M^2 \mid \mathcal{G}_t] \le 2\,\mathrm{rank}(\Sigma) \le 2d$. Both arguments hold for an arbitrary posterior on Θ.

Third, for a log-concave posterior, (13) and a union bound give $\Pr(M > u \mid \mathcal{G}_t) \le 2Ke^{1-u}$ for $u > 1$, without any independence requirement among its coordinates. With $L_K = 1 + \log(2K)$, tail

integration yields

$$\mathbb{E}[M^2 \mid \mathcal{G}_t] \le L_K^2 + \int_{L_K}^{\infty} 2u(2K)e^{1-u}\,du$$
$$= L_K^2 + 2L_K + 2.$$

Taking the minimum proves $\mathbb{E}[M^2 \mid \mathcal{G}_t] \le C_K^2$.

Conditional Cauchy–Schwarz in (17) now gives

$$\mathbb{E}[r_t \mid \mathcal{G}_t] \le C_K\sqrt{\mathbb{E}[s_t(A_t)^2 \mid \mathcal{G}_t]} + 2\sqrt{2}Se_t.$$

Take expectations, apply Jensen, sum over rounds and use Cauchy–Schwarz. The information bound for the actual selected actions gives

$$\mathbb{E}R_T \le C_K\sqrt{T\,\mathbb{E}\sum_t s_t(A_t)^2} + 2\sqrt{2}S\sum_t \mathbb{E}e_t \le C_K\sqrt{T\Gamma_T/\lambda} + 2\sqrt{2}S\sum_t \mathbb{E}e_t.$$

For an arbitrary prior, omit the third second-moment argument. All other steps remain valid, proving the final assertion. □

Marginal sampler certificates and work. The statistical theorem allows any current conditional pair law. To apply the projected-chain certificate, each coordinate must retain the single-chain transition in (44), its stated start and its within-chain independence across inner steps. For example, opposite increments $(\xi_j, -\xi_j)$ have the correct Gaussian marginal for each chain. Minibatch coupling must likewise preserve each coordinate's fresh uniform-with-replacement law and its independence from that coordinate's current Gaussian increment. The proof of Theorem B.12 concerns one chain. It therefore gives the same marginal error bound under these couplings, without invoking independence across the two chains.

For $S > 0$, choose marginal targets

$$\varepsilon_t = \min\{D/2, [2\sqrt{2}S\sqrt{T}]^{-1}\}. \tag{18}$$

Corollary B.13 or the checkable construction in Appendix B.7.3 supplies finite step counts with $e_t \le \varepsilon_t$. The residual in (7) is then at most $\sqrt{T}$. At fixed positive geometry, the checkable construction uses exactly two particles and sufficient likelihood-gradient work $\widetilde{O}(T^5)$ for fixed minibatches or $\widetilde{O}(T^6)$ for full gradients. The orders follow from $J_t = \widetilde{O}((1+t)^2\varepsilon_t^{-4})$. Coupling does not improve these worst-case sufficient powers.

There is no relative-moment sample-size threshold in this theorem. That does not ensure its numerical regret bound is below the trivial bound T for a given horizon, especially after including law error. Nor does it identify the law error of a practical short chain. With identical particles the rule reduces to exact posterior sampling when the marginals are exact and approximate posterior sampling otherwise. The result therefore establishes neither a necessity for two samples nor a uniform gain from negative dependence.

B.3 A NORMALIZED FINITE-ENSEMBLE EXTENSION

The exact maximum identity is special to two samples. There is nevertheless a canonical normalization for every fixed ensemble size. For $N \ge 2$, define

$$h_N(z) = \bar{z} + \sqrt{N-1}\left\{\frac{1}{N}\sum_{i=1}^{N}(z_i - \bar{z})^2\right\}^{1/2}, \qquad z \in \mathbb{R}^N. \tag{19}$$

The SD in (19) has denominator N. With denominator $N-1$, the equivalent coefficient is $(N-1)/\sqrt{N}$.

Put

$$Q(m) = (1+\log m)^2 + 2(1+\log m) + 2$$

and

$$\kappa_N = \max_{0\le k\le N}\frac{2k - N + 2\sqrt{(N-1)k(N-k)}}{N} \le \sqrt{N}. \tag{20}$$

Proposition B.2 (Normalized finite-ensemble OLE)**.** *Under Assumption 2.1, fix $N \geq 2$ and use $a_t = \sqrt{N-1}$ in (2). Conditional on $\mathcal{G}_t$, generate the entire fresh tuple $X_{t,1:N}$ by a parameter-independent kernel, independently of $\theta^\star$, with every marginal equal to ν_t. Dependence within the tuple is unrestricted. If $e_t = W_2(\nu_t, \Pi_{t-1})$, then*

$$\mathbb{E}R_T \leq \min\left\{T,\, C_{N,K}\sqrt{T\Gamma_T/\lambda} + 2S\sqrt{N}\sum_{t=1}^{T}\mathbb{E}e_t\right\}, \tag{21}$$

where

$$C_{N,K}^2 = \min\{Nd, NK, NQ(K), \kappa_N^2 Q(NK)\}. \tag{22}$$

For an arbitrary prior on Θ, the same statement holds with $C_{N,K}^2 = \min\{Nd, NK\}$.

The proposition is an upper-bound extension, not a claim that a larger ensemble helps. For $N > 2$, h_N can exceed every sampled value and is not coordinatewise monotone: $h_3(1, 1, -M) = (M+4)/3$. The worst-case constants deteriorate with N when only marginal laws are controlled. At $N = 2$, $h_2(z) = \max_i z_i$, $\kappa_2 = 1$, and $Q(2K) \leq 2Q(K)$, so Proposition B.2 recovers Theorem 3.1 exactly.

Proof. Let $u = N^{-1/2}\mathbf{1}$ and $P = I - uu^\top$. Then

$$h_N(z) = N^{-1/2}u^\top z + \sqrt{(N-1)/N}\,\|Pz\|_2. \tag{23}$$

If $j \in \arg\max_i z_i$, the centered coordinates sum to zero, so

$$(z_j - \bar{z})^2 \leq (N-1)\sum_{i\neq j}(z_i - \bar{z})^2.$$

Hence $h_N(z) \geq \max_i z_i$. The coefficient is minimal: applying $\bar{z} + a\,\mathrm{SD}_{\mathrm{pop}}(z) \geq \max_i z_i$ to $z = (1, 0, \ldots, 0)$ requires $a \geq \sqrt{N-1}$. The two orthogonal terms in (23), the reverse triangle inequality and Cauchy–Schwarz also give

$$|h_N(z) - h_N(w)| \leq \|z - w\|_2. \tag{24}$$

Fix $\mathcal{G}_t$ and suppress the round index. Use the measurable coordinatewise coupling construction from the proof of Theorem 3.1 to augment the actual tuple $X_{1:N}$ with posterior-marginal variables $Y_{1:N}$ satisfying $\mathbb{E}[\|X_i - Y_i\|^2 \mid \mathcal{G}_t] = e_t^2$. No joint law is imposed on the Y_i. Write $U^X(y)$ and $U^Y(y)$ for (19) applied to their sampled gaps. From (24),

$$\Delta := \max_y |U^X(y) - U^Y(y)| \leq S\left(\sum_{i=1}^{N}\|X_i - Y_i\|^2\right)^{1/2}, \qquad \mathbb{E}_t\Delta \leq S\sqrt{N}e_t. \tag{25}$$

Because h_N upper-envelopes every coordinate and Y_1 has the posterior marginal,

$$\mathbb{E}_t \max_y U^Y(y) \geq \mathbb{E}_t \max_y g_{Y_1}(x_t, y) = \mathbb{E}_t \max_y g_{\theta^\star}(x_t, y).$$

If A maximizes U^X, then pathwise $\max_y U^Y(y) \leq U^Y(A) + 2\Delta$. The tuple is conditionally independent of $\theta^\star$, so $\mathbb{E}_t g_{\theta^\star}(x_t, A) = \mathbb{E}_t\mu_t(A)$. Thus

$$\mathbb{E}_t r_t \leq \mathbb{E}_t[U^Y(A) - \mu_t(A)] + 2S\sqrt{N}e_t. \tag{26}$$

For $s_t(y) > 0$, set $Z_{i,y} = [g_{Y_i}(x_t, y) - \mu_t(y)]/s_t(y)$ and set $Z_{i,y} = 0$ otherwise. Translation equivariance and positive homogeneity of h_N imply

$$U^Y(A) - \mu_t(A) \leq s_t(A)M, \qquad M = \max_y [h_N(Z_{1,y}, \ldots, Z_{N,y})]_+.$$

Since $h_N(0) = 0$ and it is 1-Lipschitz, $M^2 \leq \max_y \sum_i Z_{i,y}^2$. Summing unit marginal second moments gives $\mathbb{E}_t M^2 \leq NK$. Alternatively, with the posterior covariance Σ and $W_i = \Sigma^{\dagger/2}(Y_i - \mathbb{E}_t Y_i)$, every positive-variance standardized gap is $v_y^\top W_i$ for a unit vector v_y. Therefore $M^2 \leq \sum_i \|W_i\|^2$ and $\mathbb{E}_t M^2 \leq N\,\mathrm{rank}(\Sigma) \leq Nd$.

For the log-concave refinements, let $H_i = \max_y |Z_{i,y}|$. The marginal calibration inequality and a union bound give $\Pr_t(H_i > v) \le Ke^{1-v}$ for $v > 1$. Integrating from $1 + \log K$ yields $\mathbb{E}_t H_i^2 \le Q(K)$, and hence $\mathbb{E}_t M^2 \le NQ(K)$. Finally, convexity and positive homogeneity imply that the maximum of h_N on the unit ℓ_∞ cube occurs at a vertex. A vertex with k positive coordinates has value equal to the expression in (20). Consequently $[h_N(z)]_+ \le \kappa_N \|z\|_\infty$. Applying the same absolute-tail union bound over all NK standardized gaps gives $\mathbb{E}_t M^2 \le \kappa_N^2 Q(NK)$. This proves $\mathbb{E}_t M^2 \le C_{N,K}^2$. The two absolute-tail bounds are omitted for an arbitrary prior.

Conditional Cauchy–Schwarz in (26) now yields

$$\mathbb{E}_t r_t \le C_{N,K} \sqrt{\mathbb{E}_t s_t(A_t)^2} + 2S\sqrt{N} e_t.$$

Summing, applying Cauchy–Schwarz and invoking Lemma B.1 proves (21). Measurable near-optimal quadratic couplings followed by a limit cover the case where a measurable conditional optimizer is unavailable. □

If a sampler certificate is desired, the marginal target $e_t \le (2S\sqrt{NT})^{-1}$ caps the transport contribution in (21) by $\sqrt{T}$. For even N, this scaling and the example with half the coordinates equal to a symmetric standardized Z and half to $-Z$ show why the result cannot certify a benefit from increasing N under arbitrary within-tuple dependence.

B.4 The cost and value of an informative query

We first isolate the exact finite-pair intervention. For two candidates s, p and two parameter draws, abbreviate $g_a^i = g_{\theta_i}(a)$ and define

$$M = \frac{(g_p^1 - g_s^1) + (g_p^2 - g_s^2)}{2}, \qquad B = \frac{|g_p^1 - g_p^2| - |g_s^1 - g_s^2|}{2}. \tag{27}$$

The identity $\max\{a, b\} = (a + b + |a - b|)/2$ gives

$$\max_i g_p^i - \max_i g_s^i = M + B. \tag{28}$$

Thus Mean2 selects p iff $M > 0$, while OLE selects it iff $M + B > 0$, up to the fixed tie rule. In a linear model, $B = (|x_p^\top(\theta_1 - \theta_2)| - |x_s^\top(\theta_1 - \theta_2)|)/2$. The dispersion contrast, rather than posterior variance alone, is therefore the exact hard-decision object.

Proposition B.3 (When acquired information repays exploration)**.** *In a valid instance of Assumption 2.1, let $\theta = (c, z)$ be uniform on $[1/2, 3/2] \times [-1, 1]$ and fix the anchor reward at $1/2$. In round one, a safe response has gap δc and a probe has gap $z/2$. After its BTL label, round two offers gaps $z/2$ and $-z/2$. Compare the exact IID optimistic variant (F) with the mean variant (M) using the same two fresh posterior draws. Put $m = \int_0^1 z \tanh(z/4)\, dz$ and $k = 2\delta$. For $0 < \delta < 1/8$, their probe probabilities $q_{\mathrm{F}}, q_{\mathrm{M}}$ satisfy*

$$q_{\mathrm{F}} - q_{\mathrm{M}} = \frac{1}{4} + \frac{5k}{12} - \frac{7k^2}{8} > 0, \qquad \mathbb{E}R_2^{\mathrm{F}} - \mathbb{E}R_2^{\mathrm{M}} = (q_{\mathrm{F}} - q_{\mathrm{M}})(\delta - m^2). \tag{29}$$

Thus the optimistic variant is better for $\delta < m^2$, tied at $\delta = m^2$, and worse for $m^2 < \delta < 1/8$. Numerically, $m \approx 8.23 \times 10^{-2}$ and the difference at $\delta = m^2/2$ is approximately -8.56×10^{-4}.

The individual probe probabilities are

$$q_{\mathrm{F}} = \frac{3}{4} - \frac{7k}{12} - \frac{17k^2}{48}, \qquad q_{\mathrm{M}} = \frac{1}{2} - k + \frac{25k^2}{48}, \tag{30}$$

Here $m = \int_{-1}^1 z\sigma(z/2)\, dz$, by symmetry.

Proof of Proposition B.3. The actual rewards are $1/2$ plus the displayed gaps. They lie in $[0, 1]$, the features have norm at most $1/2$, and the rectangular uniform prior is a full-dimensional log-concave density on a compact convex set. The candidate sets are deterministic, so the construction satisfies Assumption 2.1.

For two IID prior draws, put $M_c = \max(c_1, c_2)$. OLE probes exactly when $\max(z_1, z_2) > kM_c$. Since $\mathbb{E}M_c = 7/6$, $\mathbb{E}M_c^2 = 17/12$, and the CDF of $z \sim \mathrm{Unif}[-1, 1]$ is $(1 + t)/2$,

$$q_{\mathrm{F}} = 1 - \tfrac{1}{4}\mathbb{E}(1 + kM_c)^2 = \tfrac{3}{4} - \tfrac{7k}{12} - \tfrac{17k^2}{48}.$$

Mean2 probes exactly when $z_1 + z_2 > k(c_1 + c_2)$. On $t \in [0, 2]$, $\Pr(z_1 + z_2 > t) = (2 - t)^2/8$. Using $\mathbb{E}(c_1 + c_2) = 2$ and $\mathbb{E}(c_1 + c_2)^2 = 25/6$ gives q_{M} in (30). Their difference is the positive polynomial in (29) on the stated interval.

The first action is independent of $\theta^\star$. Relative to choosing s, choosing p therefore adds expected round-one regret $\mathbb{E}(\delta c - z/2) = \delta$. A safe query has likelihood depending only on c, so the posterior of z remains uniform and neither round-two rule has positive expected selected gap. After a probe and label one, Bayes' rule gives the normalized density $f_1(z) = \sigma(z/2)$ on $[-1, 1]$. Label zero gives its reflection. The two labels are equiprobable and the corresponding posterior means are m and $-m$.

On the antipodal second-round pair, OLE and Mean2 both select + iff $Z_1 + Z_2 > 0$. For IID $Z_1, Z_2 \sim f_1$,

$$\mathbb{E}\,\mathrm{sign}(Z_1 + Z_2) = 2m. \tag{31}$$

Indeed, write $f_1 = 1/2 + h$ with h odd. In the double integral against $\mathrm{sign}(z_1 + z_2)$, the constant and quadratic terms vanish by central symmetry, while each linear term equals m because $\frac{1}{2}\int_{-1}^{1} \mathrm{sign}(z + y)\, dy = z$. Conditional independence of $z^\star$ and the fresh pair now gives expected selected gap $(m/2)(2m) = m^2$ after either probe label. Averaged over the first-round label, the expected optimal second-round gap is $1/4$ after either round-one action by the tower property, so a probe reduces expected second-round regret by exactly m^2.

Thus, for either policy A, its expected total regret is a common constant plus $q_A(\delta - m^2)$. Subtraction proves (29). □

The next subsection changes the pairing of the reward samples in this same instance and separates its acquisition and terminal-decision effects.

B.5 A COUPLING WITH AN EXACT LEARNING ADVANTAGE

Proposition B.4 (An exact quantile-antithetic posterior pair)**.** *In the two-round instance of Proposition B.3, let $0 < \delta < 1/8$ and $m = \int_0^1 z \tanh(z/4)\, dz$. At each history the exact posterior factorizes as $\Pi_{t-1} = \Pi_{c,t-1} \otimes \Pi_{z,t-1}$. Generate fresh independent $U_c, U_z \sim \mathrm{Unif}(0, 1)$ and use the pair*

$$\theta^{(1)} = \big(F_{c,t}^{-1}(U_c), F_{z,t}^{-1}(U_z)\big), \qquad \theta^{(2)} = \big(F_{c,t}^{-1}(1 - U_c), F_{z,t}^{-1}(1 - U_z)\big),$$

where $F_{c,t}, F_{z,t}$ are the current posterior marginal CDFs. Each member has the exact joint posterior as its marginal law, and the fresh pair is conditionally independent of $\theta^\star$. Compare OLE with coefficient one and population SD (AO) with mean selection using the same pair (AM). Their first-round probe probabilities and total Bayesian regrets satisfy

$$q_{\mathrm{AO}} = 1 - \frac{5}{2}\delta, \qquad q_{\mathrm{AM}} = 0, \qquad \mathbb{E}R_2^{\mathrm{AO}} - \mathbb{E}R_2^{\mathrm{AM}} = \left(1 - \frac{5}{2}\delta\right)\left(\delta - \frac{m}{2}\right).$$

Thus AO is better than AM for $\delta < m/2$, tied at $\delta = m/2$, and worse for $m/2 < \delta < 1/8$.

Proof. The parameter set $[1/2, 3/2] \times [-1, 1]$ is compact, convex, and full-dimensional, and its uniform density is log-concave. Set the anchor reward to $1/2$ and add the gaps δc, $z/2$, or $-z/2$ to obtain the offered rewards. They lie in $[0, 1]$. The corresponding features have Euclidean norm at most $1/2$, and one may take $B = \sqrt{13}/2$, $S = 1/2$, $d = 2$, and $K = 2$. The shortlists are deterministic.

The first-round action depends only on fresh internal randomness and is therefore independent of the true parameter. Given this action, its BTL likelihood depends on c alone for a safe query and on z alone for a probe. Consequently the posterior remains a product. In particular, after a safe query its z marginal remains uniform. After a probe its c marginal remains the prior. Independent coordinate uniforms make each member of the displayed pair have the entire product posterior law.

New uniforms at each round ensure the required conditional independence from $\theta^\star$. Dependence within the pair is intentional.

For the initial pair, write its first member as (C, Z). Its second member is $(2 - C, -Z)$, so the mean scores of the safe response and probe are δ and zero. AM always chooses safe. The OLE scores are $\delta(1 + |C - 1|)$ and $|Z|/2$, respectively. Since $|Z| \sim \mathrm{Unif}[0, 1]$ is independent of C and $2\delta(1 + |C - 1|) < 1$,

$$q_{\mathrm{AO}} = \mathbb{E}\{1 - 2\delta(1 + |C - 1|)\} = 1 - \frac{5}{2}\delta.$$

The first-round OLE tie has probability zero.

After a probe the labels are equiprobable. Their normalized z-posterior densities on $[-1, 1]$ are

$$f_1(z) = \sigma(z/2), \qquad f_0(z) = \sigma(-z/2) = f_1(-z),$$

with means m and $-m$. Denote their CDFs and quantiles by F_b, Q_b. For $-1 < x < 1$,

$$F_1(x) - F_0(x) = \int_{-1}^{x} \tanh(z/4)\, dz = 4 \log \frac{\cosh(x/4)}{\cosh(1/4)} < 0.$$

Both densities are positive and continuous. Strict stochastic dominance, together with reflection, gives for every $0 < u < 1$,

$$Q_1(u) > Q_0(u) = -Q_1(1 - u), \qquad Q_0(u) + Q_0(1 - u) < 0.$$

Thus the sum of an antithetic posterior pair has the sign of $2b - 1$ almost surely after a probe label b.

For any pair (Z_1, Z_2) on the antipodal second-round shortlist, both rules prefer $+$ to $-$ exactly when $Z_1 + Z_2 > 0$: the difference of their mean scores is $(Z_1 + Z_2)/2$, while the difference of their maximum scores is $[\max(Z_1, Z_2) + \min(Z_1, Z_2)]/2$. Consequently either rule chooses the sign indicated by a probe label, and its conditional expected selected gap is $m/2$. After a safe query, the antithetic z pair sums to zero and both rules tie. Any fixed tie rule has expected selected gap zero because the posterior mean of z is zero.

The first action is independent of $\theta^\star$, so each probe reduces the expected first-round selected gap by δ relative to a safe query. It increases the expected second-round selected gap by $m/2$. The expected optimal gaps are common to the policies. Therefore the total expected regret of a policy with first-round probe probability q and this antithetic second-round pair is

$$C_\delta + q(\delta - m/2), \qquad C_\delta = \frac{3}{8} - \frac{\delta}{2} + \frac{13\delta^2}{24}.$$

Here C_δ is the regret of always choosing safe and then an uninformed sign: integrating a uniform z gives $\mathbb{E}\max(\delta c, z/2) = 1/8 + \delta/2 + 13\delta^2/24$, and $\mathbb{E}|z|/2 = 1/4$. Substitution of the two probe probabilities proves the result. □

Corollary B.5 (Direct comparison of the two OLE couplings)**.** *In this same instance, let IF denote OLE with two fresh IID exact posterior draws at both rounds. For every* $0 < \delta < 1/8$*,*

$$\mathbb{E}R_2^{\mathrm{AO}} - \mathbb{E}R_2^{\mathrm{IF}} = q_{\mathrm{AO}}(\delta - m/2) - q_{\mathrm{IF}}(\delta - m^2) < 0,$$
$$q_{\mathrm{IF}} = \frac{3}{4} - \frac{7\delta}{6} - \frac{17\delta^2}{12}.$$

Proof. For IID draws, the first-round calculation in Proposition B.3 gives q_{IF}, and the expected second-round selected gap after a probe is m^2. For completeness, the latter follows because after label one, $f_1(z) = 1/2 + h(z)$ with h odd, and

$$\mathbb{E}\,\mathrm{sign}(Z_1 + Z_2) = 2m, \qquad Z_1, Z_2 \overset{\mathrm{IID}}{\sim} f_1.$$

In the double integral the constant and quadratic terms vanish by central symmetry. Each linear term is m, using $\frac{1}{2}\int_{-1}^{1} \mathrm{sign}(z + y)\, dy = z$. Freshness makes the true z independent of the pair conditional on the observed history, so the expected selected gap is $(m/2)(2m) = m^2$. Reflection handles label zero. A safe query supplies zero selected gap in expectation. Hence IF has regret $C_\delta + q_{\mathrm{IF}}(\delta - m^2)$, giving the displayed exact difference.

The sign admits a uniform rational certificate. The elementary inequalities $x - x^3/3 \le \tanh x \le x$ for $x \ge 0$ imply

$$\frac{79}{960} \le m \le \frac{1}{12}.$$

Moreover $q_{\mathrm{AO}} \ge 11/16$, $0 < q_{\mathrm{IF}} \le 3/4$, and

$$\begin{aligned}\delta(q_{\mathrm{AO}} - q_{\mathrm{IF}}) &= \frac{\delta}{4} - \frac{4\delta^2}{3} + \frac{17\delta^3}{12} \\ &\le \frac{\delta}{4} - \frac{37\delta^2}{32} \le \frac{1}{74}.\end{aligned}$$

It follows that

$$\mathbb{E}R_2^{\mathrm{AO}} - \mathbb{E}R_2^{\mathrm{IF}} \le \frac{1}{74} - \frac{869}{30720} + \frac{1}{192} = -\frac{10873}{1136640} < 0.$$

□

Separating acquisition from terminal precision. Let $\mathcal{R}_{XY}$ denote expected two-round regret when OLE uses coupling X initially and coupling Y at the terminal round, with I denoting IID and A the quantile-antithetic construction. The preceding calculations give $\mathcal{R}_{XY} = C_\delta + q_X(\delta - v_Y)$, where $v_I = m^2$ and $v_A = m/2$. Since $\Delta q = q_A - q_I = 1/4 - 4\delta/3 + 17\delta^2/12 > 0$,

$$\mathcal{R}_{AA} - \mathcal{R}_{II} = \underbrace{\Delta q(\delta - m^2)}_{\text{first-round change}} - \underbrace{q_A(m/2 - m^2)}_{\text{terminal change}}.$$

With a common IID terminal rule, changing only the first-round coupling helps exactly when $\delta < m^2$. With a common antithetic terminal rule, the threshold is $m/2$. Changing only the terminal coupling always helps. Thus the corollary's uniform advantage combines acquisition and terminal precision, rather than a uniformly beneficial acquisition change.

Scope. This is an exact, coordinatewise quantile coupling of factorized posteriors in a two-dimensional, two-round model. Its effect is not confined to cancelling the matched mean selector's first-round probe score: the corollary compares OLE directly under the two couplings. The improvement uses both a changed probe probability and label-consistent second-round decisions. It does not order couplings on other instances or identify an optimal coupling. In particular, this construction is not the controlled study's rejection-based sampler and does not establish an advantage for that sampler, approximate neural posteriors, or a general antithetic procedure.

B.6 Calibration for larger independent ensembles

B.6.1 A BTL boundary posterior requiring logarithmic calibration

Proposition B.6 (Logarithmic conditional calibration is necessary)**.** *There is a one-dimensional bounded linear BTL model with prior $\pi_0(\theta) \propto e^{-\theta^2/2}\mathbf{1}_{[0,1]}(\theta)$, positive-probability histories H_n, and a candidate gap g such that, for every fixed $u \ge 0$,*

$$\lim_{n\to\infty} \Pr\{g - \mu_{n+1} > u s_{n+1} \mid H_n\} = e^{-(u+1)}.$$

Thus a nonnegative multiplier $h(\delta)$ guaranteeing conditional underestimation probability at most δ uniformly over these histories must satisfy $h(\delta) \ge \log(1/\delta) - 1$.

Proof of Proposition B.6. We give the example for every fixed $\rho \ge 0$. $\rho = 1$ gives the strongly convex prior in the statement. Take $d = 1$, $\Theta = [0, 1]$, and prior density proportional to $e^{-\rho\theta^2/2}$ on Θ. Use a fixed context, reference reward $r_{\mathrm{ref}} = 1/2$, and two possible candidates with rewards

$$r_+(\theta) = \tfrac{1}{2} + \tfrac{1}{2}\theta, \qquad r_-(\theta) = \tfrac{1}{2} - \tfrac{1}{2}\theta.$$

All rewards lie in $[0, 1]$, all instantaneous regrets are at most one, and the reference-centered gaps are $g_+(\theta) = \theta/2$ and $g_-(\theta) = -\theta/2$. Thus $B = 1$ and $S = 1/2$ suffice. Query r_+ deterministically

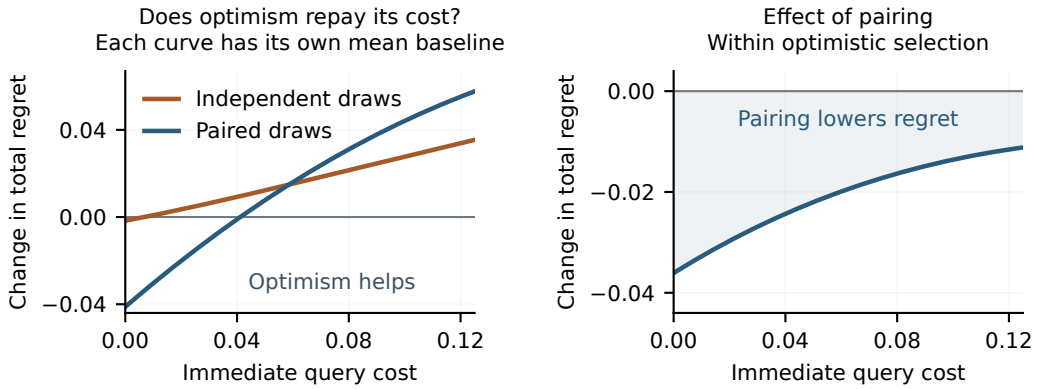

Figure 3: **The value of an informative query depends on how later decisions use it.** Exact two-round model, with lower values favoring the first named rule. Left: each optimistic rule is compared with its own mean variant under the same sampling law. The two curves therefore have different baselines. Optimism helps below zero, where the next-round benefit exceeds the immediate query cost. Right: paired and independent draws are compared within optimistic selection, using the chosen sampling law at both rounds. Pairing lowers total regret throughout the displayed range.

for the first n rounds and let H_n be the history in which all n BTL labels are one. This uses parameter-independent action kernels, and

$$\Pr(H_n) = \mathbb{E}_{\pi_0}\sigma(\theta/2)^n \geq 2^{-n} > 0.$$

The negative candidate is available at the following round. Its posterior moments are μ_{n+1} and s_{n+1} conditional on H_n.

The exact posterior density after H_n is proportional to $e^{-\rho\theta^2/2}\sigma(\theta/2)^n$ on $[0,1]$. Set $\ell(\theta) = \log\sigma(\theta/2)$ and $a = \ell'(1) = \frac{1}{2}\sigma(-1/2) > 0$. Since $\ell'(\theta) = \frac{1}{2}\sigma(-\theta/2) \geq a$ for $0 \leq \theta \leq 1$, integration gives

$$\ell(\theta) - \ell(1) \leq -a(1-\theta). \tag{32}$$

Under this posterior define $Z_n = na(1-\theta)$. Its density on $[0, na]$ is $q_n(z)/\int_0^\infty q_n(v)\,dv$, where

$$q_n(z) = \mathbf{1}_{[0,na]}(z)\exp\left[-\frac{\rho}{2}\left(1-\frac{z}{na}\right)^2 + n\left\{\ell\left(1-\frac{z}{na}\right) - \ell(1)\right\}\right].$$

The Jacobian $1/(na)$ is constant and cancels in normalization. For each fixed $z \geq 0$, differentiability of ℓ at one yields $n[\ell(1-z/(na)) - \ell(1)] \to -z$, hence $q_n(z) \to e^{-\rho/2}e^{-z}$. Since $\rho \geq 0$, (32) gives $0 \leq q_n(z) \leq e^{-z}$. Dominated convergence, also applied with weights z and z^2, shows

$$\int q_n \to e^{-\rho/2}, \qquad \mathbb{E}[Z_n \mid H_n] \to 1, \qquad \mathbb{E}[Z_n^2 \mid H_n] \to 2.$$

The normalized densities converge in L^1, again by dominated convergence and convergence to a positive normalizing constant. Consequently Z_n converges in distribution to an exponential variable E of mean one, and its variance converges to one. These conclusions include moment convergence, which distributional convergence alone would not establish.

For the negative candidate, $g_-(\theta) = -1/2 + Z_n/(2na)$, so its standardized posterior fluctuation is exactly

$$\frac{g_-(\theta) - \mu_{n+1}}{s_{n+1}} = \frac{Z_n - \mathbb{E}[Z_n \mid H_n]}{\sqrt{\mathrm{Var}(Z_n \mid H_n)}} \implies E - 1.$$

Each finite-n variance is positive, since the posterior density is positive on $(0,1)$. The limiting law has no atoms, so for every fixed $u \geq 0$ its upper-tail probability converges to $\Pr(E-1 > u) = e^{-(u+1)}$. The exact Bayesian identity makes this also the conditional tail of the true parameter's gap.

Finally, fix $\delta \in (0,1)$ and suppose a finite nonnegative multiplier $h(\delta)$ makes the conditional upper-tail probability at most δ for every n in this example. Apply the limit at the fixed value $u = h(\delta)$ to obtain $e^{-(h(\delta)+1)} \le \delta$, or $h(\delta) \ge \log(1/\delta) - 1$. For any finite constant C, the choice $h(\delta) = C\sqrt{\log(1/\delta)}$ violates this inequality for sufficiently small δ. The strict limiting violation then holds at some finite positive-probability history. The order of quantifiers is important: δ is fixed before $n \to \infty$. The histories may be rare, and this argument proves neither an unconditional coverage lower bound nor a regret lower bound. □

For the maximizing-action variant, define $R_T = \sum_{t=1}^{T}[\max_y g_{\theta^\star}(x_t, y) - g_{\theta^\star}(x_t, y_t)]$. The reference cancels from this comparator. The following results analyze the exact-moment and general independent-ensemble versions of OLE. Their simultaneous calibration requirement is separate from the two-particle expected-optimism argument in Theorem 3.1.

Theorem B.7 (Exact-moment optimism)**.** *Under Assumption 2.1, for $\delta \in (0,1)$, set $\kappa = 1 + \log(KT/\delta)$ and choose $y_t \in \arg\max_{y\in\mathcal{Y}_t}[\mu_t(y) + \kappa s_t(y)]$. Then*

$$\mathbb{E}R_T \le \min\{T,\, 2\kappa\sqrt{T\Gamma_T/\lambda} + \delta T\}, \qquad R_T = \sum_{t=1}^{T} r_t. \tag{33}$$

Proof of Theorem B.7. Conditional calibration and a union bound over at most KT candidate-round pairs show that the event

$$\mathcal{E} = \bigcap_{t\le T,\, y\in\mathcal{Y}_t} \{|g_{\theta^\star}(x_t, y) - \mu_t(y)| \le \kappa s_t(y)\}$$

has probability at least $1 - KTe^{1-\kappa} = 1 - \delta$. Adaptivity is handled by taking expectations of each conditional failure probability. Independence across rounds is unnecessary. On $\mathcal{E}$, with y_t^* a maximizing true candidate,

$$g_{\theta^\star}(x_t, y_t^*) \le \mu_t(y_t^*) + \kappa s_t(y_t^*) \le \mu_t(y_t) + \kappa s_t(y_t) \le g_{\theta^\star}(x_t, y_t) + 2\kappa s_t(y_t).$$

Since $0 \le R_T \le T$,

$$\begin{aligned}\mathbb{E}R_T &\le 2\kappa\mathbb{E}\sum_t s_t(y_t) + \delta T \\ &\le 2\kappa\sqrt{T\,\mathbb{E}\sum_t s_t(y_t)^2} + \delta T \le 2\kappa\sqrt{T\Gamma_T/\lambda} + \delta T.\end{aligned}$$

Combining with $R_T \le T$ gives the stated cap. The middle inequality is first pathwise Cauchy–Schwarz and then Jensen. The last invokes both parts of Lemma B.1. □

B.6.2 Relative concentration for exact posterior samples

For a Monte Carlo failure budget $\alpha \in (0,1)$, define

$$\begin{aligned} q &= \log(8KT/\alpha), \qquad R = 1 + \log(2NKT/\alpha), \\ a_N &= \sqrt{2q/N} + 4Rq/(3N) + e^{1-R}, \\ b_N &= \sqrt{48eq/N} + 2R^2q/(3N) + 2(R+1)e^{1-R}, \\ \epsilon_N &= \max\{a_N, b_N + a_N^2\}. \end{aligned} \tag{34}$$

Lemma B.8 (Explicit finite-ensemble moment concentration)**.** *At each round draw $N \ge 1$ fresh conditionally independent samples from Π_{t-1} after fixing $\mathcal{G}_t$. With the constants in* (34)*, the event*

$$|\widehat{\mu}_t(y) - \mu_t(y)| \le \epsilon_N s_t(y), \qquad |\widehat{s}_t(y) - s_t(y)| \le \epsilon_N s_t(y) \tag{35}$$

holds simultaneously for all candidates and rounds with probability at least $1-\alpha$. The assertion also holds for auxiliary exact samples drawn along the histories of an approximate-sampling algorithm.

Proof. Fix a pre-round history and a candidate with $s = s_t(y) > 0$, and put $Z_i = (g_{\theta_{t,i}}(x_t, y) - \mu_t(y))/s$. These are independent with $\mathbb{E}Z_i = 0$, $\mathbb{E}Z_i^2 = 1$, and $\Pr(|Z_i| > u) \le e^{1-u}$ for $u > 1$. For $0 \le u \le 1$ the same inequality follows from the trivial upper bound one. The tail integral identity consequently gives

$$\mathbb{E}Z_i^4 = \int_0^\infty 4u^3 \Pr(|Z_i| > u)\, du \le 24e. \tag{36}$$

Let $Y_i = \max(-R, \min(Z_i, R))$. Integrating the tail also gives

$$|\mathbb{E}Y_i| \le \mathbb{E}|Z_i - Y_i| \le e^{1-R},$$
$$0 \le 1 - \mathbb{E}Y_i^2 \le \mathbb{E}(Z_i^2 - Y_i^2) \le 2(R+1)e^{1-R}. \tag{37}$$

Moreover $\mathrm{Var}(Y_i) \le 1$, $|Y_i - \mathbb{E}Y_i| \le 2R$, $\mathrm{Var}(Y_i^2) \le 24e$, and $|Y_i^2 - \mathbb{E}Y_i^2| \le R^2$.

We use the following elementary Bernstein consequence: for independent centered W_i with $|W_i| \le c$ and $N^{-1}\sum_i \mathbb{E}W_i^2 \le v$, with probability at least $1 - 2e^{-q}$,

$$\left| N^{-1} \sum_i W_i \right| \le \sqrt{2vq/N} + 2cq/(3N).$$

To verify it, expand the exponential and use $|\mathbb{E}W_i^k| \le c^{k-2}\mathbb{E}W_i^2$ and $k! \ge 2\,3^{k-2}$ for $k \ge 2$. For $0 < h < 3/c$ this bounds $\log \mathbb{E}e^{h\sum_i W_i}$ by $Nvh^2/[2(1 - hc/3)]$. Chernoff optimization gives the upper tail $\exp\{-Nz^2/[2(v + cz/3)]\}$ at mean deviation z. Substitution of $z = \sqrt{2vq/N} + 2cq/(3N)$ makes its exponent at most $-q$. Apply the same argument to $-W_i$ and union bound.

Applying this inequality to Y_i and Y_i^2, then using (37), bounds the absolute sample mean by a_N and the deviation of the sample second moment from one by b_N, except on an event of probability $4e^{-q}$. The clipping changes no observation except on an event of probability at most Ne^{1-R}. The definitions give $KT\,4e^{-q} = \alpha/2$ and $KT\,Ne^{1-R} = \alpha/2$. A conditional union bound and expectation over adaptive histories therefore make these conclusions simultaneous with probability at least $1 - \alpha$.

On that event write $\bar{Z} = N^{-1}\sum_i Z_i$ and $v_N = N^{-1}\sum_i Z_i^2 - \bar{Z}^2 \ge 0$. Then

$$|\bar{Z}| \le a_N, \qquad |v_N - 1| \le b_N + a_N^2, \qquad |\sqrt{v_N} - 1| = \frac{|v_N - 1|}{\sqrt{v_N} + 1} \le b_N + a_N^2.$$

The empirical mean error divided by s is $\bar{Z}$, and the empirical standard deviation divided by s is $\sqrt{v_N}$. This proves the result. When $s = 0$, all finitely many exact draws equal the constant gap almost surely, so both errors vanish. The proof used only the conditional law of each exact ensemble. It applies unchanged to auxiliary ensembles along approximate histories.

To justify the stated sufficient order, let $h = \log(KT/\alpha) \to \infty$ and take $N = \lceil Ch^3 \rceil$ for a fixed sufficiently large constant C. Then $q = O(h)$, $R = O(h + \log N) = O(h)$, $\sqrt{q/N} = O(C^{-1/2}h^{-1})$, and $R^2 q/N = O(C^{-1})$. The remaining terms vanish or are bounded by a constant multiple of C^{-1}. Choosing C for any desired fixed positive accuracy proves sufficiency for all large h. No necessity or guarantee for small N is asserted. □

B.6.3 Inflation and approximate-sample coupling

Lemma B.9 (Empirical index stability)**.** *For paired points $X_1, \dots, X_N, Y_1, \dots, Y_N \in \Theta$, put $D_N = (N^{-1}\sum_i \|X_i - Y_i\|^2)^{1/2}$. For any S-Lipschitz gap function and empirical standard deviations with denominator N,*

$$|\widehat{\mu}_X - \widehat{\mu}_Y| \le SD_N, \qquad |\widehat{s}_X - \widehat{s}_Y| \le SD_N.$$

The same bounds hold simultaneously over all such gap functions. Hence their mean-plus-c-standard-deviation indices differ by at most $S(1 + c)D_N$ for $c \ge 0$.

Proof. For the mean use Cauchy–Schwarz and $|g(X_i) - g(Y_i)| \le S\|X_i - Y_i\|$. For the standard deviation, let $P = I - N^{-1}\mathbf{1}\mathbf{1}^\top$ be the orthogonal centering projection on $\mathbb{R}^N$. For score vectors $v_X = (g(X_i))_i$ and $v_Y = (g(Y_i))_i$,

$$|\widehat{s}_X - \widehat{s}_Y| = N^{-1/2}|\|Pv_X\|_2 - \|Pv_Y\|_2| \le N^{-1/2}\|P(v_X - v_Y)\|_2 \le SD_N.$$

This argument includes $N = 1$ and zero sample variance. The index statement is the triangle inequality. □

Theorem B.10 (Finite-ensemble regret and approximate-law transfer)**.** *Under Assumption 2.1, fix* $\delta \in (0,1)$. *Set* $\alpha = \delta/2$, $\kappa = 1 + \log(2KT/\delta)$, *and suppose* $\epsilon_N < 1$. *Use the inflated coefficient*

$$\widehat{\kappa} = (\kappa + \epsilon_N)/(1 - \epsilon_N) \tag{38}$$

and choose

$$y_t \in \arg\max_{y \in \mathcal{Y}_t} [\widehat{\mu}_t(y) + \widehat{\kappa}\widehat{s}_t(y)].$$

At each round, conditional on the previous transcript and current candidates, draw N *fresh independent samples from a common law* ν_t *on* Θ, *using randomness independent of* $\theta^\star$. *Let* $e_t = W_2(\nu_t, \Pi_{t-1})$. *Then*

$$\mathbb{E}R_T \le \min\left\{ T,\ 2\widehat{\kappa}\sqrt{T\Gamma_T/\lambda} + \delta T + 2S(1+\widehat{\kappa}) \sum_{t=1}^{T} \mathbb{E}e_t \right\}. \tag{39}$$

In particular, exact independent posterior draws have $e_t = 0$.

Proof of Theorem B.10. At each realized $\mathcal{G}_t$, choose an optimal quadratic transport coupling of ν_t and Π_{t-1}, and draw N independent pairs $(X_{t,i}, Y_{t,i})$ from it, independently of $\theta^\star$ given $\mathcal{G}_t$. Compactness ensures existence of an optimal coupling. Measurable choices can equivalently be replaced by successively near-optimal measurable couplings and a limit. The X-particles have the actual algorithm's law and determine its action. The Y-particles form an auxiliary exact independent posterior ensemble. Record these auxiliary variables in the analytical transcript. Their kernel remains parameter-independent and conveys no additional information conditional on the past and the actual sampler randomness. In particular all posterior and information conclusions above remain valid.

For $D_{t,N} = (N^{-1}\sum_i \|X_{t,i} - Y_{t,i}\|^2)^{1/2}$, conditional Jensen gives the dimension-independent *coupling* bound

$$\mathbb{E}[D_{t,N} \mid \mathcal{G}_t] \le (\mathbb{E}[D_{t,N}^2 \mid \mathcal{G}_t])^{1/2} = e_t. \tag{40}$$

It is not a claim about the distance between an empirical measure and its population law.

Let $U_t^Y(y)$ and $U_t^X(y)$ be the indices with coefficient $\widehat{\kappa}$ from the two ensembles. Define $A_t^{\mathrm{err}} = S(1+\widehat{\kappa})D_{t,N}$. Lemma B.9 bounds $\sup_y |U_t^X(y) - U_t^Y(y)|$ by A_t^{err} on every path. Calibration with failure budget $\delta/2$ and Lemma B.8 with $\alpha = \delta/2$ give a joint event $\mathcal{E}$ of probability at least $1-\delta$. On this event, for every candidate,

$$\begin{aligned} U_t^Y(y) &\ge \mu_t(y) - \epsilon_N s_t(y) + \widehat{\kappa}(1-\epsilon_N)s_t(y) = \mu_t(y) + \kappa s_t(y) \ge g_{\theta^\star}(x_t, y), \\ U_t^Y(y) - g_{\theta^\star}(x_t, y) &\le [\kappa + \epsilon_N + \widehat{\kappa}(1+\epsilon_N)]s_t(y) = 2\widehat{\kappa}s_t(y). \end{aligned}$$

The equality in the second line uses the exact inflation formula. Since the actual action maximizes U_t^X,

$$\begin{aligned} g_{\theta^\star}(x_t, y_t^*) &\le U_t^Y(y_t^*) \le U_t^X(y_t^*) + A_t^{\mathrm{err}} \le U_t^X(y_t) + A_t^{\mathrm{err}} \\ &\le U_t^Y(y_t) + 2A_t^{\mathrm{err}} \le g_{\theta^\star}(x_t, y_t) + 2\widehat{\kappa}s_t(y_t) + 2A_t^{\mathrm{err}}. \end{aligned}$$

Take expectations, bound regret on $\mathcal{E}^c$ by T, and use (40). Lemma B.1 applies to the action actually selected by X, because current sampling is parameter-independent given $\mathcal{G}_t$. Cauchy–Schwarz and Jensen then yield (39). If $\nu_t = \Pi_{t-1}$, the identity coupling has $D_{t,N} = 0$ and gives the exact finite-ensemble result. The cap again follows from $R_T \le T$. □

Numerical scope of the sufficient calibration. For $K = 20$, $T = 2000$, $\delta = 0.05$, and $\alpha = 2.5 \times 10^{-2}$, (34) gives $q \approx 16.36$. At $N = 20$, $R \approx 18.97$, $a_N \approx 21.98$, $b_N \approx 206.73$, and $\epsilon_N \approx 689.86$. The base multiplier is $\kappa \approx 15.29$, but the requirement $\epsilon_N < 1$ fails, so the inflation formula supplies no valid coefficient. Even $N = 10^4$ gives $\epsilon_N \approx 1.17$, whereas $N = 10^5$ gives $\epsilon_N \approx 0.23$ and $\widehat{\kappa} \approx 20.13$. These are evaluations of a conservative sufficient inequality, not lower bounds on the required ensemble size or estimates of actual moment error. They supply no moment-calibration guarantee at $N = 20$. The two-particle theorem uses a different argument and has no such relative-moment prerequisite.

B.7 A FINITE PROJECTED-LANGEVIN BUDGET

The two-particle theorem prices a fresh marginal approximate law. The calibrated general-ensemble theorem (Theorem B.10) additionally requires independent particles. This appendix supplies a concrete sufficient budget on a stronger prior subclass. It keeps the posterior fixed during every inner chain and does not analyze reused particles. Projected Langevin sampling on compact supports was studied by Bubeck et al. (2018). We use the explicit reflected-diffusion discretization estimate of Lehec (2023, Theorem 1). The resulting constants are conservative and do not certify the empirical-head or fixed-law neural implementations.

B.7.1 POSTERIOR GEOMETRY AND THE ACTUAL CHAIN

Assumption B.11 (Smooth strongly log-concave prior and interior start)**.** In addition to Assumption 2.1, the prior density is proportional to $\exp(-V_0(\theta))\mathbf{1}_\Theta(\theta)$. The function V_0 is convex and twice continuously differentiable on a neighborhood of Θ, and, on Θ,

$$\rho I \preceq \nabla^2 V_0(\theta) \preceq M_0 I, \qquad \|\nabla V_0(\theta)\|_2 \le G_0, \qquad \rho > 0.$$

At round t each fresh chain starts at the same $\mathcal{G}_t$-measurable point $x_{t,0} \in \Theta$ with $\operatorname{dist}(x_{t,0}, \partial\Theta) \ge r_0 > 0$. At every inner step the joint innovation is fresh conditional on the past. Each chain's Gaussian innovation and uniform-with-replacement minibatch have their stated laws and are independent of one another. Across chains they may be coupled while preserving these marginal transitions. For the general-ensemble calibrated regret result only, use independent chains.

The bounds below also hold with certified history-dependent $r_{0,t} > 0$ substituted for r_0. A shared deterministic start is compatible with independent chains. Sharing minibatches can preserve single-chain laws, but it does not supply an independent ensemble. Let $D = \operatorname{diam}(\Theta) > 0$, $n_t = t - 1$, and

$$U_{t-1}(\theta) = V_0(\theta) + \sum_{i<t} \ell_i(\theta), \qquad \ell_i(\theta) = A(\psi_i^\top \theta) - b_i \psi_i^\top \theta. \tag{41}$$

No division by n_t or by a loss-range constant is made. Direct differentiation gives

$$\begin{aligned} \nabla \ell_i(\theta) &= [\sigma(\psi_i^\top \theta) - b_i]\psi_i, & \|\nabla \ell_i(\theta)\| &\le S, \\ \nabla^2 \ell_i(\theta) &= \sigma'(\psi_i^\top \theta)\psi_i \psi_i^\top, & 0 \preceq \nabla^2 \ell_i(\theta) &\preceq (S^2/4) I. \end{aligned} \tag{42}$$

Consequently U_{t-1} is ρ-strongly convex, is L_t-Lipschitz on Θ, and has Hessian bounded above by $M_t I$, where

$$L_t = G_0 + n_t S, \qquad M_t = M_0 + n_t S^2/4. \tag{43}$$

For $n_t > 0$, independently sample indices $I_{j,1}, \dots, I_{j,m_t}$ uniformly with replacement from $\{1, \dots, n_t\}$ at inner step j and use

$$\begin{aligned} \widehat{\nabla} U_{t-1}(x) &= \nabla V_0(x) + \frac{n_t}{m_t} \sum_{a=1}^{m_t} \nabla \ell_{I_{j,a}}(x), \\ x_{j+1} &= \operatorname{proj}_\Theta\Big(x_j - \eta_t \widehat{\nabla} U_{t-1}(x_j) + \sqrt{2\eta_t}\, \xi_{j+1} \Big), \qquad \xi_{j+1} \sim \mathcal{N}(0, I_d). \end{aligned} \tag{44}$$

When $n_t = 0$, the estimator is the exact prior gradient. Exact full-gradient LMC is a separate option replacing the estimator by ∇U_{t-1}. Using a random with-replacement batch of size n_t is not the same option. Write ν_t for the law of x_{J_t} conditional on $\mathcal{G}_t$.

Theorem B.12 (Explicit law error for fresh projected chains)**.** *Under Assumption B.11, let $J_t \ge 1$, $m_t \ge 1$, and $0 < \eta_t < d/(2L_t^2)$ with $\eta_t \le 1/M_t$. Define*

$$A_t(J) = (2\sqrt{e} + 1)\frac{(1 + L_t D/d)\sqrt{d + 2\log J}}{r_0} + \frac{7 L_t}{6\sqrt{d}}. \tag{45}$$

Write $q_t = 1 - \rho\eta_t \in [0, 1)$ and

$$G_t(J) = \sum_{j=0}^{J-1} q_t^{2j} = \frac{1 - q_t^{2J}}{1 - q_t^2} \le \min\left\{ J, \frac{1}{\rho\eta_t(2 - \rho\eta_t)} \right\}. \tag{46}$$

Then, almost surely with respect to the pre-round history,

$$W_2(\nu_t, \Pi_{t-1}) \le De^{-\rho J_t \eta_t} + \sqrt{dA_t(J_t)J_t(2\eta_t)^{3/2}} + \frac{\eta_t n_t S}{\sqrt{m_t}}\sqrt{G_t(J_t)}. \tag{47}$$

For exact full gradients, remove the final term. The entire right-hand side can be replaced by its minimum with D, since both laws are supported on Θ. The last term is bounded by both $\eta_t n_t S\sqrt{J_t/m_t}$ and $n_t S\sqrt{\eta_t/[m_t\rho(2-\rho\eta_t)]}$: strong convexity prevents the coupled stochastic-gradient error from accumulating indefinitely with the number of steps.

Proof. Fix the complete pre-round history. In this proof abbreviate $U = U_{t-1}$, $L = L_t$, $M = M_t$, $\eta = \eta_t$, $J = J_t$, and $n = n_t$. All comparisons concern this same fixed potential.

First compare the full-gradient projected chain to the reflected diffusion. We state the imported estimate precisely to identify the step convention and boundary dependence. For a log-concave density proportional to e^{-U} on a convex set with nonempty interior, with U Lipschitz there with constant L, let

$$dZ_u = dB_u - \tfrac{1}{2}\nabla U(Z_u)\,du - d\Phi_u, \quad Z_0 = x_0,$$

where Φ is the outward normal reflection term. Its projected Euler chain is $z_{j+1} = \mathrm{proj}_\Theta(z_j - \frac{h}{2}\nabla U(z_j) + \sqrt{h}\,\xi_{j+1})$. For $h < d/L^2$, Lehec (2023, Theorem 1) gives

$$W_2^2(\mathrm{Law}(z_J), \mathrm{Law}(Z_{Jh})) \le d\left[(2\sqrt{e}+1)\frac{(1+\sigma_0)\sqrt{d+2\log J}}{r(x_0)} + \frac{7L}{6\sqrt{d}}\right] Jh^{3/2}, \tag{48}$$

where $r(x_0) = \mathrm{dist}(x_0, \partial\Theta)$ and $\sigma_0 = [U(x_0) - \inf_\Theta U]/d$. The assumptions here ensure $r(x_0) \ge r_0$ and $\sigma_0 \le LD/d$. Take $h = 2\eta$. This is exactly the noise and drift normalization in (44) with full gradients. The bound (48) is at most $dA_t(J)J(2\eta)^{3/2}$.

Next we give the diffusion mixing argument. No log-Sobolev constant is assumed. Rescaling time, let $X_u = Z_{2u}$, so

$$dX_u = \sqrt{2}\,dB_u - \nabla U(X_u)\,du - d\Phi_u.$$

Its invariant law is Π_{t-1}. Indeed its generator is $\Delta - \langle\nabla U, \nabla\cdot\rangle$ with reflecting Neumann boundary, and integration by parts against e^{-U} makes the generator symmetric and its integral zero. Existence and invariance for the compact convex reflecting diffusion are also part of the reflected process construction used in (48). Couple $X_0 = x_0$ with $Y_0 \sim \Pi_{t-1}$ and drive the two reflecting diffusions by the same Brownian motion. Their difference has no Brownian part. If Φ^X and Φ^Y are their reflection terms, convexity of Θ implies

$$\langle X_u - Y_u, d\Phi_u^X\rangle \ge 0, \qquad \langle Y_u - X_u, d\Phi_u^Y\rangle \ge 0.$$

These are inequalities of finite-variation measures: an outward normal at a boundary point has nonnegative inner product with the vector from any point in Θ to that boundary point. Strong convexity therefore gives

$$\begin{aligned} d\|X_u - Y_u\|^2 &= -2\langle X_u - Y_u, \nabla U(X_u) - \nabla U(Y_u)\rangle\,du \\ &\quad - 2\langle X_u - Y_u, d\Phi_u^X\rangle - 2\langle Y_u - X_u, d\Phi_u^Y\rangle \\ &\le -2\rho\|X_u - Y_u\|^2\,du. \end{aligned}$$

Multiplication by $e^{2\rho u}$ or the finite-variation Gronwall inequality gives $\|X_u - Y_u\| \le e^{-\rho u}\|x_0 - Y_0\|$. Since Y is stationary and both initial points are in a set of diameter D,

$$W_2(\mathrm{Law}(X_{J\eta}), \Pi_{t-1}) \le De^{-\rho J\eta}. \tag{49}$$

This establishes the first two terms of (47) by the Wasserstein triangle inequality.

Finally compare the minibatch chain x_j to a full-gradient chain z_j from the same point and with the same Gaussian increments. Given the past of both chains, the minibatch error $\zeta_j = \widehat{\nabla}U(x_j) - \nabla U(x_j)$ satisfies

$$\mathbb{E}[\zeta_j \mid \text{past}] = 0, \qquad \mathbb{E}[\|\zeta_j\|^2 \mid \text{past}] \le n^2S^2/m_t. \tag{50}$$

To see the second inequality, the m_t independent centered summands have zero cross terms, and a uniformly sampled likelihood gradient has second moment at most S^2. The exact prior gradient contributes no noise. The formula is also valid for $n = 0$, with zero right-hand side.

For $x, z \in \Theta$ let $H_{x,z} = \int_0^1 \nabla^2 U(z + v(x-z))\, dv$. It is symmetric with spectrum in $[\rho, M]$, and

$$x - z - \eta(\nabla U(x) - \nabla U(z)) = (I - \eta H_{x,z})(x - z).$$

For $\eta \le 1/M$ its norm is at most $q\|x - z\|$, where $q = 1 - \rho\eta \in [0, 1)$. Using nonexpansiveness of Euclidean projection, the common Gaussian increment cancels before projection, giving

$$\|x_{j+1} - z_{j+1}\|^2 \le \|x_j - z_j - \eta(\nabla U(x_j) - \nabla U(z_j)) - \eta\zeta_j\|^2.$$

Take conditional expectation before drawing the new minibatch. The cross term with ζ_j vanishes by (50). Thus

$$\mathbb{E}\|x_{j+1} - z_{j+1}\|^2 \le q^2 \mathbb{E}\|x_j - z_j\|^2 + \eta^2 n^2 S^2 / m_t.$$

The shared start makes the initial difference zero. Iterating and taking square roots proves $W_2(\mathrm{Law}(x_J), \mathrm{Law}(z_J)) \le (\eta n S/\sqrt{m_t})\sqrt{\sum_{j=0}^{J-1} q^{2j}}$. Combining the three comparisons proves (47). The diameter cap follows by coupling any two laws supported on Θ. □

B.7.2 An explicit, noncircular budget

Corollary B.13 (A closed sufficient choice of inner steps)**.** *Under Assumption B.11, fix any roundwise target $0 < \varepsilon_t < D$ and any integer minibatch size $m_t \ge 1$. Let*

$$\begin{aligned}
&\tau_t = \rho^{-1} \log(3D/\varepsilon_t),\\
&C_t = (2\sqrt{e} + 1)\frac{1 + L_t D/d}{r_0}, \qquad \overline{A}_t = C_t(\sqrt{d} + 2) + \frac{7L_t}{6\sqrt{d}},\\
&J_t = 1 + \left\lceil \max\left\{1, \frac{2\tau_t L_t^2}{d}, \tau_t M_t, \left(\frac{9d\overline{A}_t(2\tau_t)^{3/2}}{\varepsilon_t^2}\right)^{8/3}, \frac{9\tau_t^2 n_t^2 S^2}{m_t \varepsilon_t^2}\right\}\right\rceil, \qquad \eta_t = \frac{\tau_t}{J_t}.
\end{aligned} \tag{51}$$

Then $W_2(\nu_t, \Pi_{t-1}) \le \varepsilon_t$. For full gradients, omit the last entry in the maximum and evaluate all n_t likelihood gradients at each step. For independent particles satisfying the conditions of Theorem B.10, including $\epsilon_N < 1$ and its inflated bonus, inserting these budgets in (39) *gives*

$$\mathbb{E}R_T \le \min\left\{T,\ 2\widehat{\kappa}\sqrt{T\Gamma_T/\lambda} + \delta T + 2S(1 + \widehat{\kappa})\sum_t \varepsilon_t\right\}. \tag{52}$$

Proof. For every $J \ge 1$, $\log J \le (4/e)J^{1/4}$: maximize $ue^{-u/4}$ over $u = \log J \ge 0$. Consequently

$$\sqrt{d + 2\log J} \le \sqrt{d} + \sqrt{2\log J} \le \sqrt{d} + \sqrt{8/e}\, J^{1/8} \le (\sqrt{d} + 2)J^{1/8},$$

and hence $A_t(J) \le \overline{A}_t J^{1/8}$. The additive one in (51) guarantees $\eta_t < d/(2L_t^2)$ and $\eta_t \le 1/M_t$. The first error term in (47) is $De^{-\rho\tau_t} = \varepsilon_t/3$. For the squared second term, substitution of $\eta_t = \tau_t/J_t$ yields

$$dA_t(J_t)J_t(2\eta_t)^{3/2} \le d\overline{A}_t(2\tau_t)^{3/2}J_t^{-3/8} \le \varepsilon_t^2/9.$$

The final term is at most $\tau_t n_t S/\sqrt{J_t m_t} \le \varepsilon_t/3$ by $G_t(J_t) \le J_t$ and the last entry in the maximum. Their sum is at most ε_t. The same argument with the minibatch term removed proves the full-gradient case. The deterministic bound is uniform over histories because L_t, M_t, D, r_0 are certified bounds, so Theorem B.10 gives (52) directly. □

For example, when $S > 0$, choosing

$$\varepsilon_t = \min\{D/2, [2S(1 + \widehat{\kappa})T]^{-1}\}$$

makes the total computational term in (52) at most one. Combined with $\delta = 1/T$ for $T > 1$, this gives $\mathbb{E}R_T \le 2\widehat{\kappa}\sqrt{T\Gamma_T/\lambda} + 2$ with a finite explicit budget. This is a sufficient existence guarantee, not an efficient complexity claim. The extra power $8/3$ in the budget is a deliberate conservative removal of $\log J_t$ from an implicit step-count constraint. Smaller certified budgets may be obtained by checking the original explicit inequality (47). Making η_t small with J_t fixed does not improve its mixing term.

B.7.3 A CHECKABLE BUDGET AND ITS COMPUTATIONAL SCALING

The closed formula above guarantees termination without an inner search. A deterministic integer search gives a smaller sufficient order by retaining the logarithm in $A_t(J)$. For the same target and τ_t, define

$$J_{t,0} = 1 + \left\lceil \max\left\{1, \frac{2\tau_t L_t^2}{d}, \tau_t M_t, \frac{9\tau_t^2 n_t^2 S^2}{m_t \varepsilon_t^2}\right\}\right\rceil. \tag{53}$$

For full gradients omit the final entry. Starting at $J = J_{t,0}$, double J until

$$dA_t(J)(2\tau_t)^{3/2} J^{-1/2} \le \varepsilon_t^2/9, \qquad \eta_t = \tau_t/J. \tag{54}$$

This procedure uses only the certified geometry constants and target. It does not run trial chains or estimate Wasserstein distance. The initial lower bounds ensure both step-size conditions and a stochastic-gradient term at most $\varepsilon_t/3$ throughout the search. The mixing term is $\varepsilon_t/3$, and (54) bounds the squared discretization term by $\varepsilon_t^2/9$. Thus the resulting chain has $W_2(\nu_t, \Pi_{t-1}) \le \varepsilon_t$. The search terminates because $A_t(J) = O(\sqrt{\log J})$ at a fixed round.

To state the resulting work explicitly, hold $d, D, r_0, \rho, M_0, G_0, S > 0$ fixed and consider $\varepsilon_t \downarrow 0$. Then $A_t(J) = O((1+t)\sqrt{\log(eJ)})$, so squaring the check shows that

$$J \ge C(1+t)^2 \tau_t^3 \varepsilon_t^{-4} \log(eJ)$$

is sufficient for a constant C depending only on the fixed geometry. For $H = C(1+t)^2\tau_t^3\varepsilon_t^{-4}$, increasing C if necessary, an order-$H\log(e+H)$ choice satisfies this inequality. The doubling search overshoots a sufficient threshold by at most a factor two, unless its initial step count already exceeds that threshold. For integer minibatch sizes $m_t \ge 1$, the initial count in (53) is no larger in polynomial order as $\varepsilon_t \downarrow 0$. Consequently the procedure supplies

$$J_t = \widetilde{O}((1+t)^2 \varepsilon_t^{-4}). \tag{55}$$

The tilde hides logarithms in t and $1/\varepsilon_t$, not geometry constants. This is a sufficient upper bound for this fresh-chain construction, not a sampling lower bound or a practical efficiency claim.

For comparison, the closed expression (51) supplies $J_t = \widetilde{O}((1+t)^{8/3}\varepsilon_t^{-16/3})$ under the same fixed-geometry convention. For the maximizing-action regret in (39), take $\delta = 1/T$ with $T > 1$ and K fixed or polynomial in T. Choose relative calibration $\epsilon_N \le \epsilon_0 < 1$, a polylogarithmic ensemble size, and

$$\varepsilon_t = \min\left\{D/2, \frac{1}{2S(1+\widehat{\kappa})\sqrt{T}}\right\}. \tag{56}$$

The linear computational residual is then at most $\sqrt{T}$. Thus $\varepsilon_t = \widetilde{\Theta}(T^{-1/2})$ at fixed positive geometry, and the search construction gives sufficient total likelihood-gradient work $\widetilde{O}(T^5)$ with a fixed minibatch size, or $\widetilde{O}(T^6)$ with full gradients. The corresponding powers from the closed expression are $T^{19/3}$ and $T^{22/3}$. The geometric-noise refinement improves the certificate's stochastic term but does not remove its dominant reflected-diffusion discretization cost. These high sufficient powers explain why the certificate establishes finite computation without certifying the small budgets used in the experiments.

For the box $\Theta = [-b, b]^d$, prior $V_0(\theta) = \|\theta\|^2/2$, and start $x_{t,0} = 0$, all geometry constants are explicit: $D = 2b\sqrt{d}$, $\rho = M_0 = 1$, $G_0 = b\sqrt{d}$, and $r_0 = b$. A history-dependent starting estimate can instead be projected into $[-(b-r_0), b-r_0]^d$ for a chosen $0 < r_0 < b$. The same certificate applies. The bound on $U_{t-1}(x_{t,0}) - \inf U_{t-1}$ uses $L_t D$, so no numerical posterior minimization is needed.

A numerical sufficient budget, not an experimental law-error estimate. Take the valid one-dimensional subclass $\Theta = [-1, 1]$, $V_0(\theta) = \theta^2/2$, $x_{t,0} = 0$, $S = 1/2$ and $T = 2000$. For example, a reference reward $1/2$ and candidate rewards $1/2 \pm \theta/2$ realize these gap bounds. Then $D = 2$, $r_0 = \rho = M_0 = G_0 = 1$. The two-particle target $\varepsilon = [2\sqrt{2}S\sqrt{T}]^{-1} \approx 1.58 \times 10^{-2}$ gives $\tau = \log(3D/\varepsilon) \approx 5.94$. The exact doubling search in (54) returns, with full gradients, $J_1 \simeq 2.93 \times 10^{16}$ and $J_{2000} \simeq 2.68 \times 10^{22}$ per chain. At the last round its mixing and discretization bounds are 5.27×10^{-3} and 4.72×10^{-3}, with no minibatch term. Two chains at that round alone

require about 1.07×10^{26} likelihood-gradient evaluations under this sufficient schedule. These deliberately literal evaluations show the conservatism of the certificate, rather than a recommended implementation or a computational lower bound. The empirical-head and fixed-law neural studies do not establish this appendix's geometry or conditional e_t. Their inference budgets cannot be substituted into this certificate.

B.7.4 What remains conditional outside this subclass

Theorem B.10 needs only Assumption 2.1, whereas Theorem B.12 and Corollary B.13 need Assumption B.11. For a merely log-concave prior, the independent-ensemble transfer remains valid with the explicit residual $2S(1+\widehat{\kappa})\sum_t \mathbb{E}W_2(\nu_t, \Pi_{t-1})$. The present paper supplies no universal replacement of this residual by a function of (J_t, η_t, m_t) for that broader class. Uniform priors, for example, satisfy the statistical assumption but have $\rho = 0$, so the strong-convexity contraction argument above cannot be invoked. One-step reused populations and nonconvex neural parameters remain outside this sampler certificate. Shared minibatches do not supply independent particles for the general-N calibrated result. The two-particle theorem permits them when the fresh marginal transitions specified above are preserved.

C Coupling constructions and selector precision

C.1 Coupling as a selector-precision design variable

Structured dependence can reduce the leading error of a fixed reward selector. We hold the reward law, candidates, base policy and temperature fixed and derive the relevant covariance criterion. This is a selector-emulation result, separate from the sequential regret guarantee.

C.1.1 A selector-sensitive coefficient for dependent blocks

Let $\mathcal{Y}$ be a finite candidate list, let $X \sim \nu$ and let $G(X) = (g_y(X) : y \in \mathcal{Y})$ be a bounded reward vector. Write $\mu_y = \mathbb{E}g_y(X)$ and $s_y^2 = \mathrm{Var}(g_y(X))$. Fix $a \geq 0$, $\beta > 0$, a full-support base distribution p, and $c(u) = \max\{-1, \min\{1, u\}\}$. We allow either the identity map $h(u) = u$ or the clipped map $h(u) = c(u)$. Put

$$u_y = \mu_y + a s_y, \qquad v_y = h(u_y), \qquad \pi = q_v, \qquad q_v(y) = \frac{p(y)e^{v_y/\beta}}{\sum_z p(z)e^{v_z/\beta}}. \tag{57}$$

For clipping, assume $u_y \notin \{-1, 1\}$ whenever $s_y > 0$. For identity scores no clipping-margin assumption is needed. With $Z_y = g_y - \mu_y$, define the centered score influence

$$\xi_y(X) = d_y \left[Z_y + \frac{a}{2s_y}(Z_y^2 - s_y^2) \right], \qquad d_y = \begin{cases} 1, & h(u) = u, \\ \mathbf{1}_{\{|u_y|<1\}}, & h(u) = c, \end{cases} \tag{58}$$

and set $\xi_y = 0$ when $s_y = 0$. In that case $g_y = \mu_y$ almost surely, so the exact score is $h(\mu_y)$ and has no sampling error. Let $H = \mathrm{diag}(\pi) - \pi\pi^\top$ and $\mathcal{A}(\nu) = \mathrm{tr}(H\mathbb{E}[\xi\xi^\top])/(2\beta^2)$.

Consider iid blocks $(X_{i1}, \ldots, X_{ib})$ of size $b \geq 2$, $i = 1, \ldots, m$, where each member has marginal law ν but the members in a block may be dependent. Pool the $N = bm$ rewards and compute the original empirical mean and standard deviation, with either denominator N or $N - 1$. Thus, writing $\widehat{\mu}_y$ and $\widehat{s}_y$ for these pooled estimates, the implemented score is explicitly

$$\widehat{v}_y = h(\widehat{\mu}_y + a\widehat{s}_y), \qquad q_{\widehat{v}_N}(y) = \frac{p(y)e^{\widehat{v}_y/\beta}}{\sum_z p(z)e^{\widehat{v}_z/\beta}}. \tag{59}$$

Define

$$\bar{\xi}_i = \frac{1}{b}\sum_{j=1}^{b} \xi(X_{ij}), \qquad \Omega_\gamma = b\,\mathrm{Cov}(\bar{\xi}_i), \qquad \mathcal{A}_\gamma = \frac{\mathrm{tr}(H\Omega_\gamma)}{2\beta^2}, \tag{60}$$

where γ denotes the selected block coupling.

Proposition C.1 (Coupling-sensitive fixed-law precision)**.** *Under the fixed-law assumptions above, as $m \to \infty$ with fixed b (equivalently, $N = bm \to \infty$ through multiples of b),*

$$\mathbb{E}\,\mathrm{KL}(q_{\widehat{v}_N} \| q_v) = \frac{\mathcal{A}_\gamma}{N} + O(N^{-3/2}). \tag{61}$$

The remainder depends on the fixed law, the block coupling, the finite candidate list, a, β, and the positive standard deviations. When $h = c$, it also depends on the clipping margins. It is not uniform as any of these regularity quantities degenerates. Moreover,

$$0 \le \mathcal{A}_\gamma \le b\mathcal{A}(\nu). \tag{62}$$

For $b = 2$, writing $C = \mathbb{E}[\xi(X_{i1})\xi(X_{i2})^\top]$ gives

$$\mathcal{A}_\gamma = \mathcal{A}(\nu) + \frac{\mathrm{tr}(H\,\mathrm{Sym}\,C)}{2\beta^2}. \tag{63}$$

Thus a pair reduces the leading selector error exactly when $\mathrm{tr}(H\,\mathrm{Sym}\,C) < 0$.

Proof. For a candidate with $s_y > 0$, the pooled empirical variance is $s_y^2 + B_{N,y} - A_{N,y}^2$, where $A_{N,y}$ and $B_{N,y}$ are the pooled empirical means of Z_y and $Z_y^2 - s_y^2$. The block averages of these bounded centered variables are iid across blocks. A Taylor expansion on the event that the variance remains at least $s_y^2/2$, together with bounded-variable concentration on its complement, gives the vector score expansion

$$\widehat{v}_N - v = \frac{1}{m}\sum_{i=1}^{m} \bar{\xi}_i + R_N, \qquad \mathbb{E}\|R_N\|_2^2 = O(N^{-2}), \qquad \mathbb{E}\|\widehat{v}_N - v\|_2^3 = O(N^{-3/2}).$$

To make the remainder orders explicit, the fourth moments of the centered bounded block averages give $\mathbb{E}A_{N,y}^4 + \mathbb{E}B_{N,y}^4 = O(m^{-2})$. On $|A_{N,y}| \le s_y/2$ and $|B_{N,y}| \le s_y^2/4$, the variance is at least $s_y^2/2$, and Taylor's formula yields

$$\widehat{s}_y - s_y - \frac{B_{N,y}}{2s_y} = O(A_{N,y}^2 + B_{N,y}^2).$$

The complement has exponentially small probability by concentration of bounded independent block averages. All empirical scores remain bounded, so this complement contributes a smaller order. The fourth-moment bounds also imply third absolute moments $O(m^{-3/2})$ by Hölder's inequality. Since b is fixed, these are the displayed orders in $N = bm$. The same first-order expansion holds for the correction-one standard deviation because the factor $\sqrt{N/(N-1)}$ changes the score by $O(N^{-1})$. For $h = c$, the strict clipping margin keeps the empirical score on the population branch except on an exponentially small event. For h equal to the identity, this branch argument is unnecessary and $d_y = 1$ directly. The zero-variance case is exact by the convention above.

For a finite Gibbs selector,

$$\mathrm{KL}(q_{v+h} \| q_v) = \frac{h^\top H h}{2\beta^2} + O(\|h\|_2^3).$$

Writing $L_N = m^{-1}\sum_i \bar{\xi}_i$, Cauchy–Schwarz gives

$$|\mathbb{E}L_N^\top H R_N| \le \|H\|\sqrt{\mathbb{E}\|L_N\|^2\mathbb{E}\|R_N\|^2} = O(N^{-3/2}), \qquad \mathbb{E}R_N^\top H R_N = O(N^{-2}).$$

Since $\mathrm{Cov}(m^{-1}\sum_i \bar{\xi}_i) = \Omega_\gamma/N$, taking expectations proves (61). For $W_{ij} = H^{1/2}\xi(X_{ij})$, Jensen's inequality gives

$$\mathbb{E}\left\| b^{-1}\sum_j W_{ij} \right\|_2^2 \le \mathbb{E}\|W_{i1}\|_2^2,$$

which proves (62). When $b = 2$, $\Omega_\gamma = \mathbb{E}[\xi(X)\xi(X)^\top] + \mathrm{Sym}\,C$, giving (63). □

Regularized selection objective. For any probability vector ρ, define $J_v(\rho) = \sum_y \rho(y)v_y - \beta\,\mathrm{KL}(\rho\|p)$, with $0\log 0 = 0$. The full-support Gibbs rule satisfies $\log(q_v(y)/p(y)) = v_y/\beta - F(v)$, where $F(v) = \log\sum_y p(y)e^{v_y/\beta}$. Substitution gives

$$J_v(\rho) = \beta F(v) - \beta\,\mathrm{KL}(\rho\|q_v).$$

Thus q_v maximizes J_v and $J_v(q_v) - J_v(q_{\widehat{v}_N}) = \beta\,\mathrm{KL}(q_{\widehat{v}_N}\|q_v)$. The preceding expansion therefore quantifies the leading expected loss in this fixed regularized objective. It does not identify that score with a true downstream reward.

Forward accuracy and repeat disagreement. The leading coefficient also describes the forward-KL metric used in the neural experiments. Let $F(v) = \log\sum_y p(y)e^{v_y/\beta}$. Then $\nabla^2 F(v) = H/\beta^2$ and

$$\mathrm{KL}(q_v\|q_{v+h}) = F(v+h) - F(v) - \nabla F(v)^\top h = \frac{h^\top H h}{2\beta^2} + O(\|h\|^3).$$

The moment bounds in the proof therefore give $\mathbb{E}\,\mathrm{KL}(q_v\|q_{\widehat{v}_N}) = \mathcal{A}_\gamma/N + O(N^{-3/2})$. For two independent constructions $\widehat{v}_N$ and $\widetilde{v}_N$ under the same law, expand the same Bregman divergence in both arguments. Its quadratic term is $(\widehat{v}_N - \widetilde{v}_N)^\top H(\widehat{v}_N - \widetilde{v}_N)/(2\beta^2)$. The first-order centered influences are independent and the score-remainder bounds above still apply. Consequently

$$\mathbb{E}\,\mathrm{KL}(q_{\widehat{v}_N}\|q_{\widetilde{v}_N}) = \frac{2\mathcal{A}_\gamma}{N} + O(N^{-3/2}).$$

Thus forward target accuracy and independent-repeat disagreement share a leading coupling criterion. Neither formula removes finite-target error or gives a uniform finite-N certificate.

Candidate-contrast interpretation. For any covariance matrix Ω,

$$\mathrm{tr}(H\Omega) = \tfrac{1}{2}\sum_{y,z}\pi(y)\pi(z)(\Omega_{yy} + \Omega_{zz} - 2\Omega_{yz}).$$

Consequently

$$\mathcal{A}_\gamma = \frac{b}{4\beta^2}\sum_{y,z}\pi(y)\pi(z)\,\mathrm{Var}(\bar{\xi}_y - \bar{\xi}_z). \tag{64}$$

Common shifts cancel. Relative-score noise matters in proportion to the corresponding selection probabilities.

C.1.2 Balanced masks and the interaction spectrum

Suppose a finite reward network is driven by L independent dropout coordinates. Let $D_\ell \sim \mathrm{Bernoulli}(1/b)$ and, for each block, draw J_ℓ uniformly from $\{1,\dots,b\}$ and set $D_{j\ell} = \mathbf{1}\{J_\ell = j\}$. Each member has the original product-Bernoulli marginal law. Write $\chi_\ell = (D_\ell - 1/b)/\sqrt{(1/b)(1-1/b)}$ and expand the centered influence in the orthonormal product basis as

$$\xi(D) = \sum_{\varnothing\neq S\subseteq[L]} b_S\chi_S(D), \qquad \chi_S = \prod_{\ell\in S}\chi_\ell.$$

Define the selector-weighted order energies

$$\mathcal{A}_k = \frac{1}{2\beta^2}\sum_{|S|=k} b_S^\top H b_S, \qquad \mathcal{A}_{\mathrm{iid}} = \sum_{k=1}^{L}\mathcal{A}_k.$$

Theorem C.2 (Balanced-mask selector spectrum)**.** *For the balanced blocks above, the leading coefficient in Proposition C.1 is*

$$\mathcal{A}_b = \sum_{k=1}^{L}\left[1 + \frac{(-1)^k}{(b-1)^{k-1}}\right]\mathcal{A}_k, \qquad 0 \le \mathcal{A}_b \le \frac{b}{b-1}\mathcal{A}_{\mathrm{iid}}. \tag{65}$$

In particular, first-order influence is removed, while second-order influence is multiplied by $b/(b-1)$. The construction is therefore beneficial precisely when the signed sum in (65) *is smaller than $\mathcal{A}_{\mathrm{iid}}$. Balanced dependence is not universally beneficial.*

Proof. For two distinct members of a block, the standardized coordinate correlation is $\rho = -1/(b-1)$. Independence across coordinates and orthogonality of the product basis give

$$\mathbb{E}[\chi_S(D_j)\chi_T(D_{j'})] = \mathbf{1}_{\{S=T\}}\rho^{|S|}.$$

Substituting this identity into $b\,\mathrm{Cov}(b^{-1}\sum_j \xi(D_j))$ gives the multiplier $1 + (b-1)\rho^k$ for order k, which is the displayed expression. All multipliers are nonnegative and their maximum is $1+1/(b-1)$, proving the upper bound. At $b = 10$, the first three weights are 0, $10/9$ and $80/81$. □

C.1.3 A SAFE FINITE-FAMILY COUPLING CHOICE

The coefficient gives a decision-relevant selection rule when a small finite family of couplings is available. Let $\mathcal{G}$ be such a family, with each member preserving the same marginal reward law. Use an independent pilot to obtain estimates $\widehat{\mathcal{A}}_\gamma$ and radii $r_\gamma \geq 0$ satisfying, simultaneously with probability at least $1 - \delta$,

$$\left|\widehat{\mathcal{A}}_\gamma - \mathcal{A}_\gamma\right| \leq r_\gamma \qquad (\gamma \in \mathcal{G}). \tag{66}$$

Choose an independent production coupling by

$$\widehat{\gamma} \in \arg\min_{\gamma\in\mathcal{G}}\{\widehat{\mathcal{A}}_\gamma + r_\gamma\}. \tag{67}$$

Lemma C.3 (Finite-family oracle inequality)**.** *On the event in* (66)*,*

$$\mathcal{A}_{\widehat{\gamma}} \leq \min_{\gamma\in\mathcal{G}}\{\mathcal{A}_\gamma + 2r_\gamma\}. \tag{68}$$

The bound is for the leading fixed-law coefficient. Pilot evaluations and production evaluations must both be charged when comparing total computation.

Proof. For any $\gamma \in \mathcal{G}$, the simultaneous event and the choice rule give

$$\mathcal{A}_{\widehat{\gamma}} \leq \widehat{\mathcal{A}}_{\widehat{\gamma}} + r_{\widehat{\gamma}} \leq \widehat{\mathcal{A}}_\gamma + r_\gamma \leq \mathcal{A}_\gamma + 2r_\gamma.$$

Taking the minimum proves the claim. □

The oracle inequality is conditional on simultaneous coefficient bounds from an independent pilot. Our experiments use a fixed balanced coupling rather than estimate these bounds or select among couplings. The expansion uses the usual smooth delta method (van der Vaart, 1998). The balanced construction is related to classical stratified sampling (Stein, 1987) and scrambled sampling (Owen, 1997).

C.2 ZERO-SUM SGLD INNOVATIONS

A zero-sum Gaussian coupling removes the injected-noise component of the ensemble mean while preserving every innovation's Gaussian marginal. For $N \geq 2$, let $Z_i \overset{\text{iid}}{\sim} \mathcal{N}(0, I_d)$, $\bar{Z} = N^{-1}\sum_i Z_i$, and $P_N = I_N - \mathbf{1}\mathbf{1}^\mathsf{T}/N$. Define

$$\widetilde{Z} = \sqrt{\frac{N}{N-1}}P_N Z, \qquad \widetilde{Z}_i = \sqrt{\frac{N}{N-1}}(Z_i - \bar{Z}).$$

The full-coupling implementation applies independent copies of this transform to initialization and each SGLD innovation. The noise-only implementation keeps independent initialization.

Proposition C.4 (Marginal innovation law and negative covariance)**.** *Each $\widetilde{Z}_i \sim \mathcal{N}(0, I_d)$, and for $i \neq j$,*

$$\mathrm{Cov}(\widetilde{Z}_i, \widetilde{Z}_j) = -\frac{I_d}{N-1}, \qquad \sum_i \widetilde{Z}_i = 0.$$

Proof. The transform is Gaussian. Since $P_N = P_N^\mathsf{T} = P_N^2$,

$$\mathrm{Cov}(\widetilde{Z}) = \left(\frac{N}{N-1}I_N - \frac{\mathbf{1}\mathbf{1}^\mathsf{T}}{N-1}\right) \otimes I_d.$$

Its diagonal blocks equal I_d, its off-diagonal blocks equal $-I_d/(N-1)$, and $\mathbf{1}^\mathsf{T}P_N = 0$. □

Proposition C.5 (Conditional elimination of mean diffusion)**.** *Let $\mathcal{F}_t$ contain the particle and feedback history before an update. If $b_{t,i}$ and scalar σ_t are $\mathcal{F}_t$-measurable and $\widetilde{Z}_t$ is formed from a fresh Gaussian vector independent of $\mathcal{F}_t$, then the update $\theta_{t+1,i} = b_{t,i}(\mathcal{F}_t) + \sigma_t \widetilde{Z}_{t,i}$ preserves each conditional innovation marginal and satisfies*

$$\frac{1}{N}\sum_i \theta_{t+1,i} = \frac{1}{N}\sum_i b_{t,i}(\mathcal{F}_t).$$

Likewise, $d\widetilde{W}_t = \sqrt{N/(N-1)}P_N\, dW_t$ gives zero Brownian diffusion in the ensemble mean.

Proof. Condition on $\mathcal{F}_t$ and apply Proposition C.4. Summation removes the noise. The diffusion statement follows from $\mathbf{1}^\mathsf{T} P_N = 0$. □

Scope of the preservation. These are conditional innovation identities. Coupling changes the selected actions and subsequent feedback, so it need not preserve the state marginal of an IID SGLD particle or the IID interpretation of its empirical SD. The fixed-law trace criterion in Proposition C.1 therefore does not certify this dynamic sampler's regret.

D Controlled sequential preference learning

D.1 Comparison methods and computational accounting

The common OLE interface uses a mean-plus-SD score with explicitly specified inference. The exact-pair experiment has population SD and coefficient one. The neural fixed-law study uses sample SD and coefficient two. The latter is a selector-precision experiment, not an implementation of the regret theorem.

Posterior-sampling controls. Exact Thompson sampling selects the largest gap under the first exact current-history draw (Russo & Van Roy, 2014). The common evaluation procedure obtains two accepted draws for every arm, including Fisher-UCB. Its timing therefore includes sampling work that their native implementations do not require.

Curvature controls. In the exact-pair study, local Fisher-UCB maximizes $x_y^\top \widehat{\theta}_t + \sqrt{x_y^\top V_t x_y}$, where $\widehat{\theta}_t$ is the box MAP and V_t its inverse observed Hessian. Generalized-linear bandits (Filippi et al., 2010) provide the broader family, and logistic bandits (Faury et al., 2020) the specific feedback setting. The local radius and implementation are defined here.

Ablations and work. Mean2 isolates the optimistic score while keeping the same pair law and draw budget. Every learner receives feedback for its own chosen comparison. IID and balanced dropout instead compare Monte Carlo implementations of the same fixed neural selector at matched projection work.

D.2 Closed-loop acquisition with exact posterior pairs

This study tests the exact two-sample score in a setting with both sequential feedback and known regret. Frozen Qwen2.5-Instruct features (1.5B, 3B and 7B) define comparisons, and a bounded logistic teacher generates labels, following the controlled-generative principle of Osband et al. (2022). The learned posterior head has two dimensions. Encoder sizes therefore probe fixed feature maps, rather than full-network posterior inference.

Candidates and bounded features. Programs come from the PPE MBPP-Plus Best-of-K release (Frick et al., 2024), whose candidate panel was enriched using natural correctness. The experiment neither executes programs nor uses their correctness labels. Of 208 designated tasks, 32 fit the projection, 48/16 support development acquisition/evaluation, and 80/32 are reserved for acquisition/evaluation. A fixed hash selects one generator row per task. Exact Python-AST duplicates are collapsed, syntactically invalid programs remain eligible, and each row requires a valid reference

and two distinct alternatives. Complete prompt, separator and program tokens are encoded without truncation, with BF16 inference and FP32 mean pooling over the program span. All arms reuse the resulting representations.

Let W be the whitened leading two eigenvectors of the uncentered second moment of anchor-relative features on the 32 fitting tasks, with signs fixed by the largest loading. For reference a and alternative y, set

$$z_{ay} = 4W^\top(h_y - h_a), \qquad x_{ay} = \frac{z_{ay}}{2\max\{1, \|z_{ay}\|_2\}}. \tag{69}$$

Four was the smallest scale in $\{1, 2, 4\}$ meeting feature-support requirements before teacher feedback. With $r_\theta(a) = 1/2$ and $r_\theta(y) = 1/2 + x_{ay}^\top\theta$, every feasible reward is in $[0, 1]$ on $\Theta = [-1/\sqrt{2}, 1/\sqrt{2}]^2$.

Informative-challenger geometry. For each axis $j \in \{1, 2\}$ and sign $s \in \{-1, 1\}$, the prior is a box-truncated Gaussian with mean $0.4s$ on axis j and precision 96 on j and 2 on the other axis k. Distinct safe and informative challengers satisfy

$$|x_{\mathrm{safe},k}| \le 0.1, \quad sx_{\mathrm{safe},j} \ge 0.15, \quad |x_{\mathrm{probe},k}| \ge 0.3, \quad 5 \times 10^{-3} \le s(x_{\mathrm{safe},j} - x_{\mathrm{probe},j}) \le 0.06. \tag{70}$$

Among eligible triplets, maximize $|x_{\mathrm{probe},k}| - |x_{\mathrm{safe},k}|$, with deterministic index ties. This explicitly conditions on a small immediate cost and an informative alternative. Antipodal alternatives would make the score ablation inactive: $\max(a, b) - \max(-a, -b) = a + b$, so both scores choose the same action.

Every eligible context is retained and repeatedly visited in shuffled cycles. Counts in axis/sign order $(1, -), (1, +), (2, -), (2, +)$ are $43, 23, 33, 32$ for 1.5B, $4, 23, 19, 16$ for 3B, and $33, 16, 39, 39$ for 7B. Their union contains 66 of the 80 reserved tasks, with two tasks in every cell. Repeat labels are independently noisy, and neither the reference nor a duplicate is selectable. Inference is over new Bayesian episodes conditional on this finite geometry.

Exact sampling under either pair law. Each episode draws $\theta_\star$ from the prior and reveals $b_t \sim \mathrm{Bernoulli}(\sigma(x_t^\top\theta_\star))$ after selection. Exact rejection sampling then draws a fresh pair from that arm's current posterior. If U is its negative log density and Q the prior precision, the uniform Hessian lower bound is

$$H = Q + \sum_j \sigma(L_j)\sigma(-L_j)x_j x_j^\top, \qquad L_j = \|x_j\|_2.$$

At the constrained mode c, the quadratic $q(\theta) = U(c) + \nabla U(c)^\top(\theta - c) + (\theta - c)^\top H(\theta - c)/2$ satisfies $q \le U$. A Gaussian proposal proportional to e^{-q} accepts in-box points with probability e^{q-U}. Alternatively, a uniform-box proposal uses the exact box minimum of q for its rejection envelope. The tighter envelope is chosen before sampling. A 100,000-proposal cap aborts without fallback and is never reached. Exactness is mathematical, with a checked floating-point envelope tolerance in the implementation.

IID streams give independent draws. Coupled streams use opposite Gaussian or centered-box proposals and complementary acceptance uniforms. Each member returns its own first accepted proposal, without discarding or restarting its partner. Consequently both marginals are the ordinary rejection law, including for asymmetric posteriors. A 120,000-pair check compares each marginal with quadrature for symmetric, skewed and boundary-concentrated fixtures. This instantiates Theorem 3.1 with $d = K = 2$, $B = 1$, $S = 0.5$, $T = 512$ and exact-arithmetic $e_t = 0$. Its conservative ceiling is approximately $366.56 < 512$.

Arms, endpoints and inference. OLE maximizes the two sampled gaps and its matched control maximizes their mean. The population-SD and sample-SD/$\sqrt{2}$ identities agree below 10^{-12}. Each arm maintains its own feedback history. Exact Thompson sampling (Russo & Van Roy, 2014) and coefficient-one MAP/Hessian Fisher-UCB (Faury et al., 2020) are additional controls. All arms receive 512 labels. The common evaluation procedure obtains two accepted draws even for these controls. Its timing therefore includes sampling work that their native implementations do not require.

Query regret uses the best feasible challenger. On the 32 held-out tasks, all nonreference representatives enter evaluation against a hash-fixed reference. The final posterior predictive $p_T(x)$ is assessed

by $-p_\star(x)\log p_T(x) - (1-p_\star(x))\log(1-p_T(x))$, with equal task weights and $p_\star(x) = \sigma(x^\top\theta_\star)$. Secondary utility shortfall evaluates the posterior-mean choice against the best true utility.

A 64-episode development set supplies the covariance estimate used to select 512 evaluation episodes from $\{128, 256, 512, 1024\}$ before their feature execution or teacher generation. Teacher, context and potential-feedback streams are paired across arms. Effects first average equally over three encoders and four orientations. The four primary contrasts compare OLE with its matched mean rule on regret and predictive loss under each law. One-sided Bonferroni bounds control nominal familywise error at 0.01. A law succeeds only if both the Student-t and 100,000-resample paired percentile-bootstrap upper bounds are negative for both endpoints. There is no interim continuation or minimum-effect threshold.

Table 3: **Controlled acquisition with 1.5B–7B language-model features.** Both rules use coupled fresh pairs, a 2D learned head and 512 labels. Means average 512 paired episodes and four prior orientations under synthetic logistic feedback. GPU-h sums shared extraction-worker times. Peak denotes allocated CUDA memory. Full baseline outcomes and paired intervals are in Tables 4–6.

Frozen	Query regret ↓		Excess loss (10^{-4}) ↓		Feature extraction	
encoder	OLE (mean)	**OLE (optimistic)**	OLE (mean)	**OLE (optimistic)**	Worker GPU-h	Peak GiB
1.5B	19.07	**17.12**	7.27	**4.76**	0.03	3.07
3B	20.11	**17.54**	7.59	**4.72**	0.04	6.08
7B	20.30	**17.63**	7.57	**5.22**	0.07	14.60

Learning effect and controls. Coupled OLE lowers regret by 2.40 (12.09%) and predictive loss by 2.58×10^{-4}, passing both corrected endpoint tests. The respective Student-t/bootstrap upper bounds are $-1.46/-1.46$ and $-2.20/-2.21 \times 10^{-4}$. The loss reduction is 0.04% of control loss and 34.5% of its excess above teacher entropy. IID OLE improves predictive loss (-1.09×10^{-4}, adjusted upper bound -8.14×10^{-5}), but its regret difference -0.17 has upper bound 0.34. Thus the coupled-law conjunction succeeds, while the all-four-test conjunction does not.

Fisher-UCB (Faury et al., 2020) attains regret 14.30, coupled OLE 17.43, and Thompson sampling (Russo & Van Roy, 2014) 20.21. Secondary coupled-OLE contrasts against the independent mean and IID OLE are -0.96 $[-1.40, -0.52]$ and -0.79 $[-1.06, -0.53]$, with unadjusted 95% intervals. The paired score-by-law interaction is -2.23 $[-2.75, -1.71]$ for regret and -1.49 $[-1.78, -1.20] \times 10^{-4}$ for loss. It includes a coupled-minus-IID change of $+1.44$ $[0.97, 1.90]$ for the mean control, so it does not imply that coupling improves every score. The 42.94% utility-shortfall reduction is secondary. Table 7 gives checkpoint and encoder sensitivity without asserting a scaling law.

Learning curves and encoder-specific outcomes. Figure 4 traces the optimistic-minus-mean contrast under each sampling law. Under coupled sampling, optimistic selection reduces regret by 10.2%, 12.8% and 13.1% at 1.5B, 3B and 7B, respectively, with improved prediction at every size (Table 3). These sizes vary the frozen representation, while only the two-dimensional head learns. The pooled reduction of 2.40 over 512 queries corresponds to 4.68×10^{-3} per query. The four primary tests use zero-effect nulls without a prespecified minimum effect size. Against Fisher-UCB (Faury et al., 2020), the optimistic variant improves both loss and shortfall at 3B, while Fisher-UCB has lower shortfall at 7B.

Numerical accuracy and extraction cost. Order-64/96 quadrature differs by at most 5.56×10^{-16}. Extraction uses 2,035 representations and 808,174 tokens per encoder, totaling 6,105 representations and 2,424,522 tokens. Total GPU allocation is 0.39 GPU-hours, with 0.14 GPU-hours of feature-extraction work including model loading and verification. The latter excludes process startup and allocation overhead. All acquisition arms share these features. Table 3 reports worker times and peak allocated CUDA memory by encoder. The geometry was selected during development and fixed before evaluation. Experimental acquisition and analysis use no program-correctness labels.

Scope of the evaluation. Near-tied candidate choices can depend on the numerical-library version. Full reproduction requires encoder inference and projection fitting in addition to the acquisition procedure. The 208-task cohort lacks a fully specified construction rule and a verified selection date.

Table 4: Closed-loop OLE across three frozen encoders, with a two-dimensional learned head and synthetic logistic feedback. All methods acquire 512 labels. Pointwise 95% Student-t intervals use 512 paired episodes after averaging four orientations. Encoder-specific comparisons are descriptive. Excess NLL subtracts known teacher entropy and is descriptive. Encoders are fixed feature maps, not full-network posterior models. Lower is better for every metric. Citations identify the TS and logistic-UCB families. The local implementations are defined in Appendix D.1.

Encoder	Method	Query regret	Excess NLL (10^{-4})	Utility shortfall (10^{-3})
1.5B	OLE (mean), independent	18.31 [17.60, 19.01]	6.26 [5.88, 6.64]	14.09 [13.01, 15.18]
	OLE (optimistic), independent	18.15 [17.53, 18.77]	5.17 [4.86, 5.49]	11.32 [10.38, 12.26]
	OLE (mean), coupled	19.07 [18.30, 19.84]	7.27 [6.86, 7.68]	17.28 [16.02, 18.54]
	OLE (optimistic), coupled	17.12 [16.49, 17.74]	4.76 [4.48, 5.04]	10.09 [9.28, 10.90]
	Exact TS (Russo & Van Roy, 2014)	19.98 [19.35, 20.61]	5.91 [5.57, 6.25]	13.14 [12.15, 14.13]
	Local Fisher-UCB (Faury et al., 2020)	14.14 [13.49, 14.80]	4.91 [4.59, 5.22]	10.83 [9.88, 11.79]
3B	OLE (mean), independent	18.12 [17.47, 18.76]	5.96 [5.58, 6.34]	13.44 [12.24, 14.64]
	OLE (optimistic), independent	18.11 [17.48, 18.74]	4.94 [4.62, 5.25]	10.49 [9.55, 11.44]
	OLE (mean), coupled	20.11 [19.27, 20.95]	7.59 [7.11, 8.07]	18.68 [17.12, 20.23]
	OLE (optimistic), coupled	17.54 [16.90, 18.18]	4.72 [4.41, 5.02]	9.83 [8.90, 10.77]
	Exact TS (Russo & Van Roy, 2014)	20.01 [19.39, 20.63]	5.95 [5.58, 6.33]	13.30 [12.16, 14.44]
	Local Fisher-UCB (Faury et al., 2020)	14.28 [13.56, 15.01]	5.23 [4.87, 5.60]	12.30 [11.03, 13.58]
7B	OLE (mean), independent	18.75 [18.04, 19.45]	6.53 [6.12, 6.95]	13.78 [12.48, 15.07]
	OLE (optimistic), independent	18.41 [17.76, 19.06]	5.37 [5.04, 5.69]	10.63 [9.61, 11.66]
	OLE (mean), coupled	20.30 [19.47, 21.13]	7.57 [7.10, 8.04]	17.70 [16.16, 19.23]
	OLE (optimistic), coupled	17.63 [16.98, 18.28]	5.22 [4.91, 5.54]	10.69 [9.66, 11.72]
	Exact TS (Russo & Van Roy, 2014)	20.63 [19.96, 21.31]	6.35 [5.97, 6.74]	13.16 [11.95, 14.37]
	Local Fisher-UCB (Faury et al., 2020)	14.46 [13.79, 15.12]	5.02 [4.71, 5.33]	9.58 [8.59, 10.56]

Inference therefore conditions on the chosen task cohort and feature geometry, without accounting for uncertainty in task selection.

D.3 Closed-loop replay with human rankings

OpenAssistant (OASST1) supplies human rankings (Köpf et al., 2023). Complete conversation trees are split after excluding exact normalized-context and frozen token-5 Jaccard overlaps with HelpSteer2, RewardUQ and previous OLE prompts. A label-blind hash selects one prompter context, three nonsynthetic English assistant siblings and an anchor. Triplets require three distinct consistent integer ranks. A query reveals whether the chosen alternative outranks the anchor, then updates that arm's posterior. Aggregate ranks are fixed replay labels, without modeling annotator disagreement.

Table 5: **Score effects under independent and coupled posterior pairs.** Differences are OLE (optimistic) minus the matched OLE (mean) variant. U_t and $U_{\text{bootstrap}}$ are one-sided Student-t and paired-bootstrap upper bounds. Bonferroni level is 2.5×10^{-3} per comparison, giving familywise level 0.01. Inference uses 512 paired episodes after equal weighting of the twelve encoder/orientation settings. Joint improvement under one pair law requires negative upper bounds for both endpoints and both estimators.

Pair law	Endpoint	Difference	U_t	$U_{\text{bootstrap}}$
IID	Query regret	-0.17	$+0.34$	$+0.33$
IID	Log loss $\times 10^4$	-1.09	-0.81	-0.82
Coupled	Query regret	-2.40	-1.46	-1.46
Coupled	Log loss $\times 10^4$	-2.58	-2.20	-2.21

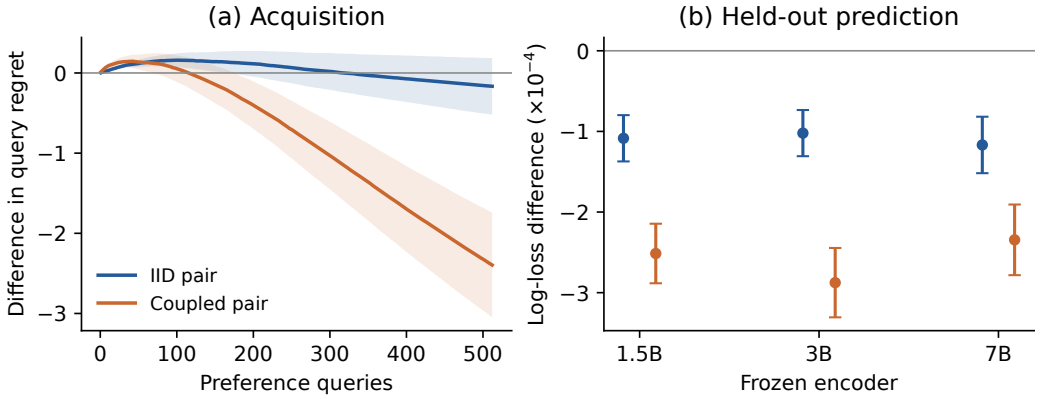

Figure 4: **Sampling and score interact in controlled preference learning.** The optimistic variant minus its matched mean variant under each pair law. Left: cumulative query regret. Right: final preference log loss by frozen encoder. Shading and bars are descriptive 95% paired-episode intervals. Only the 2D head learns, from synthetic logistic feedback.

Qwen3-0.6B response-span features use two axes fit on separate HelpSteer2 data and gap norms capped at $1/2$. The head has support $[-1/\sqrt{2}, 1/\sqrt{2}]^2$, mean $(0.4, 0)$ and precision $\text{diag}(96, 2)$, as in the controlled study. Exact rejection supplies IID or antithetic-stream pairs. Human ranks need not follow this BTL model. No OpenAssistant labels fit the features or candidate sets, although encoder pretraining exposure cannot be excluded.

Design and inference. A 16-episode development set precedes the frozen comparison on 96 disjoint episodes, each with 16 acquisition and eight held-out trees. Development/confirmation contain 384/2,304 trees, with 393 trees reserved. Development admission checks geometry, numerics, prequential prediction and action separation, without selecting on efficacy signs. All arms receive 16 labels. OLE and its two-draw mean control have matched accepted-draw budgets, with one-draw Thompson sampling (Russo & Van Roy, 2014) and local Fisher-UCB (Faury et al., 2020) as additional controls.

The four primary contrasts are OLE minus mean in acquisition mistakes (choosing the worse-ranked alternative) and held-out predictive NLL under each law. Centered studentized paired bootstrap tests use 199,999 replicates and Holm correction at familywise level 0.01. A law must improve both endpoints and reach floors of 0.5 fewer mistakes and 5×10^{-3} lower NLL. At least one law must pass, with no continuation. Power uses 10^5 simulations with effects twice these floors, simultaneous upper SD bounds, and a 0.80 lower Monte Carlo bound for the full decision. Episode inference conditions on the frozen tree sampling frame.

Results. Table 8 reports all primary effects. IID OLE adds 0.42 mistakes (paired SD 1.24) and changes NLL by -2.39×10^{-4} (SD 1.01×10^{-3}). Antithetic differences are $+0.41$ mistakes (SD

Table 6: Descriptive paired contrasts for the coupled optimistic OLE variant across three frozen encoders, with a two-dimensional learned head and synthetic logistic feedback. All methods acquire 512 labels. Pointwise 95% Student-t intervals use 512 paired episodes after averaging four orientations. Encoder-specific comparisons are descriptive. Excess NLL subtracts known teacher entropy and is descriptive. Encoders are fixed feature maps, not full-network posterior models. Lower is better for every metric. Citations identify baseline families, with local implementations in Appendix D.1.

Encoder	Optimistic minus control	Query regret	Excess NLL (10^{-4})	Utility shortfall (10^{-3})
		-1.19	-1.50	-4.00
1.5B	OLE (mean), independent	$[-1.82, -0.56]$	$[-1.86, -1.15]$	$[-5.13, -2.88]$
		-1.03	-0.42	-1.23
	OLE (optimistic), independent	$[-1.45, -0.62]$	$[-0.70, -0.14]$	$[-2.13, -0.34]$
		-1.95	-2.51	-7.19
	OLE (mean), coupled	$[-2.74, -1.16]$	$[-2.88, -2.14]$	$[-8.38, -6.01]$
		-2.86	-1.16	-3.05
	Exact TS (Russo & Van Roy, 2014)	$[-3.44, -2.29]$	$[-1.47, -0.85]$	$[-4.04, -2.07]$
		2.97	-0.15	-0.75
	Local Fisher-UCB (Faury et al., 2020)	$[2.54, 3.41]$	$[-0.41, 0.11]$	$[-1.62, 0.13]$
		-0.58	-1.24	-3.60
3B	OLE (mean), independent	$[-1.21, 0.06]$	$[-1.58, -0.90]$	$[-4.84, -2.37]$
		-0.57	-0.22	-0.66
	OLE (optimistic), independent	$[-0.98, -0.15]$	$[-0.48, 0.04]$	$[-1.63, 0.31]$
		-2.57	-2.88	-8.84
	OLE (mean), coupled	$[-3.53, -1.61]$	$[-3.31, -2.44]$	$[-10.36, -7.33]$
		-2.47	-1.24	-3.46
	Exact TS (Russo & Van Roy, 2014)	$[-3.06, -1.88]$	$[-1.56, -0.92]$	$[-4.56, -2.36]$
		3.26	-0.52	-2.47
	Local Fisher-UCB (Faury et al., 2020)	$[2.71, 3.81]$	$[-0.81, -0.23]$	$[-3.67, -1.27]$
		-1.12	-1.31	-3.09
7B	OLE (mean), independent	$[-1.73, -0.50]$	$[-1.68, -0.94]$	$[-4.37, -1.80]$
		-0.78	-0.14	0.06
	OLE (optimistic), independent	$[-1.22, -0.35]$	$[-0.45, 0.16]$	$[-1.01, 1.12]$
		-2.67	-2.35	-7.01
	OLE (mean), coupled	$[-3.49, -1.84]$	$[-2.78, -1.91]$	$[-8.58, -5.44]$
		-3.00	-1.13	-2.47
	Exact TS (Russo & Van Roy, 2014)	$[-3.59, -2.41]$	$[-1.51, -0.76]$	$[-3.76, -1.18]$
		3.17	0.20	1.11
	Local Fisher-UCB (Faury et al., 2020)	$[2.75, 3.60]$	$[-0.05, 0.46]$	$[0.17, 2.06]$

Table 7: **OLE learning dynamics by encoder.** Optimistic minus mean selection within OLE under coupled sampling: cumulative regret with paired, unadjusted 95% intervals. Each row averages four fixed prior orientations.

Encoder	$t = 64$	$t = 256$	$t = 512$
1.5B	$+0.17$ $[+0.04, +0.30]$	-0.53 $[-0.97, -0.09]$	-1.95 $[-2.74, -1.16]$
3B	$+0.17$ $[+0.02, +0.33]$	-0.66 $[-1.19, -0.14]$	-2.57 $[-3.53, -1.61]$
7B	$+0.04$ $[-0.08, +0.16]$	-1.04 $[-1.48, -0.60]$	-2.67 $[-3.49, -1.84]$

1.37) and -1.57×10^{-4} NLL (SD 1.17×10^{-3}). Neither law meets the joint statistical and practical criterion. Action disagreement is 160/1,536 under IID and 187/1,536 under antithetic sampling, confirming that the scores induce different queries. The maximum two-sample identity error is 5.56×10^{-17}.

OLE makes fewer mistakes than Fisher-UCB (Faury et al., 2020) by -0.50 $[-0.86, -0.14]$ (IID) and -0.48 $[-0.78, -0.18]$ (antithetic), with NLL differences $+3.01 \times 10^{-5}$ $[-2.42, 3.03] \times 10^{-4}$

Table 8: **OpenAssistant human-rank replay over 16 acquisitions.** Each entry pairs the optimistic and mean OLE variants and gives their difference Δ. Lower is better. One-sided studentized-bootstrap p_{H} values are Holm-adjusted over all four contrasts at FWER 0.01. Prespecified effect floors are $\Delta \le -0.5$ mistakes and $\Delta \le -5 \times 10^{-3}$ nats. Neither law passes.

Pair law	Mistakes Opt./mean	Δ	NLL $\times 10^4$ Opt./mean	Δ (nats)	$(p_{\mathrm{H}}^{\mathrm{mis}}, p_{\mathrm{H}}^{\mathrm{NLL}})$
IID	$6.35/5.94$	$+0.42$	$6695.55/6697.94$	-2.39×10^{-4}	$(1.00, 2.32 \times 10^{-2})$
Antithetic	$6.38/5.97$	$+0.41$	$6697.45/6699.03$	-1.57×10^{-4}	$(1.00, 0.24)$

and $+2.21 \times 10^{-4}$ $[-0.76, 5.18] \times 10^{-4}$. These descriptive 95% intervals do not establish predictive noninferiority. Mistake contrasts with one-draw Thompson sampling (Russo & Van Roy, 2014) are inconclusive. This natural-label replay therefore measures an acquisition–prediction tradeoff, without demonstrating the matched-control learning benefit found in the conditioned synthetic regime.

D.4 Correctly specified preferences on the natural-study geometry

The natural-rank study in Appendix D.3 did not establish an OLE benefit. Two possible explanations are the candidate geometry and short query horizon, and misspecification of human preferences by the working Bradley–Terry–Luce (BTL) model. We examine these explanations in two controlled studies with synthetic preferences. Their designs were informed by the preceding studies and fixed before their own outcomes. Neither study is an independent natural-label confirmation.

Matched geometry and correct specification. The first study retains all 96 confirmation geometry blocks, tree order, 16-query horizon, prior, posterior family, and policies of the natural-rank study. It replaces human ranks with labels generated by the learner's own bounded two-dimensional logistic model. The teacher is drawn from the same Gaussian prior, truncated to $[-1/\sqrt{2}, 1/\sqrt{2}]^2$, with mean $(0.4, 0)$ and precision $\mathrm{diag}(96, 2)$. Feature norms are at most 0.5. Four independent teachers per geometry block give 384 episodes, each with 16 acquisition and eight held-out trees. No natural ranks enter this study.

OLE maximizes the larger score from two fresh exact posterior samples. Mean2 maximizes their average under the same IID or antithetic pair law. The methods share law-specific pair streams at coincident histories and potential-outcome uniforms when they choose the same edge. TS1 (Russo & Van Roy, 2014) and Fisher-UCB (Faury et al., 2020) are descriptive controls. Fisher-UCB uses the box-MAP score plus one posterior Hessian standard deviation, with bonus coefficient one.

Endpoints and inference. The primary endpoints are true cumulative query regret and terminal predictive negative log likelihood (NLL) on common synthetic labels for both alternatives of each held-out tree. Teacher replicates are averaged within geometry block, leaving $n = 96$ paired inference units. The four one-sided beneficial-direction tests compare OLE with Mean2 for both endpoints and pair laws. They use a centered studentized paired bootstrap with 199,999 replicates and four-way Holm correction at familywise level 0.01, without a practical-effect floor. The estimand is conditional on these fixed geometry blocks. Two-sided Student-t intervals describe effect size and are not additional corrected tests.

Table 9: **Correctly specified BTL on the frozen natural-study geometry.** Effects are OLE minus Mean2 over 16 acquisitions. Lower is better. Intervals are unadjusted paired Student-t 95% block intervals. Only the one-sided beneficial-direction bootstrap tests receive Holm correction, and their adjusted p-values all equal one.

Pair law	Endpoint	Effect	95% interval
IID	True query regret	$+0.03$	$[+0.01, +0.05]$
IID	Held-out NLL	-3.33×10^{-5}	$[-1.52, 0.86]\times10^{-4}$
Antithetic	True query regret	$+0.07$	$[+0.05, +0.09]$
Antithetic	Held-out NLL	$+7.52\times10^{-5}$	$[-0.67, 2.18]\times10^{-4}$

Result and interpretation. OLE increases mean true regret from 0.28 to 0.31 under IID sampling (10.3%) and from 0.23 to 0.30 under antithetic sampling (30.1%). Both regret intervals lie above zero, while both predictive NLL intervals include zero. Teacher-expected held-out NLL differences are also near zero. The methods disagree on 10.27% of IID decisions and 14.24% of antithetic decisions, so the intervention changes acquisition. OLE has lower descriptive regret point estimates than TS1 (Russo & Van Roy, 2014) and Fisher-UCB (Faury et al., 2020) here, but Mean2 is best.

Thus the adverse direction can arise on the same geometry and 16-query horizon even when the teacher and learner share the BTL model. This does not separate geometry from horizon or exclude an additional role for misspecification in the natural-rank study.

D.5 A PROSPECTIVE GEOMETRY–HORIZON TEST

Question and design. The second study tests whether informative candidate geometry and a longer horizon jointly favor OLE. Its design was informed by the preceding matched study and the controlled learning result, then fixed before generating outcomes. It uses 393 OpenAssistant feature trees unused in the natural-rank study's development and confirmation episodes. Sixteen hash-selected trees are held out for evaluation. Each of 512 paired prior-predictive BTL episodes compares horizons 16 and 128 on recurrent 48-tree acquisition banks. The broad bank is sampled from the remaining 377 trees. The informative bank contains 48 of the 52 trees satisfying a prior-score gap at most 0.05 and an uncertain-axis gap at least 0.15, selected without outcomes. Both geometries use independently permuted 48-round cycles, matching recurrence.

The teacher and learner share the bounded prior and BTL model specified above. OLE and Mean2 use fresh exact posterior pairs after each label, with IID and antithetic laws analyzed separately. Fisher-UCB (Faury et al., 2020) is a descriptive control. The 16-query trajectory is a prefix of its 128-query counterpart. The endpoints are average true query regret and terminal teacher-expected predictive NLL on both alternatives of the 16 common held-out trees.

Primary contrasts. For each pair law and endpoint, the primary family contains the OLE-minus-Mean2 effect in the informative bank at horizon 128 and its geometry-by-horizon interaction,

$$[D_{\mathrm{inf}}(128) - D_{\mathrm{inf}}(16)] - [D_{\mathrm{broad}}(128) - D_{\mathrm{broad}}(16)],$$

where $D_G(H)$ denotes the paired endpoint difference for geometry G and horizon H. For average query regret, the prefix relation gives

$$D_G(128) - D_G(16) = \tfrac{7}{8}\{\overline{d}_{G,17:128} - \overline{d}_{G,1:16}\}.$$

The regret interaction therefore measures a difference in late-versus-early relative performance, not a direct long-horizon benefit. Each episode is one paired inference unit. The eight primary tests use 199,999 centered studentized paired bootstrap replicates, one-sided beneficial alternatives, and Holm correction at familywise level 0.01. A pair law passes only if all four of its effects are negative and rejected. No practical-effect floor was specified.

Table 10: **Prospective moderator family.** Effects are OLE minus Mean2. Lower is better. Intervals are unadjusted paired Student-t 95% intervals. $G\times H$ is the displayed geometry-by-horizon interaction. Effects and interval endpoints are multiplied by 10^4. Only the antithetic regret interaction rejects after eight-way Holm correction.

Law	Endpoint	Contrast	Effect $\times 10^4$ [95% interval]	Raw p	Holm p
IID	Regret	Informative H128	$+3.86\ [-7.77, 15.48]$	0.74	1.00
IID	Regret	$G\times H$	$-8.72\ [-25.85, 8.40]$	0.16	0.47
IID	NLL	Informative H128	$-0.61\ [-1.16, -0.06]$	1.25×10^{-2}	8.76×10^{-2}
IID	NLL	$G\times H$	$-0.76\ [-1.45, -0.08]$	1.25×10^{-2}	8.76×10^{-2}
Antithetic	Regret	Informative H128	$+13.98\ [-2.24, 30.20]$	0.95	1.00
Antithetic	Regret	$G\times H$	$-38.56\ [-61.63, -15.49]$	3.45×10^{-4}	2.76×10^{-3}
Antithetic	NLL	Informative H128	$-0.72\ [-1.38, -0.06]$	1.59×10^{-2}	8.76×10^{-2}
Antithetic	NLL	$G\times H$	$-0.48\ [-1.28, 0.32]$	0.12	0.47

Result and scope. Neither pair law passes the joint criterion. The antithetic regret interaction indicates greater attenuation of OLE's regret cost from early to late rounds in the informative bank.

However, the direct informative-H128 regret effects are $+3.86 \times 10^{-4}$ under IID sampling and $+1.40 \times 10^{-3}$ under antithetic sampling, with intervals crossing zero. No NLL contrast survives familywise correction. OLE and Mean2 disagree on 25.7% of informative-H128 decisions under IID sampling and 33.5% under antithetic sampling. Fisher-UCB (Faury et al., 2020) has lower informative-H128 regret than OLE by 4.52×10^{-3} under IID sampling and 3.42×10^{-3} under antithetic sampling.

The study therefore does not validate a general geometry–horizon explanation or establish joint regret and prediction superiority. Its inference is conditional on the fixed feature banks and synthetic prior-predictive law. The unadjusted direct-effect intervals rule out per-query regret improvements larger than 7.77×10^{-4} under IID sampling and 2.24×10^{-4} under antithetic sampling on this bank, but do not establish zero effects. Together, these controls locate limits of the observed learning benefit without attributing the natural-rank result to a unique cause.

E FIXED-LAW NEURAL SELECTOR STUDY

E.1 PROTOCOL, ESTIMAND AND FIXED REWARD FITS

This study asks whether dependence among a fixed number of stochastic reward evaluations improves implementation of a specified selector. It does not update a reward posterior or acquire preference labels. The estimand is conditional on four fitted RewardUQ MCD-DPO models (Yang et al., 2026): Qwen3-1.7B BF16 seed 42, Qwen3-1.7B FP32 seeds 42 and 43, and Qwen3-4B FP32 seed 42. All use BF16 autocast, one training epoch, DPO coefficient 0.01, dropout 0.1, learning rate 10^{-5}, cosine decay, warmup 0.05 and batch size 64. Thirty exact benchmark-prompt overlaps and 134 overlength training pairs are removed, leaving 61,971 pairs and 969 updates. These four configurations probe fit sensitivity. They are not a random sample of model training runs. Reusing the 4B study's cohort for the three 1.7B fits was chosen after observing the 4B results. The sampling draws, reference draws, settings and analysis for each additional fit were fixed before evaluation. These comparisons therefore assess conditional fit sensitivity on a reused cohort.

The evaluation cohort contains 144 text-hash-selected RewardBench 2 groups (Malik et al., 2026), 24 per domain and 943 complete responses. Every response fits the 2,048-token limit without truncation, and there is no exact overlap with reward-model training. Exact filtering does not exclude semantic or pretraining contamination. Each stochastic reward equals 0.01 times the fitted-minus-reference completion log-probability sum. For response y, the implemented score is

$$\widehat{u}_y = \widehat{\mu}_y + 2\widehat{s}_y, \qquad \widehat{q}(y) \propto \exp(\widehat{u}_y/\beta),$$

where $\widehat{s}_y$ is the sample SD of $N = 20$ dropout rewards. The primary temperature is $\beta = 1$. $\beta = 0.2$ is secondary.

IID dropout draws each mask coordinate independently. Balanced sampling partitions the 20 members into two ten-member blocks. For each coordinate in each block, exactly one uniformly chosen member is dropped. Thus every member-coordinate mask retains the original Bernoulli marginal and usual inverted-dropout scaling. Dropout is applied to the final hidden states before the vocabulary projection, with a separate random coordinate for each token and hidden dimension. Complete-vocabulary log-softmax gives the completion log probabilities used in the reward, without score clipping. The implementation uses complete ten-member blocks and independent candidate streams. Masks are independent across candidates, so the construction is a fixed-selector numerical device rather than a coherent posterior reward function shared by all responses.

Two independent reference banks contain 1,024 IID draws per response each. Pooling them defines $u_y^{\mathrm{ref}} = \bar{r}_y^{\mathrm{ref}} + 2\,\mathrm{SD}(r_y^{\mathrm{ref}})$ and $q^{\mathrm{ref}}(y) \propto \exp(u_y^{\mathrm{ref}}/\beta)$. The primary error is $\mathrm{KL}(q^{\mathrm{ref}} \| \widehat{q})$. Each of 32 paired repetitions averages two independently constructed $N = 20$ selectors and all groups before the ratio of mean errors is formed. Bootstrap resamples preserve the pairing and recompute the ratio within each resample. The 4B comparison uses a 95% interval. The three 1.7B comparisons form a family with 98.33% marginal intervals for nominal 95% Bonferroni coverage. These intervals condition on the pooled finite reference bank as well as the fitted models and cohort. They do not propagate target Monte Carlo error. The experiment measures fidelity to this finite reference target, rather than estimating the asymptotic coefficient $\mathcal{A}_\gamma$ or its interaction spectrum from Appendix C.1.

E.2 SELECTOR FIDELITY AND MATCHED WORK

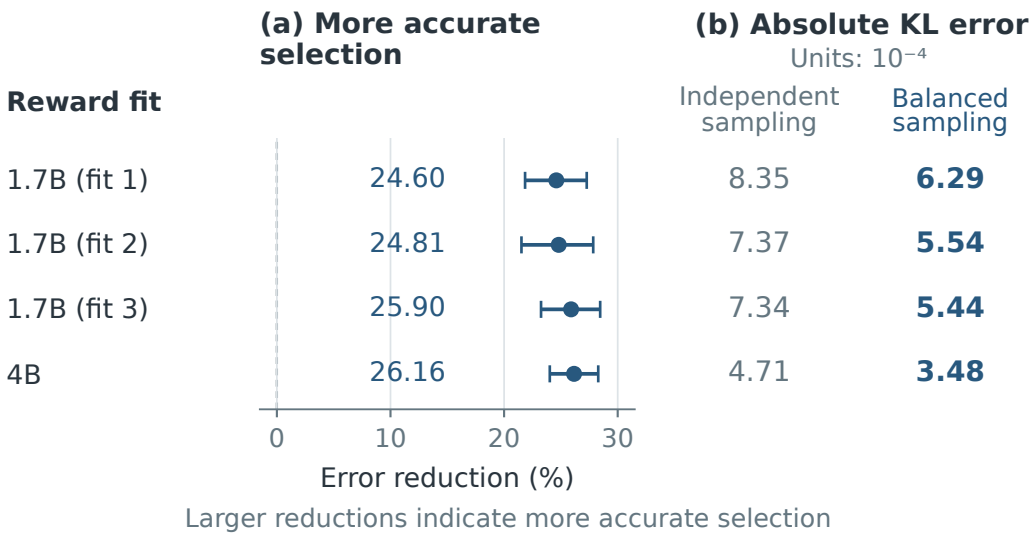

Figure 5: **Balanced sampling reduces error in neural response selection.** Left: percentage reduction in forward-KL error compared with independent sampling. Intervals are 98.33% for the three 1.7B fits and 95% for 4B, conditional on the fitted model, cohort and finite reference bank. Right: the corresponding absolute errors, in units of 10^{-4}, without absolute-error intervals. Fits 1–3 use 1.7B BF16 seed 42, 1.7B FP32 seed 42, and 1.7B FP32 seed 43, respectively. The 4B fit uses FP32 seed 42. Both sampling laws use 20 reward evaluations at matched projection work.

For the 4B fit, IID and balanced forward KL are 4.71×10^{-4} and 3.48×10^{-4}, giving ratio 0.74 $[0.72, 0.76]$. Balanced sampling also increases agreement with the reference selector's top response by 0.79 percentage points $[0.44, 1.15]$ and lowers its reference-score shortfall by 3.72×10^{-4} $[1.76, 5.76] \times 10^{-4}$. These latter intervals are paired descriptive summaries of prespecified secondary endpoints.

The three 1.7B forward-KL ratios, expressed as percentages, are 75.40%, 75.19% and 74.10%. All adjusted upper endpoints are below 100% (Table 12). This consistency across the four fixed fits does not estimate a population effect over model retraining or prompts.

Sensitivity to the finite reference bank. As a post hoc check, we repeat the temperature-one comparison with each independent 1,024-draw bank as the target, keeping all 144 groups and 32 paired repetitions. Balanced/IID forward-KL ratios remain below one for both banks in every fit (Table 13). Neither bank replaces the pooled primary target. These point estimates assess sensitivity to the realized reference draws. They neither propagate target uncertainty into the primary intervals nor establish population accuracy.

For the 4B fit, each sampling law projects 290,161,920 token–member rows. IID and balanced stochastic projection take 6,416.5 and 6,365.5 seconds, respectively (1.78 and 1.77 GPU-hours). Target construction uses additional work and is excluded from the matched-work comparison. Because the candidate set, fitted reward law, number of members and projection rows are identical, the result supports better fidelity at matched projection work. It does not establish higher human-preference accuracy, fewer projections at a fixed target error, or lower total end-to-end training cost.

Table 11: **Fixed-rule approximation and agreement with the reference selector.** Fixed 4B fit, 144 held-out groups, $N = 20$, $\beta = 1$, pooled 2048-draw target. The evaluation cohort and both sampling protocols were fixed before the evaluation draws. Conditional 95% paired intervals resample 32 complete repetitions. Differences are balanced minus IID, in percentage points for agreement. Choice and shortfall are prespecified descriptive endpoints.

Endpoint	IID	**Balanced**	Difference [95% CI]
Target KL $\times 10^3 \downarrow$	0.47	$\mathbf{0.35}$	-0.12 $[-0.14, -0.11]$
Top-choice agreement (%) $\uparrow$	96.54	$\mathbf{97.33}$	$+0.79$ $[+0.44, +1.15]$
Score shortfall $\times 10^3 \downarrow$	1.28	$\mathbf{0.91}$	-0.37 $[-0.58, -0.18]$

Table 12: Balanced-to-IID forward-KL ratios for the three 1.7B reward fits on the same 144 response groups. Ratios and interval endpoints are expressed as percentages. Each fit uses 32 paired repetitions. Primary intervals at $\beta = 1$ use Bonferroni correction for nominal 95% familywise coverage across fits. The $\beta = 0.2$ comparison is secondary. Intervals condition on the fitted models and observed cohort.

Reward fit	$\beta = 1$ forward KL		$\beta = 0.2$ forward KL		Repetitions
	Ratio	98.33% marginal Bonferroni CI	Ratio	95% CI	
Qwen3-1.7B BF16, seed 42	**75.40**	[72.73, 78.15]	76.30	[73.40, 79.20]	32
Qwen3-1.7B FP32, seed 42	**75.19**	[72.15, 78.47]	73.40	[69.30, 77.40]	32
Qwen3-1.7B FP32, seed 43	**74.10**	[71.54, 76.75]	72.80	[69.80, 76.10]	32

Table 13: **Reference-bank sensitivity of fixed-law selector fidelity.** Balanced/IID ratios of mean forward KL at $\beta = 1$, expressed as percentages. Both separate banks contain 1,024 draws per response. The pooled target contains 2,048. Post hoc descriptive point estimates condition on the same four fits, 144 groups and 32 paired repetitions.

Fixed Qwen3 fit	Bank 0	Bank 1	Pooled
1.7B BF16, seed 42	75.66	75.62	75.40
1.7B FP32, seed 42	75.19	75.63	75.19
1.7B FP32, seed 43	74.38	74.32	74.10
4B FP32, seed 42	73.92	74.26	73.84

F Additional implementation studies

F.1 Additional learning, cost and reward-quality studies

These studies examine finite-compute implementations and the distinction between selector precision and useful decisions. Their samplers, cohorts and endpoints differ from the fresh-posterior experiment in Appendix D.2 and the four-fit, 144-group study in Appendix E.1. Designs and sample sizes were informed by earlier pilots. The within-study intervals do not provide multiplicity control over the entire research sequence.

Learning with a retained ensemble. An eight-dimensional bounded-$\tanh$ reward model supplies 20 fixed feature vectors, including an anchor. Both selection and the regret oracle exclude the anchor, leaving 19 eligible responses and preventing self-comparisons. The study uses 1,536 new environments, two repetitions per method and $T = 2000$ queries. Ensembles retain $N = 20$ particles across rounds, with minibatch size 32, one update per round, step size $1/t$, clipping at three and temperature one. OLE adds 0.3 times the sample SD, using divisor $N - 1$, to the sampled mean. Initialization has scale 0.3, while the drift uses a unit-Gaussian prior. These are implementation settings, not a claim that initialization samples the target prior.

OLE achieves mean cumulative regret 228.89, compared with 247.16 for matched mean selection. The primary paired difference is -18.27 $[-26.05, -10.49]$, a 7.39% reduction. Repetitions are averaged within environment before forming the two-sided 95% Student-t interval over 1,536 independent environments. Only this contrast is confirmatory within the study. Online Fisher-UCB

(Faury et al., 2020) has lower regret, 214.12. Each ensemble trajectory uses 1,270,080 likelihood-gradient terms, compared with 63,504 for Fisher, whose covariance and candidate-gradient work is additional. The result supports the optimism ablation in this retained-state implementation, without establishing the fresh posterior-marginal conditions of Theorem 3.1.

Learning with few comparisons. A separate comparison examines zero-sum and independent innovations in the retained-ensemble implementation of Appendix C.2. It uses 512 independent eight-dimensional nonlinear preference environments, four repetitions per method and $T = 250$ queries. Both arms retain 20 particles and use the same one-update-per-round gradient computation. Zero-sum sampling couples initialization and every Gaussian innovation. Averaging repetitions within each environment gives a paired regret difference of -2.25 with a two-sided 99% Student-t interval $[-3.72, -0.79]$. The horizon was chosen after exploratory short-horizon results and fixed before the new cohort. Longer-horizon comparisons did not establish an advantage. The result identifies a benefit of this sampling design in a limited-feedback setting, without establishing superiority to Fisher-UCB (Faury et al., 2020) or exact posterior sampling.

An additional fresh-posterior control. A bounded linear-logistic control uses a standard Gaussian prior truncated to $[-1/\sqrt{2}, 1/\sqrt{2}]^2$, eight candidate features per round including a zero-feature reference, and continuous posterior rejection sampling. The reference is excluded from selection and the regret oracle. Across 384 Bayesian environments and $T = 256$ queries, the IID optimistic variant minus its matched two-draw mean variant has cumulative regret $+1.41$ with a 95% paired Student-t interval $[-0.13, 2.94]$. The interval includes zero. This additional setting illustrates why an accurate posterior sampler alone does not ensure an advantage over mean selection.

A separate tuned cost comparison. Each method selects one of four configurations on 32 development environments with two repetitions. The grid is $\{0.1, 0.3, 1, 3\}$ for the OLE and Fisher-UCB (Faury et al., 2020) bonuses, or the step multiplier for the other methods. Evaluation then uses 256 independent environments and two repetitions per environment. Table 14 reports all five methods. The paired OLE-minus-Fisher regret difference is $+19.48$ $[5.57, 33.40]$, using a 97.5% interval. OLE is also slower in the serial timing comparison. Timings are medians of 24 trajectories, comprising eight fixed test environments and three repetitions in rotated method order, with one CPU core and one BLAS thread on a shared host. They exclude tuning and are descriptive, rather than an isolated hardware benchmark.

Precision at different projection budgets. A separate neural study uses 128 label-blindly selected RewardBench pairs, 32 per canonical section (Lambert et al., 2025), and four fitted Qwen3 RewardUQ models (Yang et al., 2026). Its endpoint is repeat disagreement: the average of the two directed KL divergences between soft selectors constructed independently under the same sampling law. It is distinct from forward KL to the finite reference selector in the 144-group study. Scores use mean plus twice the sample SD, a uniform base and softmax temperatures $\beta \in \{1, 0.2\}$. Balanced dropout forms complete ten-member blocks with drop probability 0.1.

A retrospective extension examines $N \in \{10, 20, 30, 40, 60, 80\}$ after the $N = 20, 80$ results were known. Across 48 matched-work comparisons, balanced/IID repeat-KL ratios range from 0.62 to 0.79. Pointwise 95% intervals resample 32 complete paired repeat indices, after averaging the fixed response pairs, with 10,000 bootstrap draws. They condition on the fitted models and cohort. The full $N = 80$ bank uses 340,587,520 token–member vocabulary projections per method across 64 constructions, with prefix work proportional to N. However, balanced $N = 20$ versus IID $N = 30$ and balanced $N = 40$ versus IID $N = 60$ give 16 ratios ranging from 1.06 to 1.20. None has a 95% upper endpoint at or below one. The precision improvement therefore does not establish that one-third fewer projections preserve repeatability. Projection counts also exclude training, backbone computation, reference scoring and pilot draws, and are not elapsed-time speedups.

External reward quality. A separate evaluation uses the fixed Qwen3-4B FP32 seed-42 RewardUQ fit on 1,864 eligible RewardBench 2 groups and 8,973 responses (Malik et al., 2026). One group is excluded by the 2,048-token limit. Combining the 51 reference/tied group pairs in the Ties domain gives 1,813 official scoring units. IID, antithetic and balanced sampling each use $N = 20$, both mean-only and mean-plus-2SD scores, identical projection work and eight paired repetitions.

Table 14: **Retained-ensemble regret and serial cost.** The 256-environment comparison is separate from the 1,536-environment anchor-excluded study. Regret intervals are 95% environment-level intervals. Time is the median milliseconds per trajectory and excludes tuning. Citations identify method families. Fisher TS denotes the local Fisher-preconditioned sampling variant, not a separately attributed algorithm.

Method	Mean regret	95% interval	Time (ms)
Mean, $N = 20$	243.56	[225.58,261.55]	370
OLE, $N = 20$	233.49	[222.43,244.56]	373
Fisher-UCB (Faury et al., 2020)	214.01	[203.75,224.27]	239
Thompson sampling (Russo & Van Roy, 2014), $N = 1$	293.16	[282.02,304.31]	205
Fisher TS (local) (Russo & Van Roy, 2014)	284.40	[275.35,293.45]	292

Balanced optimistic selection minus IID optimistic selection changes official macro accuracy by $+0.02$ percentage points, with a paired group-by-repeat 97.5% interval $[-0.32, 0.35]$. Balanced optimistic selection minus balanced mean selection gives -0.36 points $[-0.86, 0.14]$. Neither contrast establishes improvement, and their joint criterion requiring both lower endpoints to exceed zero is not met. This evaluates immediate externally labeled response quality, not the information obtained by acquiring labels and updating a learner. Together with the projection-budget comparison, it separates three claims: more precise implementation of a selector, lower computation, and better decisions. Evidence for the first does not establish the other two.

G AGENTIC OLE IMPLEMENTATION

G.1 AGENTIC OLE: IMPLEMENTATION AND EVALUATION

Agentic OLE learns from trajectories containing reasoning, retrieval and memory actions. The agent can retrieve evidence, manage memory and produce an answer. Rewards evaluate answer quality, action format, retrieval relevance and memory operations. Reward-sensitive rejection selects experience for policy updates. The application therefore instantiates OLE at the level of generated training trajectories.

G.1.1 TRAINING AND SELECTION

Training questions are curated from the HotpotQA (Yang et al., 2018) and 2WikiMultiHopQA (Ho et al., 2020) training splits. For each candidate question, ten TC-RAG (Jiang et al., 2025) trajectories are generated. Questions are excluded if all trajectories are incorrect, at least half are correct, or the question is solved with at most one retrieval. This procedure concentrates training on questions that require retrieval and leave room for improvement. The retrieval corpus is a Wikipedia dump dated November 1, 2023.

Trajectory and prompt-group selection. For prompt b, let $r_{b,i}$ be a generated trajectory's reward, and let μ_b and σ_b^2 be the group's empirical reward mean and variance. The trajectory-level rule ranks trajectories by $\widetilde{r}_{b,i} = \mu_b + \sigma_b^2 r_{b,i}$ and retains a global top fraction. The prompt-group implementation computes $\mu_b + k\sigma_b$ and sets the advantages of groups below a threshold to zero. The threshold for the next batch is the current batch's $(1-p)$-quantile for retained fraction p. The latter implementation processes the full actor batch even when a group's advantages are zero.

These implementations use variation across realized trajectory rewards. Their selection unit and filtering operation differ. Which variant was used in the rejection ablation is unresolved. The contrasts therefore do not identify the effect of either selection rule or separate reward aggregation from retention.

Optimization and comparison protocols. The nominal training protocol uses GRPO (Shao et al., 2024), learning rate 10^{-6}, batch size 16, one epoch and 300 steps. Generation uses temperature 0.7, a 4,096-token context and at most 2,048 new tokens. The number of rollouts per prompt is unresolved,

with conflicting values of four and ten. The batch size and training duration of the evaluated models also cannot be verified. These uncertainties limit exact replication of the training comparison.

The QA comparison includes direct and chain-of-thought prompting, fixed-interval retrieval, prompted agents and trained search agents. FS-RAG and FL-RAG denote this evaluation's sentence-based and fixed-token retrieval controls, respectively. IRCoT (Trivedi et al., 2023) interleaves retrieval with generated reasoning sentences. ReSearch (Chen et al., 2025) uses GRPO in its original method. Search-R1 (Jin et al., 2025) supports both PPO (Schulman et al., 2017) and GRPO. The optimizer configurations of the evaluated baseline models are unspecified. Per-method retrieval and token budgets are not established as identical. The tables compare the resulting systems under their respective training and search procedures.

G.1.2 Complete QA comparisons

The evaluation contains seven QA datasets. HotpotQA and 2WikiMultiHopQA are the two in-domain data sources. Bamboogle (Press et al., 2023), FRAMES (Krishna et al., 2025), MuSiQue (Trivedi et al., 2022), Natural Questions (Kwiatkowski et al., 2019) and TriviaQA (Joshi et al., 2017) are the out-of-domain tasks in this study. This designation concerns the agentic training sources, without asserting that the pretrained language models have never encountered these datasets. Tables 15–17 give every task score for all thirteen methods at each model size.

F1 is reported in percentages. Each mean is the arithmetic average of the seven task scores. The Search-R1 (Jin et al., 2025) 7B mean is 34.08. Dataset-level means do not estimate variation across training runs. Per-seed results and confidence intervals are unavailable, and evaluated item counts are unspecified.

G.1.3 Rejection, memory and reward ablations

Table 18 includes the complete memory/rejection comparison and three reward-component ablations. Rejection raises mean F1 by 8.12 points with memory actions and 3.13 points without them. The latter contrast includes declines on Bamboogle and FRAMES. Removing retrieval or memory rewards reduces mean F1 from 33.55 to 25.53 or 24.39, respectively. Removing both yields 19.94. Thus selection operates within a system whose memory actions and intermediate rewards also affect performance. The rejection ablation does not compare a dispersion bonus against mean-only ranking or random retention.

G.1.4 Broader applications

DeepResearch Bench (Du et al., 2025) evaluates multi-document report generation. RACE measures report quality. FACT evaluates effective citations and claim accuracy. Table 19A reports all three metrics and the RAG, ReAct (Yao et al., 2023) and Enterprise Deep Research (Prabhakar et al., 2025) baselines. OLE improves report quality, effective citation and claim accuracy over the strongest listed comparator by 1.39, 8.14 and 8.56 points, respectively. These outcomes evaluate the complete retrieval, memory, reward and policy system.

ALFWorld (Shridhar et al., 2021) and WebShop (Yao et al., 2022) evaluate interaction with household and shopping environments. Table 19B gives all ALFWorld categories. Pick, Look, Cool and Pick2 improve, while Clean and Heat decline. Overall success is reported alongside the category rates. WebShop's score and success rate are reported separately in Table 19C. The DeepResearch model checkpoint and the ALFWorld/WebShop backbone, episode counts and adaptation procedure are unspecified. These results therefore describe application performance, without establishing frozen-checkpoint or zero-shot transfer.

The medical-QA evaluation uses Qwen2.5-7B-Instruct (Qwen et al., 2025) on MedQA (Jin et al., 2021), with Qwen2.5-72B as the judge. Table 19D separates format compliance from answer correctness. OLE achieves 92% format accuracy and 87% answer accuracy. The answer gain over the strongest comparator, Base, is three points.

Table 15: **Agentic QA with Qwen2.5-1.5B.** F1 (%) on all seven tasks. Training uses the two in-domain data sources. Means are recomputed from the seven displayed task scores. Bold marks column maxima.

	In-domain		Out-of-domain					
Method	2Wiki	HotpotQA	Bamboogle	FRAMES	MuSiQue	NQ	TriviaQA	Mean
Base (Qwen et al., 2025)	21.95	20.20	5.87	9.51	8.26	11.31	32.81	15.70
CoT (Wei et al., 2022)	16.75	17.90	19.74	8.43	7.22	9.82	25.17	15.00
FS-RAG†	21.37	26.55	15.31	11.40	9.50	17.74	44.73	20.94
FL-RAG†	26.08	28.14	14.84	12.58	10.30	21.94	48.59	23.21
ReAct (Yao et al., 2023)	12.43	23.48	12.54	8.63	7.99	21.06	33.91	17.15
IRCoT (Trivedi et al., 2023)	21.79	26.68	17.49	9.32	8.73	22.93	44.19	21.59
TC-RAG (Jiang et al., 2025)	25.23	22.29	11.63	10.27	7.90	14.21	34.05	17.94
ReSearch (Chen et al., 2025)	25.55	30.45	**22.16**	9.21	9.54	29.07	44.65	24.38
Search-R1 (Jin et al., 2025)	26.51	20.66	13.80	9.65	8.36	11.93	25.76	16.67
AEPO (Dong et al., 2025a)	17.34	14.75	11.58	8.98	7.11	9.55	23.02	13.19
ARPO (Dong et al., 2025b)	20.02	15.48	12.38	9.11	6.16	9.19	23.96	13.76
Mem1 (Zhou et al., 2025)	14.43	14.34	12.03	8.18	6.08	18.41	22.06	13.65
OLE	**29.53**	**30.88**	20.00	**13.46**	**16.48**	**31.35**	**51.83**	**27.65**

† FS-RAG and FL-RAG are sentence-based and fixed-token retrieval controls. See Appendix G.1.

Table 16: **Agentic QA with Qwen2.5-3B.** F1 (%) on all seven tasks. Training uses the two in-domain data sources. Means are recomputed from the seven displayed task scores. Bold marks column maxima.

	In-domain		Out-of-domain					
Method	2Wiki	HotpotQA	Bamboogle	FRAMES	MuSiQue	NQ	TriviaQA	Mean
Base (Qwen et al., 2025)	23.98	24.08	9.45	8.01	9.70	14.27	38.84	18.33
CoT (Wei et al., 2022)	18.90	23.82	20.80	7.16	10.47	16.53	39.62	19.61
FS-RAG†	15.47	25.85	10.48	10.42	7.64	19.84	45.38	19.30
FL-RAG†	16.80	26.78	11.05	9.19	7.29	21.93	48.50	20.22
ReAct (Yao et al., 2023)	25.09	34.37	24.86	10.53	13.92	27.19	46.04	26.00
IRCoT (Trivedi et al., 2023)	15.89	24.50	25.27	6.79	12.43	27.86	49.19	23.13
TC-RAG (Jiang et al., 2025)	28.47	21.94	17.59	7.69	8.99	20.74	50.46	22.27
ReSearch (Chen et al., 2025)	27.23	33.96	15.09	10.00	9.47	34.61	53.93	26.33
Search-R1 (Jin et al., 2025)	29.90	37.24	29.90	10.76	13.53	34.73	55.08	30.16
AEPO (Dong et al., 2025a)	23.01	28.71	22.09	12.26	11.70	26.76	47.78	24.62
ARPO (Dong et al., 2025b)	29.55	36.48	27.32	13.49	13.38	33.29	53.66	29.60
Mem1 (Zhou et al., 2025)	18.06	20.15	5.19	4.99	4.47	19.18	33.09	15.02
OLE	**32.92**	**44.00**	**31.48**	**16.23**	**16.48**	**37.15**	**56.62**	**33.55**

† FS-RAG and FL-RAG are sentence-based and fixed-token retrieval controls. See Appendix G.1.

Table 17: **Agentic QA with Qwen2.5-7B.** F1 (%) on all seven tasks. Training uses the two in-domain data sources. Means are recomputed from the seven displayed task scores. Bold marks column maxima.

	In-domain		Out-of-domain					
Method	2Wiki	HotpotQA	Bamboogle	FRAMES	MuSiQue	NQ	TriviaQA	Mean
Base (Qwen et al., 2025)	25.41	26.63	17.86	12.52	12.15	19.72	49.08	23.34
CoT (Wei et al., 2022)	23.55	29.10	37.56	17.60	14.35	22.47	49.33	27.71
FS-RAG†	17.71	29.21	16.86	12.52	10.74	16.82	35.02	19.84
FL-RAG†	19.78	34.42	24.10	12.10	12.46	19.72	42.66	23.61
ReAct (Yao et al., 2023)	27.51	42.81	27.63	15.29	19.34	**30.01**	54.55	31.02
IRCoT (Trivedi et al., 2023)	36.45	26.29	21.90	6.78	8.39	19.63	49.43	24.12
TC-RAG (Jiang et al., 2025)	29.70	40.83	25.13	16.46	17.56	29.01	54.78	30.50
ReSearch (Chen et al., 2025)	30.03	30.39	30.42	15.61	12.58	23.69	48.25	27.28
Search-R1 (Jin et al., 2025)	35.03	38.89	42.04	18.01	19.08	29.59	55.91	34.08
AEPO (Dong et al., 2025a)	19.88	13.85	13.24	7.24	5.85	9.93	17.53	12.50
ARPO (Dong et al., 2025b)	30.71	25.20	32.94	12.18	12.71	17.80	40.16	24.53
Mem1 (Zhou et al., 2025)	25.29	29.98	36.50	14.15	14.13	26.38	51.04	28.21
OLE	**38.34**	**45.15**	**49.21**	**19.44**	**22.01**	23.60	**58.45**	**36.60**

† FS-RAG and FL-RAG are sentence-based and fixed-token retrieval controls. See Appendix G.1.

Table 18: **Complete agentic ablations.** Qwen2.5-3B F1 (%). The upper panel crosses memory actions and rejection. The lower panel removes retrieval and memory reward components from the full system.

Configuration	2Wiki	HotpotQA	Bamboogle	FRAMES	MuSiQue	NQ	TriviaQA	Mean
OLE	32.92	44.00	31.48	16.23	16.48	37.15	56.62	33.55
Without rejection	26.86	28.33	27.69	12.98	12.68	23.67	45.80	25.43
Without memory actions	27.54	28.77	24.60	9.83	13.86	35.10	51.91	27.37
Without both	24.56	28.61	25.49	11.61	12.54	22.07	44.83	24.24
Without retrieval reward	23.70	34.71	19.68	7.70	11.45	29.07	52.39	25.53
Without memory reward	27.49	33.78	17.14	7.41	7.63	22.00	55.29	24.39
Without both rewards	22.98	26.42	15.74	5.33	6.07	18.62	44.40	19.94

Table 19: **Broader applications of Agentic OLE.** The comparisons evaluate complete systems, including OLE's retrieval, memory, reward and policy components. DeepResearch Bench metrics and WebShop score use their native units, while ALFWorld and WebShop success rates and both medical QA accuracies are percentages. ALFWorld Overall is the evaluation aggregate, and medical QA uses Qwen2.5-7B-Instruct judged by Qwen2.5-72B. † RAG and Prompt denote evaluation-specific controls whose detailed configurations are unspecified.

A. DeepResearch Bench

Method	Report quality	Effective citation	Claim accuracy
RAG†	36.68	2.93	10.86
ReAct (Yao et al., 2023)	42.21	4.69	23.78
EDR (Prabhakar et al., 2025)	40.99	12.28	23.70
OLE	**43.60**	**20.42**	**32.34**

B. ALFWorld success (%)

Method	Pick	Look	Clean	Heat	Cool	Pick2	Overall
Prompt†	45.8	48.4	**34.5**	**41.0**	19.6	31.7	36.5
OLE	**84.7**	**80.6**	24.1	33.3	**37.0**	**51.2**	**51.1**

C. WebShop

Method	Score	Success (%)
Prompt†	14.7	0.8
OLE	**27.8**	**2.3**

D. Medical QA

Method	Format accuracy (%)	Answer accuracy (%)
Base (Qwen et al., 2025)	39	84
ReAct (Yao et al., 2023)	56	79
OLE	**92**	**87**