ADAPTIVE PAIRWISE-FEEDBACK SAMPLING FOR MULTI-PROMPT KL-REGULARIZED POLICIES

Anonymous authors
Paper under double-blind review

ABSTRACT

Pairwise comparisons can be easier to obtain than calibrated rewards, yet using them to construct response policies usually involves fitting a reward model or a parameterized policy. We study a complementary setting: each prompt has a fixed finite set of candidate responses, and a limited feedback budget must be distributed across prompts with different deployment importance and sampling difficulty.

We introduce a direct sampling procedure that constructs KL-regularized response policies from fresh comparisons. Under the Bradley–Terry model, accepting a proposed response after repeated comparison wins yields a Markov chain with the desired stationary policy at an implementable temperature. We analyze the empirical occupation policy actually returned by the sampler, establishing a finite-sample objective-gap bound and an exact asymptotic leading term. This characterization reveals a square-root allocation rule balancing deployment weights and local sampling difficulties. An adaptive scheduler estimates these difficulties online and converges to the oracle allocation.

At deterministic horizons of T global update rounds, the adaptive procedure achieves an expected aggregate objective gap of $O(T^{-1/2})$ and an $O_p(T^{-1})$ rate matching the oracle's order. Synthetic experiments show objective gaps comparable to the known-difficulty oracle and smaller than round-robin allocation, supporting the practical value of the allocation rule. Experiments on SQLite questions further illustrate improved pairwise preference performance over round-robin under LLM evaluation. Together, these results connect direct comparison-based sampling, policy-quality guarantees, and adaptive feedback allocation.

1 INTRODUCTION

Large language models can produce several plausible responses to the same prompt, for example through different decoding seeds, system prompts, or checkpoints. Yet a system may still lack a trustworthy scalar score for deciding how these responses should be used. In this setting, a human annotator or preference model may reliably answer a simpler question: which of two responses is better? Pairwise comparisons therefore provide a natural form of supervision when absolute quality is difficult to define or calibrate. This perspective underlies preference-based reinforcement learning and modern reinforcement learning from human feedback (RLHF) (Christiano et al., 2017; Ouyang et al., 2022).

RLHF and preference-optimization methods commonly use comparisons to fit a reward model or a parameterized policy (Christiano et al., 2017; Ouyang et al., 2022; Rafailov et al., 2023). Comparisons may be collected in advance or during learning, but they serve as training data for the fitted model. When a finite candidate set is already available for each prompt, model fitting and policy optimization can be unnecessary or operationally costly. We instead ask whether fresh pairwise comparisons can directly construct a useful policy, without fitting a reward model or repeatedly retraining a parameterized language model.

We study a finite collection of prompts $\mathcal{X}$. Each prompt x has a finite candidate-response set $\mathcal{S}_x$, a deployment weight f_x, and a full-support reference policy q_x. The reference policy specifies the baseline distribution over available candidate responses, while f_x measures the prompt's importance in the deployment population. Pairwise outcomes are generated from an unobserved prompt-specific utility r_x, but that utility is never observed by the algorithm.

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.

Our scope is feedback allocation over these fixed candidate sets, without generalization to unseen prompts.

Our objective is not merely to identify one best response per prompt. A deployed system may need to randomize among high-quality responses, retain useful diversity, and remain close to a trusted reference policy. We therefore seek policies $\pi = (\pi_x)_{x\in\mathcal{X}}$ that maximize the deployment-weighted KL-regularized value at a temperature $\widetilde{\tau}$ that the sampler can implement exactly:

$$J(\pi) := \sum_{x\in\mathcal{X}} f_x \left[\sum_{s\in\mathcal{S}_x} \pi_x(s) r_x(s) - \widetilde{\tau}\,\mathrm{KL}(\pi_x \| q_x) \right].$$

Its optimizer is the exponential tilt $\pi_x^*(s) \propto q_x(s)\exp(r_x(s)/\widetilde{\tau})$. Thus, the objective formalizes a utility–conservatism tradeoff: it favors responses that are preferred under the latent comparison model while penalizing excessive movement away from the reference policy. Our goal is to return a policy with a small objective gap $J(\pi^*) - J(\widehat{\pi})$ using only a finite stream of noisy comparisons.

The key observation is that fresh comparisons can determine Markov-chain transitions. For one prompt, we propose a candidate from the reference policy and accept it only after it wins m independent Bradley–Terry comparisons against the current response. This repeated-win rule defines a finite Markov chain whose stationary distribution is exactly the KL-regularized response policy at the implementable temperature $\widetilde{\tau} = 1/m$. The method requires no utility values, reward-model scores, or repeated policy fitting; its returned policy is the empirical occupation measure of the chain. We also identify the exact leading term of the expected objective gap after n local updates. This characterization reveals a prompt-specific sampling difficulty that determines how feedback should be allocated across prompts.

A finite comparison budget must be divided among prompts. Some prompts matter more in deployment, while others require more updates before their returned policies stabilize. A uniform allocation can therefore use comparisons inefficiently. We develop an adaptive scheduler that uses the comparisons observed so far to decide which prompt to update next, while ensuring that every prompt continues to receive updates. Section 4 establishes an objective-gap rate in probability matching the order of an ideal allocation that knows the prompt-specific difficulties in advance. The procedure may be stopped after any number of rounds and returns a policy from the data collected so far.

We evaluate the allocation rule in synthetic and SQLite experiments. Learned allocation yields synthetic objective gaps comparable to the known-difficulty oracle and smaller than round-robin. On fixed SQLite questions, fresh, counterbalanced LLM judgments favor adaptive allocation over round-robin under a prespecified deployment distribution. This preference metric does not directly measure the KL-regularized objective gap.

1.1 Contributions

- **Direct KL-regularized policy sampling from pairwise feedback.** We show that a repeated-win comparison rule has the implemented KL-regularized response policy as its exact stationary distribution. This gives a direct alternative to reward-model fitting and repeated policy optimization in the finite candidate-response setting. We also distinguish the desired temperature from the implementable temperature and quantify their discrepancy.
- **Finite-sample guarantees and asymptotic sharpness.** Starting from an arbitrary response, we prove an expected objective-gap upper bound for the empirical policy returned after any finite number n of local updates. The bound has a leading term of order $1/n$, with a lower-order correction for initialization and finite-horizon effects. We also establish an asymptotically sharp $1/n$ expansion, identifying the exact leading coefficient of the expected objective gap. Thus, the result evaluates the policy that the sampler actually returns, rather than only its limiting distribution.
- **Learning to allocate comparisons across prompts.** We introduce CTM, an adaptive scheduler that learns from the comparisons it observes which prompts require more updates, while continuing to explore every prompt. From arbitrary initial responses, we prove an expected aggregate KL-regularized objective-gap bound of order $T^{-1/2}$ and a sharper $O_p(T^{-1})$ rate for the returned policy collection at deterministic horizons of T global update rounds. The latter matches the oracle's order without requiring knowledge of the prompt-specific difficulties.

1.2 Related work

Preference optimization and online RLHF. RLHF uses comparisons to train reward models and policies (Christiano et al., 2017; Ouyang et al., 2022); direct preference optimization replaces the explicit reward-model stage but still fits a policy from comparison data (Rafailov et al., 2023). Recent online methods collect fresh preferences while iteratively learning a policy, reward model, or general preference model (Xiong et al., 2024; Munos et al., 2024; Calandriello et al., 2024). Our method uses fresh comparisons to determine finite-state Markov-chain transitions and returns an empirical occupation policy.

Preference bandits and comparison-driven chains. Dueling and contextual dueling bandits use relative feedback to identify a preferred action or control regret (Yue et al., 2012; Dudík et al., 2015). Our target is instead the full KL-regularized response distribution for every prompt. Our local construction is also related to Markov-chain sampling and ranking from pairwise comparisons (Barker, 1965; Sanborn & Griffiths, 2007; Fotakis et al., 2022; Negahban et al., 2017); our analysis concerns the finite-budget occupation policy rather than a single selected winner or an estimator of unregularized item strengths.

Adaptive allocation. The square-root allocation is related to optimal computing-budget allocation (Chen et al., 2000). Here, however, the loss is a deployment-weighted KL objective gap and the difficulty factor is the integrated autocorrelation quantity d_x. The adaptive component allocates feedback among several distribution-estimation problems while preserving the fixed local transition kernels.

2 Problem formulation

Prompts and pairwise feedback. Let $\mathcal{X}$ be a finite prompt set. Each prompt $x \in \mathcal{X}$ has a finite candidate-response set $\mathcal{S}_x = [K_x]$, where $K_x \geq 2$. We refer to a candidate response $s \in \mathcal{S}_x$ as an arm when convenient. Prompt x also has a *deployment weight* $f_x > 0$, with $\sum_{x\in\mathcal{X}} f_x = 1$, a *reference policy* $q_x \in \Delta(\mathcal{S}_x)$ satisfying $q_x(s) > 0$ for every $s \in \mathcal{S}_x$, and an unobserved utility function $r_x : \mathcal{S}_x \to \mathbb{R}$. Here, $\Delta(\mathcal{S}_x)$ denotes the set of probability distributions on $\mathcal{S}_x$.

Following the Bradley–Terry model for paired comparisons (Bradley & Terry, 1952), a comparison between two distinct responses $s, s' \in \mathcal{S}_x$ produces

$$Y_x(s, s') \sim \text{Bernoulli}(p_x(s, s')), \qquad p_x(s, s') := \frac{e^{r_x(s)}}{e^{r_x(s)} + e^{r_x(s')}}. \tag{1}$$

The event $Y_x(s, s') = 1$ means that s is preferred to s'. Conditional on the entire history before a query and its selected prompt and ordered pair (x, s, s'), the fresh outcome has the Bernoulli law in (1). The odds satisfy $p_x(s, s')/p_x(s', s) = e^{r_x(s)-r_x(s')}$.

Policy objective. For distributions μ and q on the same finite set, with q having full support, define $\mathrm{KL}(\mu\|q) := \sum_s \mu(s)\log(\mu(s)/q(s))$, with the convention $0\log(0/q(s)) = 0$.

The repeated-win update requires an integer confirmation level m. For a requested temperature $\tau > 0$, we therefore choose $m := \max\{1, \mathsf{round}(1/\tau)\}$, where $\mathsf{round}(1/\tau) := \lfloor 1/\tau + 1/2 \rfloor$. Our guarantees concern the implemented temperature $\widetilde{\tau} := 1/m$.

For comparison with the policy requested at temperature τ, define the corresponding exponential tilt

$$\pi_x^{(\tau)}(s) := \frac{q_x(s)e^{r_x(s)/\tau}}{\sum_{s'\in\mathcal{S}_x} q_x(s')e^{r_x(s')/\tau}}. \tag{2}$$

For a policy $\pi_x \in \Delta(\mathcal{S}_x)$, define

$$J_x(\pi_x) := \sum_{s\in\mathcal{S}_x} \pi_x(s)r_x(s) - \widetilde{\tau}\,\mathrm{KL}(\pi_x\|q_x), \tag{3}$$

and aggregate performance across prompts as $J(\pi) := \sum_{x\in\mathcal{X}} f_x J_x(\pi_x)$, where $\pi = (\pi_x)_{x\in\mathcal{X}}$.

Proposition 2.1 (Target policy and exact objective-gap identity)**.** *The policy*

$$\pi_x^*(s) = \frac{q_x(s)e^{mr_x(s)}}{\sum_{s' \in \mathcal{S}_x} q_x(s')e^{mr_x(s')}} \tag{4}$$

is the unique maximizer of J_x over $\Delta(\mathcal{S}_x)$. Moreover, for every $\pi_x \in \Delta(\mathcal{S}_x)$,

$$J_x(\pi_x^*) - J_x(\pi_x) = \widetilde{\tau}\,\mathrm{KL}(\pi_x \| \pi_x^*). \tag{5}$$

Consequently, $\pi^* := (\pi_x^*)_{x \in \mathcal{X}}$ is the unique maximizer of J, and

$$J(\pi^*) - J(\pi) = \widetilde{\tau} \sum_{x \in \mathcal{X}} f_x\,\mathrm{KL}(\pi_x \| \pi_x^*). \tag{6}$$

The policy–reward correspondence in (4) also underlies direct preference optimization (DPO; Rafailov et al., 2023); here it yields exact objective-gap identities (proof in Appendix B.1). The next result separates temperature-discretization error from sampling error.

Proposition 2.2 (Effect of temperature discretization)**.** *Let $R_x := \max_{s \in \mathcal{S}_x} r_x(s) - \min_{s \in \mathcal{S}_x} r_x(s)$. Then, for every prompt x,*

$$\max\left\{\mathrm{KL}\left(\pi_x^* \middle\| \pi_x^{(\tau)}\right), \mathrm{KL}\left(\pi_x^{(\tau)} \middle\| \pi_x^*\right)\right\} \le \left|m - \frac{1}{\tau}\right| R_x. \tag{7}$$

Consequently,

$$\left\|\pi_x^* - \pi_x^{(\tau)}\right\|_{\mathrm{TV}} \le \sqrt{\frac{R_x}{2}\left|m - \frac{1}{\tau}\right|}. \tag{8}$$

In particular, the two policies coincide when $1/\tau$ is an integer.

The proof appears in Appendix B.2.

Sequential task. Starting from arbitrary responses, the procedure sequentially selects prompts for local updates and returns an empirical policy collection $\widehat{\pi}_T$ after T rounds, without requiring T in advance. Each update uses m fresh comparisons, or none if the proposal equals the current response; thus $B := mT$ bounds the query count. Our goal is a small deployment-weighted gap $J(\pi^*) - J(\widehat{\pi}_T)$.

3 Single-prompt sampler

For a fixed prompt x, the *Repeated-Win Pairwise Sampler* (RWPS) proposes $s' \sim q_x$ and replaces its current response $C_x^{(i)}$ only if s' wins all m independent fresh comparisons (Algorithm 1). The resulting Markov chain's empirical response frequencies form the returned policy.

Algorithm 1 Repeated-Win Pairwise Sampler (RWPS)

Require: prompt x, current response $C_x^{(i)} \in \mathcal{S}_x$, reference policy q_x, confirmation level m

1: $C_x^{(i+1)} \leftarrow C_x^{(i)}$
2: Draw a proposal $s' \sim q_x$
3: **if** $s' \neq C_x^{(i)}$ **then**
4: Query $Y_x(s', C_x^{(i)})$ independently m times
5: **if** all m comparisons favor s' **then** $C_x^{(i+1)} \leftarrow s'$
6: **return** $C_x^{(i+1)}$

For distinct responses $s, s' \in \mathcal{S}_x$, the resulting transition probability is

$$P_x(s, s') = q_x(s')p_x(s', s)^m, \qquad P_x(s, s) = 1 - \sum_{u \in \mathcal{S}_x \setminus \{s\}} P_x(s, u). \tag{9}$$

Proposition 3.1 (Exact target policy)**.** *The Markov chain with transition kernel P_x has the unique stationary distribution π_x^*.*

Appendix C.1 proves this by establishing reversibility, irreducibility, and aperiodicity.

Returned policy. Starting from an arbitrary initial response $C_x^{(0)}$, let $(C_x^{(i)})_{i\geq 0}$ be the chain generated by Algorithm 1. After n local updates, we return

$$\widehat{\pi}_{x,n}(s) := \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{C_x^{(i)} = s\}, \qquad s \in \mathcal{S}_x. \tag{10}$$

The initial response $C_x^{(0)}$ is excluded from these frequencies. Every update contributes, including updates that leave the current response unchanged.

Local difficulty. Reversibility, irreducibility, and aperiodicity imply that the eigenvalues of P_x are real and satisfy

$$1 = \lambda_{x,1} > \lambda_{x,2} \geq \cdots \geq \lambda_{x,K_x} > -1.$$

Define the *local difficulty*

$$d_x := \frac{1}{2}\sum_{j=2}^{K_x} \frac{1+\lambda_{x,j}}{1-\lambda_{x,j}} > 0. \tag{11}$$

The local difficulty reflects both the number of candidate responses and the dependence between successive updates. For independent samples from π_x^*, all nontrivial eigenvalues would vanish, giving $d_x = (K_x - 1)/2$. Eigenvalues close to one increase d_x, reflecting slowly decaying dependence along the corresponding directions of the chain. Thus prompts with the same number of candidates can require different update budgets. The results below quantify this through the leading expected objective gap.

Finite-sample guarantee. Let κ_x be a finite constant depending only on P_x, accounting for finite-horizon and initialization effects; its explicit expression appears in (28) of Appendix C.2.

Theorem 3.2 (Single-prompt finite-sample objective gap)**.** *Let the distribution of $C_x^{(0)}$ be arbitrary. For every $n \geq 1$,*

$$J_x(\pi_x^*) - \mathbb{E}[J_x(\widehat{\pi}_{x,n})] \leq \widetilde{\tau}\left(\frac{2d_x}{n} + \frac{\kappa_x}{n^2}\right). \tag{12}$$

The expectation includes the initial draw, proposal draws, and comparison outcomes. Appendix C.2 proves the bound, including its n^{-2} correction for finite-horizon and initialization effects.

Asymptotic sharpness. The next result identifies the exact leading coefficient.

Theorem 3.3 (Asymptotic sharpness)**.** *Fix a prompt x and let $C_x^{(0)} \sim \mu_x$ have any initial distribution. Let $\widehat{\pi}_{x,n}$ be the occupation policy in (10), and let $d_x > 0$ be the local difficulty in (11). As $n \to \infty$,*

$$J_x(\pi_x^*) - \mathbb{E}[J_x(\widehat{\pi}_{x,n})] = \frac{\widetilde{\tau} d_x}{n} + O(n^{-3/2}). \tag{13}$$

The implied constant may depend on the fixed prompt and its initial distribution, but not on n. Thus n times the expected gap converges to $\widetilde{\tau} d_x > 0$.

The $1/n$ rate is tight. The factor-two slack in Theorem 3.2 comes from KL $\leq \chi^2$, whereas the quadratic expansion of KL has coefficient $1/2$ (proof in Appendix C.3). Section 4 uses these leading costs to allocate updates across prompts.

4 Adaptive allocation across prompts

We allocate local updates across prompts using their deployment weights f_x and local difficulties d_x, first with known difficulties and then with estimates learned from the observed transitions.

4.1 Oracle allocation with known local difficulties

Let $X_t \in \mathcal{X}$ denote the prompt selected at global round t, and let $N_x(t) := \sum_{u=1}^{t} \mathbf{1}\{X_u = x\}$ be the number of updates assigned to prompt x in the first t global rounds. Each global round advances

exactly one local chain. Thus, after T rounds, prompt x has received $N_x(T)$ local updates, with $\sum_{x\in\mathcal{X}} N_x(T) = T$.

First consider deterministic integer allocations $n_x \geq 1$ satisfying $\sum_{x\in\mathcal{X}} n_x = T$, and write $\widehat{\pi} := (\widehat{\pi}_{x,n_x})_{x\in\mathcal{X}}$. Applying Theorem 3.2 to each prompt and summing with the deployment weights gives

$$J(\pi^*) - \mathbb{E}[J(\widehat{\pi})] \leq \widetilde{\tau} \sum_{x\in\mathcal{X}} f_x \left(\frac{2d_x}{n_x} + \frac{\kappa_x}{n_x^2} \right). \tag{14}$$

The prompt-specific trajectories need not be independent.

To identify the optimal leading allocation, fix an interior proportion vector $\omega \in \Delta(\mathcal{X})$ and deterministic integer counts $n_x(T) = T\omega_x + O(1)$ with $\sum_x n_x(T) = T$. Theorem 3.3 gives

$$J(\pi^*) - \mathbb{E}[J(\widehat{\pi})] = \frac{\widetilde{\tau}}{T} \sum_{x\in\mathcal{X}} \frac{f_x d_x}{\omega_x} + o(T^{-1}).$$

The following allocation uniquely minimizes this leading coefficient.

Proposition 4.1 (Oracle allocation)**.** *Among all interior allocation proportions $\omega \in \Delta(\mathcal{X})$, the leading coefficient $\sum_{x\in\mathcal{X}} f_x d_x/\omega_x$ is uniquely minimized by*

$$\omega_x^{\mathrm{or}} := \frac{\sqrt{f_x d_x}}{\sum_{y\in\mathcal{X}} \sqrt{f_y d_y}}, \qquad n_x^{\mathrm{or}} := T\omega_x^{\mathrm{or}}. \tag{15}$$

Here n_x^{or} denotes the continuous target count for prompt x.

For deterministic integer allocations satisfying $n_x(T) = T\omega_x^{\mathrm{or}} + O(1)$,

$$J(\pi^*) - \mathbb{E}[J(\widehat{\pi})] = \frac{\widetilde{\tau}}{T} \left(\sum_{x\in\mathcal{X}} \sqrt{f_x d_x} \right)^2 + o(T^{-1}).$$

The finite-sample upper bound (14) has twice this leading coefficient and an $O(T^{-2})$ correction. Appendix D.1 proves Proposition 4.1.

4.2 Adaptive allocation with unknown local difficulties

We estimate each transition kernel and its difficulty, then track the resulting allocation while ensuring continued exploration.

Estimating the transition kernels. For a prompt x and local update count $k \geq 0$, define

$$V_{x,k}(s,s') := \sum_{i=0}^{k-1} \mathbf{1}\{C_x^{(i)} = s,\ C_x^{(i+1)} = s'\},$$
$$V_{x,k}(s) := \sum_{u\in\mathcal{S}_x} V_{x,k}(s,u).$$

Both counts are zero when $k = 0$. For a fixed pseudocount $\alpha > 0$, set

$$\widehat{P}_{x,k}(s,s') := \frac{V_{x,k}(s,s') + \alpha q_x(s')}{V_{x,k}(s) + \alpha}. \tag{16}$$

The full support of q_x makes this estimate positive and row-stochastic, including before any updates have been observed.

Estimating the local difficulties. The local difficulty admits the equivalent trace representation

$$d_x = \mathrm{tr}\left[\left(I - P_x + \mathbf{1}(\pi_x^*)^\top \right)^{-1} \right] - 1 - \frac{K_x - 1}{2}.$$

For a positive row-stochastic matrix P, let $\pi(P)$ denote its unique stationary distribution. Replacing P_x and π_x^* by the estimated kernel and its stationary distribution gives

$$\widehat{d}_{x,k} := \mathrm{tr}\left[\left(I - \widehat{P}_{x,k} + \mathbf{1}\pi(\widehat{P}_{x,k})^\top\right)^{-1}\right] - 1 - \frac{K_x - 1}{2}. \tag{17}$$

Positivity of $\widehat{P}_{x,k}$ makes this expression well defined, without requiring reversibility. The stationary distribution $\pi(\widehat{P}_{x,k})$ is used to estimate sampling difficulty; the returned response policy remains the empirical occupation distribution. At global time t, use the clipped estimate

$$\widehat{d}_x^+(t) := \max\{\widehat{d}_{x,N_x(t)}, (t+1)^{-1/2}\}. \tag{18}$$

This positivity floor makes the following plug-in allocation well defined:

$$\widehat{\omega}_x(t) := \frac{\sqrt{f_x \widehat{d}_x^+(t)}}{\sum_{u\in\mathcal{X}} \sqrt{f_u \widehat{d}_u^+(t)}}. \tag{19}$$

Ensuring exploration. Let $M := |\mathcal{X}|$. Mix the plug-in allocation with a vanishing uniform component:

$$\varepsilon_t := \frac{1}{\sqrt{t+1}}, \qquad \bar{\omega}_x(t) := (1-\varepsilon_t)\widehat{\omega}_x(t) + \frac{\varepsilon_t}{M}. \tag{20}$$

Unlike the positivity floor in (18), this mixture ensures $\bar{\omega}_x(t) \geq \varepsilon_t/M$ independently of the difficulty estimates, guaranteeing continued exploration. Its vanishing weight allows convergence to the oracle proportions.

Tracking the cumulative targets. Define the cumulative target

$$Q_x(t+1) := \sum_{u=0}^{t} \bar{\omega}_x(u).$$

Algorithm 2 selects the prompt with the largest deficit $Q_x(t+1) - N_x(t)$, breaking ties by any fixed rule, following the cumulative-tracking approach of Garivier & Kaufmann (2016).

Algorithm 2 Cumulative tracking for multiple prompts (CTM)

Require: $(\mathcal{S}_x, q_x, f_x)_{x\in\mathcal{X}}$, temperature τ, pseudocount $\alpha > 0$

$m \leftarrow \max\{1, \mathsf{round}(1/\tau)\}$
Choose arbitrary initial responses $C_x^{(0)} \in \mathcal{S}_x$
Set $N_x(0) \leftarrow 0$ and $Q_x(0) \leftarrow 0$ for every $x \in \mathcal{X}$
Set all transition counts to zero and initialize $\widehat{P}_{x,0}$ and $\widehat{d}_{x,0}$ for every $x \in \mathcal{X}$
for $t = 0, 1, \ldots$ until interruption **do**
  Form $\widehat{d}_x^+(t)$ and $\widehat{\omega}(t)$ using (18) and (19)
  Form $\bar{\omega}(t)$ using (20) and set $Q_x(t+1) \leftarrow Q_x(t) + \bar{\omega}_x(t)$ for every x
  Choose $X_{t+1} \in \arg\max_{x\in\mathcal{X}}\{Q_x(t+1) - N_x(t)\}$
  Run RWPS (Algorithm 1) for prompt X_{t+1} and record its transition
  Set $N_x(t+1) \leftarrow N_x(t) + \mathbf{1}\{X_{t+1} = x\}$ for every x
  Update $\widehat{P}_{X_{t+1},N_{X_{t+1}}(t+1)}$ and $\widehat{d}_{X_{t+1},N_{X_{t+1}}(t+1)}$; retain the other estimates
return $\widehat{\pi}_T$ at the final round T, as defined below

After T rounds, return the collection

$$\widehat{\pi}_T := \left(\widehat{\pi}_{x,N_x(T)}\right)_{x\in\mathcal{X}}, \qquad \widehat{\pi}_{x,0} := \delta_{C_x^{(0)}}.$$

Thus a prompt with no updates returns its initial response as a point mass.

4.3 GUARANTEES FOR CTM

We first establish consistency of the estimates and allocation proportions, then bound the objective gap of the returned policy.

Proposition 4.2 (Plug-in consistency)**.** *For every prompt x and every initial state,*

$$\widehat{P}_{x,k} \longrightarrow P_x, \qquad \widehat{d}_{x,k} \longrightarrow d_x \quad \text{almost surely as } k \to \infty.$$

The proof appears in Appendix D.2. Combined with continued exploration and cumulative tracking, this consistency yields the following allocation guarantee.

Lemma 4.3 (Infinite exploration and allocation consistency)**.** *Under Algorithm 2, $N_x(t) \to \infty$ for every prompt $x \in \mathcal{X}$ on every sample path. Moreover, almost surely and simultaneously for all $x \in \mathcal{X}$,*

$$\bar{\omega}_x(t) \longrightarrow \omega_x^{\mathrm{or}}, \qquad \frac{N_x(t)}{t} \longrightarrow \omega_x^{\mathrm{or}} \qquad (t \to \infty). \tag{21}$$

Appendix D.3 gives the proof.

Theorem 4.4 (Objective-gap guarantees)**.** *Under the standing assumptions, run Algorithm 2 from arbitrary initial responses and let $\widehat{\pi}_T$ be its returned policy at deterministic horizon T. As $T \to \infty$,*

$$\mathbb{E}[J(\pi^*) - J(\widehat{\pi}_T)] = O(T^{-1/2}), \qquad J(\pi^*) - J(\widehat{\pi}_T) = O_p(T^{-1}). \tag{22}$$

For fixed m and comparison budget $B = mT$, the corresponding rates are $O(B^{-1/2})$ and $O_p(B^{-1})$.

The in-probability rate matches the oracle's T^{-1} order; equality of the leading constants is not established.

Proof sketch. Forced exploration and bounded rounding discrepancy give $N_x(T) \geq c\sqrt{T}$ for some $c > 0$, every prompt, and all sufficiently large T, on every sample path. To handle the data-dependent counts, we use the uniform occupation bound of Lemma D.3 in Appendix D.4:

$$\mathbb{E}\left[\sup_{\ell \geq k} \chi^2(\widehat{\pi}_{x,\ell} \| \pi_x^*)\right] \leq \frac{C_x}{k}, \qquad k \geq 1.$$

Whenever $N_x(T) \geq k$, the occupation error at this random index is bounded by the supremum above. Applying this bound above the exploration floor, together with the aggregate gap identity and KL $\leq \chi^2$, gives the expected $O(T^{-1/2})$ bound. For any $0 < \nu < \min_x \omega_x^{\mathrm{or}}$, Lemma 4.3 gives $\min_x N_x(T) \geq \nu T$ with probability tending to one. The same uniform bound and Markov's inequality yield $O_p(T^{-1})$. Appendix D.4 gives the details. □

5 NUMERICAL EXPERIMENTS

Experimental questions and methods. We test whether adaptive allocation improves policy quality on a synthetic weighted hard instance and in an experiment on 30 SQLite questions. Comparing RWPS+CTM with RWPS+ROUND-ROBIN isolates the allocation rule while retaining the same local sampler and occupation-policy output. The synthetic experiment also includes RWPS+ORACLE-CTM, a benchmark using the true local difficulties. A third baseline, UNIFORM+ROUND-ROBIN, cycles through prompts and candidate pairs and returns a policy based on accumulated pairwise wins. It changes both sampling and the output policy, so it is not a pure allocation comparison. We use three replications, $m = 2$, and $\tau = \widetilde{\tau} = 0.5$. Within each experiment, methods receive equal numbers of updates: $T = 6000$ for the synthetic instance and $T = 5000$ for SQLite. The nominal sampling allowance is $B = 2T$, not an equal actual call count: RWPS self-proposals require no judgments, and evaluation calls are separate. Appendix E gives the detailed protocol.

Evaluation. For the synthetic instance, known utilities allow exact evaluation of $J(\pi^*) - J(\widehat{\pi}_T)$ under the objective in Section 2; smaller gaps are better. For SQLite, utilities are unavailable. For each evaluation trial, we select prompt x with probability f_x and independently sample one answer from each method's policy at that checkpoint. Identical answers receive a score of $1/2$ without an

evaluator call. Otherwise, an LLM evaluator compares the answers twice, with their display order reversed in the second comparison. The trial scores 1 if both judgments favor the first method, 0 if both favor the second, and $1/2$ if they disagree. The win ratio averages these scores, so values above 0.5 favor the first-named method. This preference metric does not measure the KL-regularized objective gap or directly validate its convergence rate. Figure 1 reports means across three replications, with descriptive 95% bootstrap bands summarizing between-replication variability.

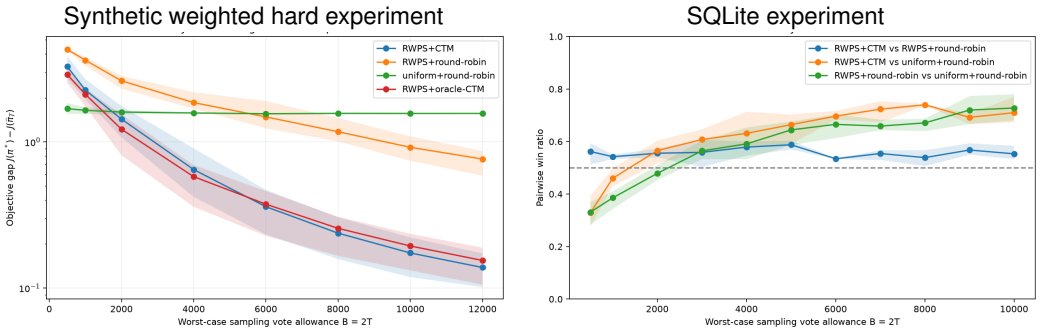

Figure 1: Synthetic weighted hard experiment (left: aggregate objective gap, logarithmic vertical axis) and SQLite experiment (right: pairwise win ratios; values above 0.5 favor the first-named method). Both horizontal axes show the nominal sampling allowance $B = 2T$, excluding evaluation calls. Shading shows descriptive 95% bootstrap intervals across three replications.

Synthetic results. At $T = 6000$ ($B = 12000$), the mean objective gap is 0.138 for RWPS+CTM, compared with 0.154 for RWPS+ORACLE-CTM and 0.763 for RWPS+ROUND-ROBIN. Learned and oracle CTM therefore have comparable observed gaps in these runs, while both improve on cyclic prompt allocation; these observations do not establish equivalence or superiority of learned CTM over the oracle. The mean gap of UNIFORM+ROUND-ROBIN is 1.566, but its policy based on accumulated wins does not directly target the KL-regularized objective. Table 2 in Appendix E.2 gives the final numerical summaries.

SQLite results and scope. At $T = 5000$ ($B = 10000$), RWPS+CTM has a mean win ratio of 0.553 against RWPS+ROUND-ROBIN. Against UNIFORM+ROUND-ROBIN, the respective RWPS win ratios are 0.710 and 0.728 (Table 3 in Appendix E.3). In this experiment, six of the 30 questions receive 60% of deployment weight, while answers labeled strong collectively receive 10% of the proposal probability within each question. This setup tests whether adaptive allocation helps when some questions matter more and strong answers are proposed infrequently. The observed advantage concerns this setting, not universal CTM superiority. The experiment uses fixed AI-generated candidate sets and the same model for sampling judgments and final evaluation, with only three replications.

6 CONCLUSION

We developed a direct method for constructing KL-regularized policies over finite response sets from pairwise feedback. RWPS has the implemented target policy as its stationary distribution, and its occupation policy admits both a finite-sample objective-gap bound and an asymptotically sharp $1/n$ expansion after n local updates, with leading coefficient $\widetilde{\tau} d_x$. The resulting local difficulties determine the square-root oracle allocation across prompts. CTM consistently learns this allocation online and achieves an expected $O(T^{-1/2})$ objective gap and an $O_p(T^{-1})$ rate at deterministic horizons of T global update rounds, matching the oracle's order in probability. Synthetic and SQLite experiments illustrate the benefits of adaptive allocation in the settings studied.

The analysis assumes fixed finite candidate sets and fresh conditional Bradley–Terry comparisons. Open theoretical questions include matching the oracle's expected $O(T^{-1})$ gap and leading constant, and obtaining guarantees at data-dependent stopping times. Practical directions include estimating difficulties without repeated full matrix inversions, growing candidate sets during sampling, and robustness to preference-model misspecification.

AI Use Statement

We used generative AI tools to assist with drafting and revising manuscript sections, organizing the exposition, developing and checking mathematical proof arguments, experimental code development, and literature search and summarization. In the SQLite experiment, LLMs also generated and reviewed candidate questions and answers and supplied pairwise judgments, as described in Appendix E.3. The authors are responsible for the accuracy and integrity of all claims, code, and results, including those developed with AI assistance.

Ethics Statement

The reported experiments use synthetic preferences and model-generated judgments over AI-generated candidate answers; they involve no human participants or human-subject data. Model-generated judgments may reflect systematic errors or biases and should not be treated as validated human preferences. Future applications using human feedback should address informed consent, privacy, and bias mitigation.

Reproducibility Statement

Sections 2–4 specify the formal setting, assumptions, and algorithms, and Appendices B–D provide the proofs. Appendix E describes the experimental constructions, methods, evaluation metrics, and uncertainty summaries. The planned anonymous supplement will include experiment code, frozen input pools, configurations, seeds and environment metadata, checkpoints, sampling and evaluation records, and plotting scripts.

References

A. A. Barker. Monte carlo calculations of the radial distribution functions for a proton–electron plasma. *Australian Journal of Physics*, 18(2):119–133, 1965.

Ralph A. Bradley and Milton E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

Daniele Calandriello, Zhaohan Daniel Guo, Rémi Munos, Mark Rowland, Yunhao Tang, Bernardo Avila Pires, Pierre Harvey Richemond, et al. Human alignment of large language models through online preference optimisation. In *Proceedings of ICML*, 2024.

Chun-Hung Chen, Jianwu Lin, Enver Yücesan, and Stephen E. Chick. Simulation budget allocation for further enhancing the efficiency of ordinal optimization. *Discrete Event Dynamic Systems*, 10: 251–270, 2000.

Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 4299–4307, 2017.

Miroslav Dudík, Katja Hofmann, Robert E. Schapire, Aleksandrs Slivkins, and Masrour Zoghi. Contextual dueling bandits. In *Proceedings of the 28th Conference on Learning Theory (COLT)*, volume 40 of *Proceedings of Machine Learning Research*, pp. 563–587. PMLR, 2015.

Dimitris Fotakis, Alkis Kalavasis, and Christos Tzamos. Perfect sampling from pairwise comparisons. In *Advances in Neural Information Processing Systems*, 2022.

Aurélien Garivier and Emilie Kaufmann. Optimal best arm identification with fixed confidence. In *Proceedings of the 29th Conference on Learning Theory*, pp. 998–1027, 2016.

Peter Hall and C. C. Heyde. *Martingale Limit Theory and Its Application*. Academic Press, 1980. doi: 10.1016/C2013-0-10818-5.

Rémi Munos, Michal Valko, Daniele Calandriello, Mohammad Gheshlaghi Azar, Mark Rowland, Zhaohan Daniel Guo, Yunhao Tang, Matthieu Geist, et al. Nash learning from human feedback. In *Proceedings of ICML*, 2024.

Sahand Negahban, Sewoong Oh, and Devavrat Shah. Rank centrality: Ranking from pairwise comparisons. *Operations Research*, 65(1):266–287, 2017.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, 2023.

Adam N. Sanborn and Thomas L. Griffiths. Markov chain monte carlo with people. In *Advances in Neural Information Processing Systems*, 2007.

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 ICML*, 2024.

Yisong Yue, Josef Broder, Robert Kleinberg, and Thorsten Joachims. The K-armed dueling bandits problem. *Journal of Computer and System Sciences*, 78(5):1538–1556, 2012.

A NOTATION

The table summarizes the main notation; auxiliary quantities used only within individual proofs are defined where they appear.

Symbol	Meaning
Setting: prompts, responses, and target policies	
$\mathcal{X},\ M = \lvert\mathcal{X}\rvert$	Finite prompt set and its size.
$\mathcal{S}_x = [K_x]$	Candidate-response set for prompt x, with $K_x \geq 2$.
f_x	Positive deployment weight, with $\sum_x f_x = 1$.
$r_x(s)$	Latent utility of response s for prompt x.
q_x	Reference policy with full support on $\mathcal{S}_x$.
$\tau,\ \widetilde{\tau} = 1/m$	Requested and implemented KL-regularization temperatures, respectively.
m	Confirmation level $m = \max\{1, \mathsf{round}(1/\tau)\}$; all m comparisons must favor the proposal for a move. Each update uses at most m queries.
$\pi_x^{(\tau)}$	Target policy at the requested temperature τ; see (2).
$\pi_x^*,\ \pi^*$	Optimal policy at $\widetilde{\tau}$ and the collection $(\pi_x^*)_{x\in\mathcal{X}}$. Each π_x^* is also the stationary distribution of P_x.
$J_x,\ J$	Per-prompt objective at $\widetilde{\tau}$ and its aggregate $J(\pi) = \sum_x f_x J_x(\pi_x)$.
Pairwise feedback and local sampler	
$Y_x(s, s')$	Bernoulli comparison label; success means that response s is preferred to s'.
$p_x(s, s')$	Bradley–Terry preference probability $\Pr(Y_x(s, s') = 1)$.
$C_x^{(i)}$	Response state after i local updates; $C_x^{(0)}$ is the initial state.
P_x	Transition kernel of the repeated-win update; see (9).
$\lambda_{x,j}$	Eigenvalues of P_x, ordered as $1 = \lambda_{x,1} > \lambda_{x,2} \geq \cdots \geq \lambda_{x,K_x} > -1$.
d_x	Local difficulty determining the leading expected occupation-policy error; see (11).

Continued on the next page

Symbol	Meaning (continued)
γ_x	Absolute spectral gap $1 - \max_{j\geq 2} \lvert\lambda_{x,j}\rvert$.
$\pi^*_{x,\min}$	Smallest stationary probability $\min_{s\in\mathcal{S}_x} \pi^*_x(s)$.
κ_x	Finite-horizon and initialization correction constant in (28).
Update counts, budgets, and empirical policies	
$k,\ n$	Local update count; n denotes the horizon in the single-prompt guarantees.
$t,\ T$	Global round index and total number of global rounds.
X_t	Prompt selected at global round t.
$N_x(t)$	Number of updates assigned to x: $\sum_{u=1}^{t} \mathbf{1}\{X_u = x\}$.
$B = mT$	Deterministic worst-case comparison budget; the realized query count is at most B.
n_x	Fixed number of updates allocated to x in the multi-prompt benchmark.
$\widehat{\pi}_{x,n}(s)$	Empirical occupation probability $n^{-1}\sum_{i=1}^{n} \mathbf{1}\{C_x^{(i)} = s\}$, averaging states after updates.
$\widehat{\pi}_T$	Returned policy collection. Its prompt-x component is $\widehat{\pi}_{x,N_x(T)}$ if $N_x(T) \geq 1$, and $\delta_{C_x^{(0)}}$ otherwise.
Estimation and adaptive allocation	
$V_{x,k}(s,s')$	Transition count $\sum_{i=0}^{k-1} \mathbf{1}\{C_x^{(i)} = s,\ C_x^{(i+1)} = s'\}$.
$V_{x,k}(s)$	Count of transition starting states: $\sum_{i=0}^{k-1} \mathbf{1}\{C_x^{(i)} = s\} = \sum_{s'} V_{x,k}(s,s')$.
α	Positive pseudocount used in the transition estimator.
$\pi(P)$	Unique stationary distribution of a positive row-stochastic matrix P.
$\widehat{P}_{x,k}$	Pseudocount-smoothed transition estimate after k local updates.
$\widehat{d}_{x,k}$	Plug-in estimate of the local difficulty d_x.
$\widehat{d}^{+}_{x}(t)$	Positive safeguard $\max\{\widehat{d}_{x,N_x(t)}, (t+1)^{-1/2}\}$.
ω_x^{or}	Oracle allocation proportion, proportional to $\sqrt{f_x d_x}$.
$\widehat{\omega}_x(t)$	Plug-in allocation proportion, proportional to $\sqrt{f_x \widehat{d}^{+}_{x}(t)}$.
ε_t	Vanishing uniform-exploration weight $(t+1)^{-1/2}$.
$\bar{\omega}_x(t)$	Allocation target with uniform exploration: $(1-\varepsilon_t)\widehat{\omega}_x(t) + \varepsilon_t/M$.
$Q_x(t)$	Cumulative target $\sum_{u=0}^{t-1} \bar{\omega}_x(u)$ used for deterministic rounding, with $Q_x(0) = 0$.

B Proofs for the Objective and Temperature-Discretization Results

B.1 Objective-gap identity

For convenience, write

$$Z_x(m) := \sum_{u\in\mathcal{S}_x} q_x(u) e^{m r_x(u)}.$$

Proposition 2.1 (restated). *For every prompt $x \in \mathcal{X}$, the policy*

$$\pi^*_x(s) := \frac{q_x(s) e^{m r_x(s)}}{Z_x(m)}$$

is the unique maximizer of J_x over $\Delta(\mathcal{S}_x)$. Moreover, for every $\pi_x \in \Delta(\mathcal{S}_x)$,

$$J_x(\pi^*_x) - J_x(\pi_x) = \widetilde{\tau}\,\mathrm{KL}(\pi_x \| \pi^*_x).$$

Consequently, $\pi^* := (\pi_x^*)_{x\in\mathcal{X}}$ is the unique maximizer of J, and

$$J(\pi^*) - J(\pi) = \widetilde{\tau} \sum_{x\in\mathcal{X}} f_x \,\mathrm{KL}(\pi_x \| \pi_x^*).$$

The proof rewrites the local objective as its optimum value minus a KL divergence, and then averages the result over prompts.

Proof. Fix a prompt x and an arbitrary policy $\pi_x \in \Delta(\mathcal{S}_x)$. Because q_x has full support, π_x^* also has full support, so $\mathrm{KL}(\pi_x \| \pi_x^*)$ is finite. Taking logarithms in the displayed definition of π_x^* gives, for every $s \in \mathcal{S}_x$,

$$\log \frac{\pi_x^*(s)}{q_x(s)} = m r_x(s) - \log Z_x(m).$$

Since $m = 1/\widetilde{\tau}$, solving for $r_x(s)$ yields

$$r_x(s) = \widetilde{\tau} \log \frac{\pi_x^*(s)}{q_x(s)} + \widetilde{\tau} \log Z_x(m). \tag{23}$$

Substitute (23) into the definition of J_x in (3) and use $\sum_{s\in\mathcal{S}_x} \pi_x(s) = 1$. The two entropy terms then combine as follows:

$$\begin{aligned} J_x(\pi_x) &= \widetilde{\tau} \log Z_x(m) + \widetilde{\tau} \sum_{s\in\mathcal{S}_x} \pi_x(s) \left(\log \frac{\pi_x^*(s)}{q_x(s)} - \log \frac{\pi_x(s)}{q_x(s)} \right) \\ &= \widetilde{\tau} \log Z_x(m) - \widetilde{\tau} \sum_{s\in\mathcal{S}_x} \pi_x(s) \log \frac{\pi_x(s)}{\pi_x^*(s)} \\ &= \widetilde{\tau} \log Z_x(m) - \widetilde{\tau} \,\mathrm{KL}(\pi_x \| \pi_x^*). \end{aligned}$$

Terms with $\pi_x(s) = 0$ contribute zero under the KL convention in Section 2.

Setting $\pi_x = \pi_x^*$ makes the KL term vanish, so

$$J_x(\pi_x^*) = \widetilde{\tau} \log Z_x(m).$$

Subtracting the expression for $J_x(\pi_x)$ gives (5). Since KL divergence is nonnegative and is zero only when its two arguments are equal, $J_x(\pi_x) \le J_x(\pi_x^*)$, with equality only for $\pi_x = \pi_x^*$. Thus π_x^* is the unique local maximizer.

Finally, for $\pi = (\pi_x)_{x\in\mathcal{X}}$, summing the local identity with the deployment weights gives

$$\begin{aligned} J(\pi^*) - J(\pi) &= \sum_{x\in\mathcal{X}} f_x \big[J_x(\pi_x^*) - J_x(\pi_x) \big] \\ &= \widetilde{\tau} \sum_{x\in\mathcal{X}} f_x \,\mathrm{KL}(\pi_x \| \pi_x^*), \end{aligned}$$

which is (6). Every summand is nonnegative and every f_x is strictly positive. Hence the aggregate gap is zero exactly when $\pi_x = \pi_x^*$ for every x, proving the uniqueness of the aggregate maximizer and completing the proof of Proposition 2.1. □

B.2 Temperature discretization

Proposition 2.2 (restated). *Let $R_x := \max_{s\in\mathcal{S}_x} r_x(s) - \min_{s\in\mathcal{S}_x} r_x(s)$. Then*

$$\max \left\{ \mathrm{KL}\left(\pi_x^* \middle\| \pi_x^{(\tau)}\right), \mathrm{KL}\left(\pi_x^{(\tau)} \middle\| \pi_x^*\right) \right\} \le \left| m - \frac{1}{\tau} \right| R_x,$$

and

$$\left\| \pi_x^* - \pi_x^{(\tau)} \right\|_{\mathrm{TV}} \le \sqrt{\frac{R_x}{2} \left| m - \frac{1}{\tau} \right|}.$$

The two policies coincide when $1/\tau$ is an integer.

The proof first controls the partition functions, then the two KL divergences, and finally applies Pinsker's inequality.

Proof. Fix x. For any real a, write

$$\pi_{x,a}(s) = \frac{q_x(s)e^{ar_x(s)}}{Z_x(a)}, \qquad Z_x(a) = \sum_{u\in\mathcal{S}_x} q_x(u)e^{ar_x(u)}.$$

Partition-function bound. Direct substitution gives the partition-function ratio

$$\frac{Z_x(\frac{1}{\widetilde{\tau}})}{Z_x(\frac{1}{\tau})} = \sum_{s\in\mathcal{S}_x} \frac{q_x(s)e^{\frac{r_x(s)}{\tau}}}{Z_x(\frac{1}{\tau})} e^{(\frac{1}{\widetilde{\tau}}-\frac{1}{\tau})r_x(s)} = \mathbb{E}_{s\sim\pi_{x,\frac{1}{\tau}}}\left[e^{(\frac{1}{\widetilde{\tau}}-\frac{1}{\tau})r_x(s)}\right]. \tag{24}$$

The values $(\frac{1}{\widetilde{\tau}}-\frac{1}{\tau})r_x(s)$ range from

$$\min_{u\in\mathcal{S}_x}\left(\frac{1}{\widetilde{\tau}}-\frac{1}{\tau}\right)r_x(u) \quad \text{to} \quad \max_{u\in\mathcal{S}_x}\left(\frac{1}{\widetilde{\tau}}-\frac{1}{\tau}\right)r_x(u).$$

Exponentiating these two extrema bounds the integrand in (24) pointwise. Taking expectations and logarithms therefore gives

$$\min_{u\in\mathcal{S}_x}\left(\frac{1}{\widetilde{\tau}}-\frac{1}{\tau}\right)r_x(u) \le \log\frac{Z_x(\frac{1}{\widetilde{\tau}})}{Z_x(\frac{1}{\tau})} \le \max_{u\in\mathcal{S}_x}\left(\frac{1}{\widetilde{\tau}}-\frac{1}{\tau}\right)r_x(u). \tag{25}$$

The length of this interval is $|\frac{1}{\widetilde{\tau}}-\frac{1}{\tau}|R_x$.

KL bounds. For every state s,

$$\log\frac{\pi_{x,\frac{1}{\widetilde{\tau}}}(s)}{\pi_{x,\frac{1}{\tau}}(s)} = (\frac{1}{\widetilde{\tau}}-\frac{1}{\tau})r_x(s) - \log\frac{Z_x(\frac{1}{\widetilde{\tau}})}{Z_x(\frac{1}{\tau})}.$$

Averaging under $\pi_{x,\frac{1}{\widetilde{\tau}}}$ gives

$$\mathrm{KL}(\pi_{x,\frac{1}{\widetilde{\tau}}}\|\pi_{x,\frac{1}{\tau}}) = (\frac{1}{\widetilde{\tau}}-\frac{1}{\tau})\mathbb{E}_{s\sim\pi_{x,\frac{1}{\widetilde{\tau}}}}[r_x(s)] - \log\frac{Z_x(\frac{1}{\widetilde{\tau}})}{Z_x(\frac{1}{\tau})}. \tag{26}$$

The first term on the right is an average of the same values and hence lies between their minimum and maximum. The logarithmic partition ratio belongs to the same interval by (25). Their difference is nonnegative because it is a KL divergence and is at most the interval length $|\frac{1}{\widetilde{\tau}}-\frac{1}{\tau}|R_x$. Interchanging $\frac{1}{\tau}$ and $\frac{1}{\widetilde{\tau}}$ gives the same upper bound for the reverse KL divergence. Since $\pi_x^* = \pi_{x,\frac{1}{\widetilde{\tau}}}$ and $\pi_x^{(\tau)} = \pi_{x,\frac{1}{\tau}}$, this proves (7).

Total-variation bound. Finally, total variation is $\|\xi-\zeta\|_{\mathrm{TV}} := \frac{1}{2}\sum_s|\xi(s)-\zeta(s)|$. Pinsker's inequality states $\|\xi-\zeta\|_{\mathrm{TV}} \le \sqrt{\mathrm{KL}(\xi\|\zeta)/2}$. Applying it with either of the two KL bounds above proves (8). If $\frac{1}{\widetilde{\tau}}-\frac{1}{\tau}=0$, then the two exponential tilts are identical, which proves the final assertion. □

C Proofs for the Local Sampler

Appendix C.1 establishes the stationary law of the local sampler. Appendix C.2 proves the finite-sample objective-gap bound, and Appendix C.3 identifies its asymptotically exact leading coefficient.

C.1 Establishing the stationary law of the local sampler

Proposition 3.1 (restated). *The Markov chain with transition kernel P_x has the unique stationary distribution π_x^*.*

Proof of Proposition 3.1. We establish stationarity and uniqueness by proving that P_x is reversible and irreducible. We also verify aperiodicity for the subsequent spectral arguments. Fix a prompt x and write

$$Z_x(m) := \sum_{u\in\mathcal{S}_x} q_x(u)e^{mr_x(u)}.$$

Reversibility. For two distinct states $s, s' \in \mathcal{S}_x$, the transition rule (9), the target distribution (4), and the Bradley–Terry model (1) give

$$\begin{aligned}\pi_x^*(s)P_x(s,s') &= \frac{q_x(s)e^{mr_x(s)}}{Z_x(m)} q_x(s')p_x(s',s)^m \\ &= \frac{q_x(s)q_x(s')}{Z_x(m)} \left(\frac{e^{r_x(s)}e^{r_x(s')}}{e^{r_x(s)}+e^{r_x(s')}}\right)^m .\end{aligned}$$

The last expression is symmetric in s and s'. It therefore equals $\pi_x^*(s')P_x(s',s)$, which is the detailed-balance identity. Summing detailed balance over the first coordinate shows that $(\pi_x^*)^\top P_x = (\pi_x^*)^\top$; hence π_x^* is stationary and P_x is reversible with respect to it.

Irreducibility. Full support of q_x gives $q_x(s') > 0$, and every Bradley–Terry probability lies strictly between zero and one. Thus $P_x(s,s') > 0$ whenever $s \neq s'$. Every state can therefore reach every other state in one step, so the finite chain is irreducible. Its stationary distribution is therefore unique.

Aperiodicity. For every $s \in \mathcal{S}_x$,

$$\sum_{s' \neq s} P_x(s,s') = \sum_{s' \neq s} q_x(s')p_x(s',s)^m \leq \sum_{s' \neq s} q_x(s') = 1 - q_x(s).$$

It follows that $P_x(s,s) \geq q_x(s) > 0$ for every s. An irreducible chain with a positive self-loop has period one, so the chain is aperiodic. □

C.2 Bounding the finite-sample occupation-policy error

Theorem 3.2 (restated). *For a fixed prompt x, an arbitrary initial distribution, and every integer $n \geq 1$,*

$$J_x(\pi_x^*) - \mathbb{E}[J_x(\widehat{\pi}_{x,n})] \leq \widetilde{\tau}\left(\frac{2d_x}{n} + \frac{\kappa_x}{n^2}\right).$$

Proof of Theorem 3.2. Lemmas C.1, C.2, and C.3, stated and proved below, give

$$\mathbb{E}[\chi^2(\widehat{\pi}_{x,n} \| \pi_x^*)] \leq \frac{2d_x}{n} + \frac{\kappa_x}{n^2}, \tag{27}$$

where κ_x is defined in (28). For any distribution ξ, the inequality $\log z \leq z - 1$ gives

$$\mathrm{KL}(\xi \| \pi_x^*) \leq \sum_s \xi(s)\left(\frac{\xi(s)}{\pi_x^*(s)} - 1\right) = \chi^2(\xi \| \pi_x^*),$$

with zero coordinates contributing zero. Proposition 2.1 therefore yields

$$\begin{aligned}J_x(\pi_x^*) - \mathbb{E}[J_x(\widehat{\pi}_{x,n})] &= \widetilde{\tau}\,\mathbb{E}[\mathrm{KL}(\widehat{\pi}_{x,n} \| \pi_x^*)] \\ &\leq \widetilde{\tau}\,\mathbb{E}[\chi^2(\widehat{\pi}_{x,n} \| \pi_x^*)] \\ &\leq \widetilde{\tau}\left(\frac{2d_x}{n} + \frac{\kappa_x}{n^2}\right).\end{aligned}$$

This proves the theorem. □

Roadmap for the supporting lemmas. Lemma C.1 separates the occupation error as

$$\mathbb{E}[\chi^2(\widehat{\pi}_{x,n} \| \pi_x^*)] = \underbrace{\mathcal{S}_{x,n}}_{\text{stationary contribution}} + \underbrace{\mathcal{T}_{x,n}}_{\text{initialization correction}} .$$

Lemmas C.2 and C.3 then establish, respectively,

$$\mathcal{S}_{x,n} = \frac{2d_x}{n} + O(n^{-2}), \qquad \mathcal{T}_{x,n} = O(n^{-2}),$$

with explicit constants whose sum is κ_x.

Common notation. Define the absolute spectral gap and minimum stationary mass by

$$\gamma_x := 1 - \max_{j \geq 2} |\lambda_{x,j}| > 0, \qquad \pi^*_{x,\min} := \min_{s \in \mathcal{S}_x} \pi^*_x(s).$$

The finite-horizon and initialization corrections give the constant

$$\kappa_x := 4 \sum_{j=2}^{K_x} \frac{|\lambda_{x,j}|}{|1 - \lambda_{x,j}|^2} + (K_x - 1) \left(\frac{1}{\pi^*_{x,\min}} - 1 \right) \frac{2 - \gamma_x}{\gamma_x^2}. \tag{28}$$

Fix x and write $P := P_x$, $\pi := \pi^*_x$, $K := K_x$, $C_i := C_x^{(i)}$, and $\lambda_j := \lambda_{x,j}$. Let μ_i be the law of C_i. We allow μ_1 to be arbitrary, which covers every initialization of C_0; $\mathbb{E}_{\mu_1}$ denotes expectation for the chain starting with this law at time one. For functions on $\mathcal{S}_x$, use

$$\langle u, v \rangle_\pi := \sum_s \pi(s) u(s) v(s), \qquad \|u\|^2_{2,\pi} := \langle u, u \rangle_\pi.$$

For a distribution ξ, the pointwise ratio ξ/π gives

$$\chi^2(\xi \| \pi) := \sum_s \frac{(\xi(s) - \pi(s))^2}{\pi(s)} = \left\| \frac{\xi}{\pi} - 1 \right\|^2_{2,\pi}.$$

Lemma C.1 (Spectral occupation decomposition)**.** *Let $\phi_1, \ldots, \phi_K$ be an $L^2(\pi)$-orthonormal eigenbasis of P with $\phi_1 \equiv 1$ and $P\phi_j = \lambda_j \phi_j$. For $j \geq 2$ and $i \geq 1$, define $\Delta_{j,i} := \mathbb{E}_{\mu_1}[\phi_j(C_i)^2] - 1$. Then, for every $n \geq 1$,*

$$\mathbb{E}_{\mu_1}[\chi^2(\widehat{\pi}_{x,n} \| \pi)] = \mathcal{S}_{x,n} + \mathcal{T}_{x,n}, \tag{29}$$

where

$$\mathcal{S}_{x,n} := \frac{K-1}{n} + \frac{2}{n^2} \sum_{\ell=1}^{n-1} (n - \ell) \sum_{j=2}^{K} \lambda_j^\ell,$$

$$\mathcal{T}_{x,n} := \frac{1}{n^2} \sum_{j=2}^{K} \sum_{i=1}^{n} \Delta_{j,i} \left(1 + 2 \sum_{\ell=1}^{n-i} \lambda_j^\ell \right).$$

Proof. We obtain the decomposition from the spectral representation of the occupation error and the lagged second moments of the eigenfunctions.

Step 1: Spectral representation. Reversibility ensures that the stated eigenbasis exists. For $j \geq 2$, orthonormality gives $\mathbb{E}_\pi[\phi_j] = 0$ and $\mathbb{E}_\pi[\phi_j^2] = 1$. The function $g_n := \widehat{\pi}_{x,n}/\pi - 1$ is orthogonal to ϕ_1, and its remaining coefficients are

$$\langle g_n, \phi_j \rangle_\pi = \sum_s \big(\widehat{\pi}_{x,n}(s) - \pi(s)\big) \phi_j(s) = \frac{1}{n} \sum_{i=1}^{n} \phi_j(C_i).$$

Parseval's identity therefore gives

$$\chi^2(\widehat{\pi}_{x,n} \| \pi) = \sum_{j=2}^{K} \left(\frac{1}{n} \sum_{i=1}^{n} \phi_j(C_i) \right)^2. \tag{30}$$

Step 2: Lagged second moments. For $1 \leq i \leq k \leq n$, the Markov property and eigenvalue equation give

$$\mathbb{E}[\phi_j(C_k) \mid C_i] = (P^{k-i}\phi_j)(C_i) = \lambda_j^{k-i} \phi_j(C_i).$$

Multiplying by $\phi_j(C_i)$ and taking expectations yields

$$\mathbb{E}_{\mu_1}[\phi_j(C_i)\phi_j(C_k)] = \lambda_j^{k-i}(1 + \Delta_{j,i}). \tag{31}$$

Step 3: Separate the stationary and initialization terms. Expand the squares in (30) and apply (31) to obtain

$$\mathbb{E}_{\mu_1}[\chi^2(\widehat{\pi}_{x,n}\|\pi)] = \frac{1}{n^2}\sum_{j=2}^{K}\sum_{i=1}^{n}(1+\Delta_{j,i})\left(1+2\sum_{\ell=1}^{n-i}\lambda_j^\ell\right).$$

The terms containing $\Delta_{j,i}$ equal $\mathcal{T}_{x,n}$. For the remaining terms, each lag ℓ occurs $n-\ell$ times, giving $\mathcal{S}_{x,n}$ and proving (29). □

Lemma C.2 (Stationary spectral contribution)**.** *For every $n \geq 1$,*

$$\left|\mathcal{S}_{x,n} - \frac{2d_x}{n}\right| \leq \frac{4}{n^2}\sum_{j=2}^{K}\frac{|\lambda_j|}{(1-\lambda_j)^2}.$$

Proof. For any $\lambda \neq 1$, summing the finite geometric series gives

$$\frac{1}{n} + \frac{2}{n^2}\sum_{\ell=1}^{n-1}(n-\ell)\lambda^\ell = \frac{1+\lambda}{1-\lambda}\frac{1}{n} - \frac{2\lambda(1-\lambda^n)}{(1-\lambda)^2 n^2}. \tag{32}$$

Sum over $\lambda = \lambda_j$, $j = 2, \ldots, K$, and use $2d_x = \sum_{j=2}^{K}(1+\lambda_j)/(1-\lambda_j)$ to obtain

$$\mathcal{S}_{x,n} = \frac{2d_x}{n} + \mathcal{R}_{x,n}, \qquad \mathcal{R}_{x,n} := -\frac{2}{n^2}\sum_{j=2}^{K}\frac{\lambda_j(1-\lambda_j^n)}{(1-\lambda_j)^2}.$$

Since $|\lambda_j| < 1$, the remainder satisfies

$$|\mathcal{R}_{x,n}| \leq \frac{2}{n^2}\sum_{j=2}^{K}\frac{|\lambda_j|\,|1-\lambda_j^n|}{(1-\lambda_j)^2} \leq \frac{4}{n^2}\sum_{j=2}^{K}\frac{|\lambda_j|}{(1-\lambda_j)^2}, \tag{33}$$

which proves the lemma. □

Lemma C.3 (Initialization transient)**.** *For every $n \geq 1$,*

$$|\mathcal{T}_{x,n}| \leq \frac{(K_x-1)(\pi_{x,\min}^{*-1}-1)(2-\gamma_x)}{n^2\gamma_x^2}.$$

Proof. We first bound the initialization deviations and then sum their contribution to the occupation error.

Step 1: Bound the initialization deviations. Define $g_1 := \mu_1/\pi - 1$. Reversibility implies that the density deviation at time i is $P^{i-1}g_1$, so

$$\Delta_{j,i} = \mathbb{E}_{\mu_i}[\phi_j^2] - \mathbb{E}_\pi[\phi_j^2] = \langle P^{i-1}g_1, \phi_j^2 - 1\rangle_\pi. \tag{34}$$

Both functions have π-mean zero. Since $\gamma_x = 1 - \max_{j\geq 2}|\lambda_j|$, every mean-zero function h satisfies

$$\|P^r h\|_{2,\pi} \leq (1-\gamma_x)^r\|h\|_{2,\pi}, \qquad r \geq 0. \tag{35}$$

Thus Cauchy–Schwarz gives

$$|\Delta_{j,i}| \leq (1-\gamma_x)^{i-1}\|g_1\|_{2,\pi}\,\|\phi_j^2 - 1\|_{2,\pi}.$$

The two norms satisfy

$$\|g_1\|_{2,\pi}^2 = \sum_s \frac{\mu_1(s)^2}{\pi(s)} - 1 \leq \frac{1}{\pi_{x,\min}^*} - 1,$$

using $\sum_s \mu_1(s)^2 \leq 1$, and

$$\|\phi_j^2 - 1\|_{2,\pi}^2 = \mathbb{E}_\pi[\phi_j^4] - 1 \leq \|\phi_j\|_\infty^2\mathbb{E}_\pi[\phi_j^2] - 1 \leq \frac{1}{\pi_{x,\min}^*} - 1,$$

using $\mathbb{E}_\pi[\phi_j^2] = 1$ and $\pi(s)\phi_j(s)^2 \leq 1$. Consequently,

$$|\Delta_{j,i}| \leq \left(\frac{1}{\pi_{x,\min}^*} - 1\right)(1-\gamma_x)^{i-1}. \tag{36}$$

Step 2: Sum the initialization contribution. The lag weights satisfy

$$\left|1 + 2\sum_{\ell=1}^{n-i} \lambda_j^\ell\right| \le 1 + 2\sum_{\ell\ge 1}(1-\gamma_x)^\ell = \frac{2-\gamma_x}{\gamma_x}.$$

Insert this and (36) into the definition of $\mathcal{T}_{x,n}$ in Lemma C.1. Summing $\sum_{i\ge 1}(1-\gamma_x)^{i-1} = 1/\gamma_x$ gives

$$|\mathcal{T}_{x,n}| \le \frac{(K_x-1)(\pi_{x,\min}^{*-1}-1)(2-\gamma_x)}{n^2\gamma_x^2}, \tag{37}$$

as claimed. □

C.3 Proof of asymptotic sharpness of the single-prompt objective gap

Theorem 3.3 (restated). *Fix a prompt x and let $C_x^{(0)} \sim \mu_x$ have any initial distribution. Let $\widehat{\pi}_{x,n}$ be the occupation policy in (10), and let $d_x > 0$ be the local difficulty in (11). As $n \to \infty$,*

$$J_x(\pi_x^*) - \mathbb{E}[J_x(\widehat{\pi}_{x,n})] = \frac{\widetilde{\tau} d_x}{n} + O(n^{-3/2}).$$

The implied constant may depend on the fixed prompt and its initial distribution, but not on n. Thus n times the expected gap converges to $\widetilde{\tau} d_x > 0$.

Proof of Theorem 3.3. We combine the two-sided chi-square estimate from Appendix C.2 with the KL remainder bound proved below.

By Lemma C.1 in Appendix C.2,

$$\mathbb{E}_{\mu_x}[\chi^2(\widehat{\pi}_{x,n}\|\pi_x^*)] = \mathcal{S}_{x,n} + \mathcal{T}_{x,n}.$$

Lemmas C.2 and C.3 in the same appendix bound $|\mathcal{S}_{x,n} - 2d_x/n|$ and $|\mathcal{T}_{x,n}|$, respectively. Adding these absolute bounds gives

$$\left|\mathbb{E}_{\mu_x}[\chi^2(\widehat{\pi}_{x,n}\|\pi_x^*)] - \frac{2d_x}{n}\right| \le \frac{\kappa_x}{n^2}. \tag{38}$$

These estimates allow any law of $C_x^{(1)}$, including the law induced by the arbitrary initialization $C_x^{(0)} \sim \mu_x$.

Lemma C.4, stated and proved in Appendix C.3.1 below, then yields

$$\left|\mathbb{E}_{\mu_x}[\mathrm{KL}(\widehat{\pi}_{x,n}\|\pi_x^*)] - \frac{1}{2}\mathbb{E}_{\mu_x}[\chi^2(\widehat{\pi}_{x,n}\|\pi_x^*)]\right| \le \frac{C_{\mathrm{ent},x}}{n^{3/2}}.$$

Combining this with (38), we obtain

$$\left|\mathbb{E}_{\mu_x}[\mathrm{KL}(\widehat{\pi}_{x,n}\|\pi_x^*)] - \frac{d_x}{n}\right| \le \frac{\kappa_x}{2n^2} + \frac{C_{\mathrm{ent},x}}{n^{3/2}}. \tag{39}$$

Finally, Proposition 2.1 gives

$$J_x(\pi_x^*) - \mathbb{E}_{\mu_x}[J_x(\widehat{\pi}_{x,n})] = \widetilde{\tau}\,\mathbb{E}_{\mu_x}[\mathrm{KL}(\widehat{\pi}_{x,n}\|\pi_x^*)].$$

Together with (39), this proves the claimed $O(n^{-3/2})$ expansion. Multiplying by n gives the stated positive limit, since $d_x > 0$, and hence the tightness of the $1/n$ rate. □

The supporting lemma controls the cubic KL remainder through third absolute moments of the occupation errors.

C.3.1 CONTROLLING THE KL REMAINDER

Lemma C.4 (Expected remainder in the quadratic expansion of KL)**.** *For each fixed prompt x, there is a finite constant $C_{\mathrm{ent},x}$, depending only on its kernel P_x, such that, for every initial distribution μ_x and every integer $n \geq 1$,*

$$\mathbb{E}_{\mu_x}\left|\mathrm{KL}(\widehat{\pi}_{x,n}\|\pi^*_x) - \frac{1}{2}\chi^2(\widehat{\pi}_{x,n}\|\pi^*_x)\right| \leq \frac{C_{\mathrm{ent},x}}{n^{3/2}}. \tag{40}$$

Proof. Fix x and retain $P = P_x$, $\pi = \pi^*_x$, and $C_i = C^{(i)}_x$ from Appendix C.2. Write $\widehat{\pi}_n = \widehat{\pi}_{x,n}$, $\mu = \mu_x$, and $\mathcal{F}_i = \sigma(C_0, \ldots, C_i)$.

We first bound the KL remainder by third absolute moments of the occupation errors. We then obtain these moments through a Poisson decomposition and a bounded-increment martingale estimate.

Step 1: Reduce the KL remainder to third moments. Let $b := 1/\pi^*_{x,\min} - 1$ and define

$$R(z) := (1+z)\log(1+z) - z - \frac{z^2}{2}, \qquad -1 \leq z \leq b,$$

with $0 \log 0 = 0$. Taylor's formula at zero gives $R(z) = -z^3/6 + O(z^4)$, while $R(-1) = 1/2$. Consequently, $|R(z)|/|z|^3$ extends continuously at zero and is bounded on $[-1, b]$. Thus, for some finite c_π,

$$|R(z)| \leq c_\pi |z|^3, \qquad -1 \leq z \leq b. \tag{41}$$

This bound includes zero empirical coordinates.

Set $z_s := \widehat{\pi}_n(s)/\pi(s) - 1$. Then $z_s \in [-1, b]$ and $\sum_s \pi(s) z_s = 0$, so

$$\mathrm{KL}(\widehat{\pi}_n\|\pi) - \frac{1}{2}\chi^2(\widehat{\pi}_n\|\pi) = \sum_s \pi(s) R(z_s).$$

Taking absolute values and applying (41) gives

$$\mathbb{E}_\mu\left|\mathrm{KL}(\widehat{\pi}_n\|\pi) - \frac{1}{2}\chi^2(\widehat{\pi}_n\|\pi)\right| \leq c_\pi \sum_s \frac{\mathbb{E}_\mu|\widehat{\pi}_n(s) - \pi(s)|^3}{\pi(s)^2}. \tag{42}$$

It therefore remains to bound the third absolute moment of each occupation error.

Step 2: Decompose each occupation error. For each state s, define the centered indicator $f_s(u) := \mathbf{1}\{u = s\} - \pi(s)$ and the Poisson solution

$$h_s := \sum_{r=0}^{\infty} P^r f_s.$$

The mean-zero contraction (35) gives

$$\sum_{r=0}^{\infty} \|P^r f_s\|_\infty \leq \frac{\|f_s\|_{2,\pi}}{\sqrt{\pi^*_{x,\min}}} \sum_{r=0}^{\infty} (1-\gamma_x)^r = \frac{\|f_s\|_{2,\pi}}{\sqrt{\pi^*_{x,\min}}\gamma_x} < \infty.$$

Thus h_s is bounded and $(I - P)h_s = f_s$, by telescoping the series. Set $B_s := 2\|h_s\|_\infty$ and

$$D_i := h_s(C_i) - (Ph_s)(C_{i-1}), \qquad M_k := \sum_{i=1}^{k} D_i, \qquad M_0 := 0.$$

The Markov property gives $\mathbb{E}_\mu[D_i \mid \mathcal{F}_{i-1}] = 0$ and $|D_i| \leq B_s$. Hence $(M_k, \mathcal{F}_k)$ is a martingale with $\mathbb{E}_\mu M_k^2 \leq k B_s^2$. Summing the Poisson equation gives

$$n\big(\widehat{\pi}_n(s) - \pi(s)\big) = M_n + R_{s,n}, \qquad R_{s,n} := (Ph_s)(C_0) - (Ph_s)(C_n), \tag{43}$$

where $|R_{s,n}| \leq B_s$ on every sample path.

Step 3: Bound the third moments and conclude. Taylor's formula for $u \mapsto |u|^3$, whose second derivative is $6|u|$, gives, for all real a, b,

$$|a+b|^3 \le |a|^3 + 3a|a|b + 3|a|b^2 + |b|^3.$$

Apply this with $a = M_k$ and $b = D_{k+1}$. The term linear in D_{k+1} vanishes after conditioning on $\mathcal{F}_k$, so

$$\begin{aligned}\mathbb{E}_\mu|M_{k+1}|^3 - \mathbb{E}_\mu|M_k|^3 &\le 3B_s^2\mathbb{E}_\mu|M_k| + B_s^3 \\ &\le B_s^3(3\sqrt{k}+1).\end{aligned}$$

Summing over $0 \le k < n$ yields $\mathbb{E}_\mu|M_n|^3 = O(n^{3/2})$. Using $|a+b|^3 \le 4(|a|^3+|b|^3)$ in (43) and dividing by n^3, we obtain

$$\mathbb{E}_\mu|\widehat{\pi}_n(s) - \pi(s)|^3 \le C_s n^{-3/2}, \tag{44}$$

for a finite C_s independent of μ and n.

Substituting (44) into (42) yields

$$\mathbb{E}_\mu\left|\mathrm{KL}(\widehat{\pi}_n\|\pi) - \frac{1}{2}\chi^2(\widehat{\pi}_n\|\pi)\right| \le \left(c_\pi \sum_s \frac{C_s}{\pi(s)^2}\right) n^{-3/2}.$$

The finite state space and full support of π make the constant $C_{\mathrm{ent},x} := c_\pi \sum_s C_s/\pi(s)^2$ finite. It is independent of the initialization and horizon, proving the lemma. □

D Proofs for the Adaptive Multi-Prompt Allocation

This appendix establishes the guarantees for adaptive multi-prompt allocation. Appendix D.1 derives the oracle square-root allocation. Appendix D.2 proves almost-sure consistency of the local kernel and difficulty estimates. Appendix D.3 then shows that the adaptive allocation tracks the oracle proportions almost surely. Finally, Appendix D.4 combines uniform occupation control with lower bounds on local sample sizes to establish the expected and in-probability objective-gap guarantees.

Throughout, k denotes a local transition count, while t and T denote global rounds; the local count for prompt x after t rounds is $N_x(t)$. The scheduler uses only past observations to decide which prompt to update. Each selected prompt receives a fresh proposal from q_x and fresh comparisons under Section 2's conditional feedback law. Thus scheduling changes when a local chain advances, while its transition kernel in local time remains P_x as defined in (9).

D.1 Deriving the optimal square-root allocation

Proposition 4.1 (restated). *Among all interior allocation proportions $\omega \in \Delta(\mathcal{X})$, the leading coefficient $\sum_{x\in\mathcal{X}} f_x d_x/\omega_x$ is uniquely minimized by*

$$\omega_x^{\mathrm{or}} := \frac{\sqrt{f_x d_x}}{\sum_{y\in\mathcal{X}}\sqrt{f_y d_y}}, \qquad n_x^{\mathrm{or}} := T\omega_x^{\mathrm{or}}. \tag{45}$$

Here n_x^{or} denotes the continuous target count for prompt x.

Proof. For any interior $\omega \in \Delta(\mathcal{X})$, Cauchy–Schwarz and $\sum_x \omega_x = 1$ give

$$\sum_x \frac{f_x d_x}{\omega_x} = \left(\sum_x \frac{f_x d_x}{\omega_x}\right)\left(\sum_x \omega_x\right) \ge \left(\sum_x \sqrt{f_x d_x}\right)^2.$$

Since every $f_x d_x > 0$, equality holds exactly when $\sqrt{f_x d_x/\omega_x} = c\sqrt{\omega_x}$ for all x, with a common constant $c > 0$, or equivalently when $\omega_x \propto \sqrt{f_x d_x}$. Normalization gives ω^{or} in (45), which is interior and is the unique minimizer. The minimum value is $\left(\sum_x \sqrt{f_x d_x}\right)^2$, and multiplying the optimal proportions by T gives the continuous target counts n_x^{or}. □

D.2 Estimating the local transition kernel and difficulty

Proposition 4.2 (restated). *For every prompt x and every initial state,*

$$\widehat{P}_{x,k} \longrightarrow P_x, \qquad \widehat{d}_{x,k} \longrightarrow d_x \quad \textit{almost surely as } k \to \infty.$$

Proof of Proposition 4.2. Fix a prompt x. We first prove kernel consistency from the local-count limits in Lemma D.1, stated and proved below. We then obtain difficulty consistency by continuity of the kernel-to-difficulty map.

Kernel estimation. We prove

$$\widehat{P}_{x,k} \longrightarrow P_x \quad \text{almost surely.}$$

For an ordered state pair (s, s'), define the centered transition count

$$\mathcal{M}_k(s, s') := V_{x,k}(s, s') - P_x(s, s')V_{x,k}(s).$$

Subtracting $P_x(s, s')$ from the estimator in (16) gives the exact identity

$$\widehat{P}_{x,k}(s, s') - P_x(s, s') = \frac{\mathcal{M}_k(s, s') + \alpha\big(q_x(s') - P_x(s, s')\big)}{V_{x,k}(s) + \alpha}. \tag{46}$$

Lemma D.1 gives $\mathcal{M}_k(s, s')/k \to 0$ and $V_{x,k}(s)/k \to \pi^*_x(s) > 0$ almost surely. Dividing both the numerator and denominator in (46) by k therefore proves entrywise almost-sure convergence.

Since the state space is finite, entrywise convergence gives $\widehat{P}_{x,k} \to P_x$ almost surely in any fixed matrix norm.

Difficulty estimation. We prove

$$\widehat{d}_{x,k} \longrightarrow d_x \quad \text{almost surely.}$$

For a positive row-stochastic matrix R near P_x, let $\pi(R)$ be its stationary distribution and define

$$\mathcal{G}_x(R) := \operatorname{tr}\Big[\big(I - R + \mathbf{1}\pi(R)^\top\big)^{-1}\Big] - 1 - \frac{K_x - 1}{2}.$$

At the true kernel, the matrix $I - P_x + \mathbf{1}(\pi^*_x)^\top$ has eigenvalue one on constant vectors and eigenvalues $1 - \lambda_{x,j}$ on the remaining eigenspaces. Thus, by (11),

$$\mathcal{G}_x(P_x) = \sum_{j=2}^{K_x} \left(\frac{1}{1 - \lambda_{x,j}} - \frac{1}{2}\right) = \frac{1}{2}\sum_{j=2}^{K_x} \frac{1 + \lambda_{x,j}}{1 - \lambda_{x,j}} = d_x.$$

The estimator definition (17) likewise gives $\widehat{d}_{x,k} = \mathcal{G}_x(\widehat{P}_{x,k})$.

The eigenvalue one of P_x is simple, so its normalized stationary distribution depends continuously on the kernel near P_x. The matrix in the inverse is nonsingular at P_x and remains so in a neighborhood. Continuity of matrix inversion therefore makes $\mathcal{G}_x$ continuous at P_x. Together with kernel consistency, this gives

$$\widehat{d}_{x,k} = \mathcal{G}_x(\widehat{P}_{x,k}) \longrightarrow \mathcal{G}_x(P_x) = d_x \quad \text{almost surely,}$$

completing the proof. □

We now prove the local-count limits used in the kernel-estimation argument. The transition-count error is a martingale; the state-count error is a martingale plus a bounded endpoint term.

Lemma D.1 (Local count control)**.** *Fix a prompt x and any initial state. For every ordered pair $(s, s') \in \mathcal{S}_x^2$, define*

$$\mathcal{M}_k(s, s') := V_{x,k}(s, s') - P_x(s, s')V_{x,k}(s).$$

Then, almost surely,

$$\frac{\mathcal{M}_k(s, s')}{k} \longrightarrow 0, \qquad \frac{V_{x,k}(s)}{k} \longrightarrow \pi^*_x(s). \tag{47}$$

Proof. Suppress the fixed prompt subscript and write

$$C_i := C_x^{(i)}, \quad P := P_x, \quad \pi := \pi_x^*, \quad V_k(s) := V_{x,k}(s), \quad V_k(s,s') := V_{x,k}(s,s').$$

Let $\mathcal{F}_k := \sigma(C_0, \dots, C_k)$ be the history through the kth local transition. Both parts use the following martingale strong law. If $M_k = \sum_{i=1}^k D_i$ is a martingale with respect to $(\mathcal{F}_k)$ and $|D_i| \le L$ for a deterministic constant $L < \infty$, then

$$\sum_{i=1}^{\infty} \frac{\mathbb{E}[D_i^2 \mid \mathcal{F}_{i-1}]}{i^2} \le L^2 \sum_{i=1}^{\infty} \frac{1}{i^2} < \infty \quad \text{almost surely.}$$

This conditional second-moment bound verifies the hypothesis of the martingale strong law of Hall & Heyde (1980, Theorem 2.18), with normalizing sequence $U_k = k$. Consequently $M_k/k \to 0$ almost surely.

Transition counts.

Fix an ordered pair (s, s') and write

$$\begin{aligned} \mathcal{M}_k(s,s') &= \sum_{i=0}^{k-1} \zeta_{i+1}(s,s'), \\ \zeta_{i+1}(s,s') &:= \mathbf{1}\{C_i = s\}\big(\mathbf{1}\{C_{i+1} = s'\} - P(s,s')\big). \end{aligned} \tag{48}$$

The Markov property gives

$$\mathbb{E}[\zeta_{i+1}(s,s') \mid \mathcal{F}_i] = \mathbf{1}\{C_i = s\}\big(P(C_i,s') - P(s,s')\big) = 0.$$

Also $|\zeta_{i+1}(s,s')| \le 1$. Thus $(\mathcal{M}_k(s,s'), \mathcal{F}_k)_{k \ge 0}$ is a martingale. The strong law above gives $\mathcal{M}_k(s,s')/k \to 0$ almost surely.

State frequencies.

Fix a state s and define $g_s(u) := \mathbf{1}\{u = s\} - \pi(s)$. We construct a solution to the Poisson equation explicitly. Since the chain is finite and irreducible, one is a simple eigenvalue of P, and $I - P + \mathbf{1}\pi^\top$ is invertible. Define

$$h_s := (I - P + \mathbf{1}\pi^\top)^{-1} g_s.$$

Using $\pi^\top P = \pi^\top$ and $\pi^\top \mathbf{1} = 1$, multiply $(I - P + \mathbf{1}\pi^\top)h_s = g_s$ by $\pi^\top$ to obtain $\pi^\top h_s = \pi^\top g_s = 0$. Consequently

$$h_s - P h_s = g_s.$$

The function h_s is bounded because the state space is finite. Set

$$\mathcal{U}_k(s) := \sum_{i=0}^{k-1} \big(h_s(C_{i+1}) - (Ph_s)(C_i)\big).$$

Each increment has conditional mean zero given $\mathcal{F}_i$ and absolute value at most $2\|h_s\|_\infty$. Therefore $\mathcal{U}_k(s)$ is a martingale with

$$\mathbb{E}[\mathcal{U}_k(s)^2] \le 4\|h_s\|_\infty^2 k, \qquad \frac{\mathcal{U}_k(s)}{k} \to 0 \quad \text{almost surely.}$$

The state count uses the starting states of the first k transitions: $V_k(s) = \sum_{i=0}^{k-1} \mathbf{1}\{C_i = s\}$. Substituting $g_s = h_s - Ph_s$ and telescoping gives

$$\begin{aligned} V_k(s) - k\pi(s) &= \sum_{i=0}^{k-1} g_s(C_i) \\ &= \mathcal{U}_k(s) + h_s(C_0) - h_s(C_k). \end{aligned} \tag{49}$$

The endpoint term is bounded by $2\|h_s\|_\infty$. Dividing by k and using $\mathcal{U}_k(s)/k \to 0$ almost surely gives $V_k(s)/k \to \pi(s)$ almost surely, completing (47). Neither argument requires stationary initialization, so both conclusions hold for every initial state. □

D.3 Tracking the oracle allocation

Lemma 4.3 (restated). *Under Algorithm 2, $N_x(t) \to \infty$ for every prompt $x \in \mathcal{X}$ on every sample path. Moreover, almost surely and simultaneously for all $x \in \mathcal{X}$,*

$$\bar{\omega}_x(t) \longrightarrow \omega_x^{\mathrm{or}}, \qquad \frac{N_x(t)}{t} \longrightarrow \omega_x^{\mathrm{or}} \qquad (t \to \infty).$$

Proof of Lemma 4.3. We use the cumulative rounding bound in Lemma D.2, stated below. Write $M := |\mathcal{X}|$ and $Q_x(t) := \sum_{u=0}^{t-1} \bar{\omega}_x(u)$. The smoothing rule (20) gives $\bar{\omega}_x(u) \geq 1/(M\sqrt{u+1})$. Together with Lemma D.2, this yields, for $t \geq 1$,

$$N_x(t) \geq Q_x(t) - (M-1) \geq \frac{1}{M} \sum_{u=0}^{t-1} (u+1)^{-1/2} - (M-1) \geq \frac{\sqrt{t}}{M} - (M-1).$$

The last inequality uses $(u+1)^{-1/2} \geq t^{-1/2}$ for $u < t$. Thus every $N_x(t)$ diverges on every sample path.

We next work on the probability-one event on which the consistency conclusion of Proposition 4.2 holds for all prompts. Since $N_x(t) \to \infty$, evaluating the local difficulty estimates at these counts gives $\widehat{d}_{x,N_x(t)} \to d_x$. Therefore

$$\widehat{d}_x^{+}(t) = \max\{\widehat{d}_{x,N_x(t)}, (t+1)^{-1/2}\} \longrightarrow d_x > 0.$$

Continuity of the normalized square-root map in (19) gives $\widehat{\omega}_x(t) \to \omega_x^{\mathrm{or}}$. The smoothing weight tends to zero, so (20) also gives $\bar{\omega}_x(t) \to \omega_x^{\mathrm{or}}$.

Finally, the rounding discrepancy satisfies $|N_x(t) - Q_x(t)| \leq M - 1$. Cesaro's lemma consequently gives

$$\frac{N_x(t)}{t} = \frac{1}{t} \sum_{u=0}^{t-1} \bar{\omega}_x(u) + \frac{N_x(t) - Q_x(t)}{t} \longrightarrow \omega_x^{\mathrm{or}}.$$

All limits hold on the same probability-one event, proving the claimed simultaneous almost-sure convergence. □

The following rounding bound supplies the deterministic discrepancy estimate used in the proof above.

Lemma D.2 (Cumulative online rounding)**.** *For any sequence $\rho(0), \rho(1), \ldots$ in the $|\mathcal{X}|$-simplex, define $Q(t) = \sum_{u=0}^{t-1} \rho(u)$ and choose*

$$X_{t+1} \in \arg\max_{x \in \mathcal{X}} \{Q_x(t+1) - N_x(t)\}.$$

Then, for every t,

$$\max_x |N_x(t) - Q_x(t)| \leq |\mathcal{X}| - 1. \tag{50}$$

Thus the realized count stays within a constant of the cumulative fractional target. This conclusion follows directly from (Garivier & Kaufmann, 2016, Lemma 15).

D.4 Controlling the returned policy's objective gap

Theorem 4.4 (restated). *Under the standing assumptions of Section 4, run Algorithm 2 from arbitrary initial responses and let $\widehat{\pi}_T$ be its returned policy at deterministic horizon T. As $T \to \infty$,*

$$\mathbb{E}[J(\pi^*) - J(\widehat{\pi}_T)] = O(T^{-1/2}), \qquad J(\pi^*) - J(\widehat{\pi}_T) = O_p(T^{-1}).$$

For fixed m and comparison budget $B = mT$, the corresponding rates are $O(B^{-1/2})$ and $O_p(B^{-1})$.

Proof of Theorem 4.4. Both rates follow from the uniform occupation bound in Lemma D.3, stated and proved below. Forced exploration gives a deterministic lower bound of order $\sqrt{T}$ on each local

sample size, while allocation consistency gives a lower bound of order T with probability tending to one.

Write $\Delta_T := J(\pi^*) - J(\widehat{\pi}_T)$. The aggregate gap identity (6) and $\mathrm{KL} \le \chi^2$ give

$$0 \le \Delta_T = \widetilde{\tau} \sum_{x \in \mathcal{X}} f_x \, \mathrm{KL}(\widehat{\pi}_{x,N_x(T)} \| \pi^*_x) \le \widetilde{\tau} \sum_{x \in \mathcal{X}} f_x \chi^2(\widehat{\pi}_{x,N_x(T)} \| \pi^*_x).$$

The counts $N_x(T)$ depend on the observed trajectories, so a fixed-horizon expectation bound cannot be substituted at these indices. On $\{\min_x N_x(T) \ge k\}$, each occupation error is bounded pathwise by its supremum over all local horizons $\ell \ge k$. Lemma D.3 therefore yields, for every integer $k \ge 1$,

$$\mathbb{E}\Big[\mathbf{1}\{\min_x N_x(T) \ge k\} \Delta_T\Big] \le \frac{C}{k}, \qquad C := \widetilde{\tau} \sum_{x \in \mathcal{X}} f_x C_x < \infty. \tag{51}$$

Expected rate.

The proof of Lemma 4.3 (Appendix D.3) gives

$$N_x(T) \ge \frac{\sqrt{T}}{|\mathcal{X}|} - (|\mathcal{X}| - 1).$$

Thus, with $c := 1/(2|\mathcal{X}|)$, every count is at least $k_T := \lceil c\sqrt{T} \rceil$ on every sample path for all sufficiently large T. Applying (51) with $k = k_T$ gives

$$\mathbb{E}[\Delta_T] \le \frac{C}{k_T} = O(T^{-1/2}),$$

which proves (22).

In-probability rate.

Fix $0 < \nu < \min_x \omega^{\mathrm{or}}_x$ and define $\mathcal{E}_T := \{\min_x N_x(T) \ge \nu T\}$. The allocation consistency in Lemma 4.3 gives $\mathbb{P}(\mathcal{E}_T^c) \to 0$. Since the counts are integers, (51) with $k = \lceil \nu T \rceil$ gives $\mathbb{E}[\mathbf{1}\{\mathcal{E}_T\}\Delta_T] \le C/(\nu T)$. For every $R > 0$, Markov's inequality consequently yields

$$\mathbb{P}(T\Delta_T > R) \le \mathbb{P}(\mathcal{E}_T^c) + \frac{T}{R}\mathbb{E}[\mathbf{1}\{\mathcal{E}_T\}\Delta_T] \le \mathbb{P}(\mathcal{E}_T^c) + \frac{C}{\nu R}.$$

Taking the limit superior as $T \to \infty$ and then letting $R \to \infty$ proves $T\Delta_T = O_p(1)$, or equivalently $\Delta_T = O_p(T^{-1})$.

Finally, substituting $T = B/m$ with fixed m gives the corresponding rates $O(B^{-1/2})$ in expectation and $O_p(B^{-1})$ in probability. □

We now prove the uniform occupation bound used above, using the Poisson decomposition and martingale second-moment estimate from Lemma D.1.

Lemma D.3 (Uniform occupation control)**.** *For every prompt x, there is a finite constant C_x, depending only on P_x, such that, from any initial state and for every integer $k \ge 1$,*

$$\mathbb{E}\Big[\sup_{\ell \ge k} \chi^2(\widehat{\pi}_{x,\ell} \| \pi^*_x)\Big] \le \frac{C_x}{k}. \tag{52}$$

Proof. Fix a prompt x and write $P := P_x$, $\pi := \pi^*_x$, and $C_i := C_x^{(i)}$. We apply the indicator-based Poisson decomposition from Lemma D.1, then sum over dyadic blocks.

For each state $s \in \mathcal{S}_x$, let h_s be the bounded solution of $h_s - Ph_s = \mathbf{1}\{\cdot = s\} - \pi(s)$ constructed there, and let $\mathcal{U}_n(s)$ be its martingale. The centered occupation sum satisfies

$$S_n(s) := \sum_{i=1}^{n} \big(\mathbf{1}\{C_i = s\} - \pi(s)\big) = \mathcal{U}_n(s) + (Ph_s)(C_0) - (Ph_s)(C_n).$$

This is (49) with the counted states shifted from $C_0, \dots, C_{n-1}$ to $C_1, \dots, C_n$. The endpoint term is bounded by $2\|h_s\|_\infty$, and $\mathbb{E}[\mathcal{U}_n(s)^2] \le 4\|h_s\|_\infty^2 n$. Thus Doob's L^2 inequality gives, for every integer $q \ge 1$,

$$\mathbb{E}\left[\max_{1\le n\le q} |S_n(s)|^2\right] \le 8\mathbb{E}[\mathcal{U}_q(s)^2] + 8\|h_s\|_\infty^2 \le A_s q,$$

where $A_s < \infty$ depends only on P, uniformly over the initial state.

Partitioning the integers $n \ge k$ into the blocks $[2^r k, 2^{r+1}k)$ yields

$$\begin{aligned}\mathbb{E}\left[\sup_{n\ge k} \frac{|S_n(s)|^2}{n^2}\right] &\le \sum_{r=0}^{\infty} \frac{\mathbb{E}[\max_{1\le n\le 2^{r+1}k} |S_n(s)|^2]}{(2^r k)^2} \\ &\le \frac{2A_s}{k}\sum_{r=0}^{\infty} 2^{-r} = \frac{4A_s}{k}.\end{aligned}$$

Finally,

$$\chi^2(\widehat{\pi}_{x,n}\|\pi) = \sum_{s\in\mathcal{S}_x} \frac{|S_n(s)|^2}{n^2\pi(s)}.$$

Taking the supremum and expectation, then summing the statewise bounds, proves (52) with $C_x := 4\sum_{s\in\mathcal{S}_x} A_s/\pi(s) < \infty$. □

E Experimental Details

This appendix documents the two experiments in Section 5. We first describe the common methods, budget convention, and uncertainty summaries, then give the synthetic instance and the SQLite experiment. The anonymous supplement will include experiment code, frozen input pools, configurations, environment metadata, checkpoints, raw call logs, execution and evaluation records, and plotting scripts.

E.1 Common protocol

Both experiments use three replications and requested temperature $\tau = 0.5$, so $m = 2$ and $\widetilde{\tau} = 1/m = 0.5$.

Methods. RWPS+CTM uses the cumulative-tracking rule in Algorithm 2, transition pseudocount $\alpha = 0.5$, and $\varepsilon_t = (t+1)^{-1/2}$. Its initial difficulty estimate is $(K_x - 1)/2$ and subsequent estimates are computed from empirical transition counts. RWPS+ROUND-ROBIN uses the same within-prompt RWPS update but visits prompts cyclically. UNIFORM+ROUND-ROBIN also visits prompts cyclically; within a prompt it traverses a seeded random permutation of all unordered candidate pairs and reshuffles after a complete cycle. Let $W_x(s)$ denote the accumulated win credit of response s, initialized to zero. For a compared pair (s, u), agreement of both counterbalanced judgments in favor of s adds one to $W_x(s)$ and zero to $W_x(u)$, and conversely when both favor u. If the judgments disagree, each response receives half a win. After at least one pair has been compared for prompt x, the returned policy is

$$\widehat{\pi}_x(s) = \frac{W_x(s)}{\sum_{u\in\mathcal{S}_x} W_x(u)}.$$

The synthetic experiment additionally includes RWPS+ORACLE-CTM. It uses the same cumulative-tracking and forced-exploration mechanism, but replaces the estimated difficulty by d_x computed from the known synthetic kernel in (9), using (11). It forms the allocation directly from these true difficulties, without applying the time-dependent difficulty floor. It is a known-difficulty allocation benchmark, not a finite-budget lower bound on the gaps of other methods.

Sampling and returned policies. For each replication, the initial response for every prompt is drawn from q_x. Proposal streams are drawn from q_x in advance and shared across RWPS+CTM and RWPS+ROUND-ROBIN, as well as RWPS+ORACLE-CTM in the synthetic experiment. Each non-self RWPS update uses two judgments, one in each A/B order, and accepts the proposal only if

both prefer it. The returned RWPS policy is the empirical occupation distribution. Write $(\widehat{\pi}_T)_x$ for the prompt-x component of a method's policy at global checkpoint T. For RWPS, this is $\widehat{\pi}_{x,N_x(T)}$, with the zero-update convention from Section 4.

Within each experiment, all methods receive the same number T of prompt updates. Thus $B = mT = 2T$ is a nominal sampling-judgment allowance, not the actual call count: RWPS self-proposals use no calls. Checkpoint evaluation calls are additional and are not included in B. The comparisons therefore use equal update counts, not matched realized API expenditures.

Uncertainty summaries. Each curve has three replication-level values at each checkpoint, generated with distinct replication seeds. We report their arithmetic mean. For each method or matchup and checkpoint, we draw 10,000 bootstrap samples of size three with replacement from these three values and plot the 2.5th and 97.5th percentiles of the bootstrap means. These descriptive intervals summarize between-replication variation, rather than providing precise population confidence intervals or formal tests.

E.2 Synthetic weighted hard experiment

Instance. The frozen pool contains 30 prompts and 187 candidates. Each K_x is sampled uniformly from $\{2, \dots, 10\}$. The first three prompts each receive deployment weight 0.2, and the other 27 each receive $0.4/27$. For each of the first three prompts, $q_{x,i} \propto \mathrm{geomspace}(1, 0.025)_i$ and the candidate utilities are equally spaced from -6 to 6. Thus high-utility, high-index answers have low proposal probability. For each remaining prompt, we independently draw unnormalized candidate weights $w_{x,i}$ such that $\log w_{x,i} \sim \mathcal{N}(0, 0.9^2)$. We lower-bound every weight by 0.02 times the within-prompt mean and normalize the resulting weights to sum to one, yielding a non-uniform reference distribution q_x with full support. Candidate utilities are equally spaced from -3 to 3. Pairwise outcomes follow the Bradley–Terry probability $\Pr(i \succ j) = \{1 + \exp(r_{x,j} - r_{x,i})\}^{-1}$, with independent draws for sampling judgments. Each method receives $T = 6000$ updates.

Exact objective-gap evaluation. Because the synthetic utilities are known, no additional pairwise judge is needed for the plotted metric. For every saved checkpoint policy $\widehat{\pi}_T$, we calculate

$$J(\widehat{\pi}_T) = \sum_x f_x \left[\langle (\widehat{\pi}_T)_x, r_x \rangle - \widetilde{\tau}\ \mathrm{KL}((\widehat{\pi}_T)_x \| q_x) \right], \qquad \widetilde{\tau} = 0.5.$$

The exact optimizer is $\pi^*_{x,i} \propto q_{x,i} \exp(r_{x,i}/\widetilde{\tau})$. We report $J(\pi^*) - J(\widehat{\pi}_T)$, with final-checkpoint results in Table 2.

Table 2: Synthetic objective gaps at $T = 6000$ ($B = 12000$). Means and descriptive 95% bootstrap intervals use three replications; smaller gaps are better.

Method	Mean gap	95% interval
RWPS+CTM	0.138	$[0.101, 0.172]$
RWPS+ORACLE-CTM	0.154	$[0.105, 0.189]$
RWPS+ROUND-ROBIN	0.763	$[0.587, 0.865]$
UNIFORM+ROUND-ROBIN	1.566	$[1.563, 1.571]$

E.3 SQLite experiment

Questions and candidates. We constructed 30 questions from official SQLite documentation: six each on transactions, foreign keys, backups, write-ahead logging, and in-memory databases. Questions and candidate answers were generated with `gpt-4.1-mini-2025-04-14`. Candidate counts were drawn independently and uniformly from $\{2, \dots, 10\}$, yielding 179 candidate answers. `gpt-5.6-luna` reviewed question clarity, answer distinctness, factual support, difficulty, and answer-quality labels (strong, acceptable, weak, or incorrect), with rejected items repaired and reviewed again until approved.

Deployment and proposal distributions. The configuration was fixed before the formal run, without using experimental outcomes. Questions are ranked first by the initial CTM proxy $(K_x - 1)/2$, then by reviewed difficulty, with the fixed question identifier breaking remaining ties. The highest-ranked six questions each receive deployment weight 0.1; the other 24 each receive $0.4/24$. Let $q_x^{(0)}$ denote the original positive proposal distribution generated from seeded logistic-normal weights, and let $\mathcal{S}_x^{\text{strong}}$ be the set of answers labeled strong. For each question, strong answers collectively receive proposal probability 0.1, divided in proportion to their original probabilities. The remaining 0.9 is allocated among non-strong answers after multiplying their original probabilities by $a(s)$, where $a(s) = 0.5$, 1, or 2 for acceptable, weak, or incorrect answers, respectively. Thus

$$q_x(s) = \begin{cases} 0.1 \dfrac{q_x^{(0)}(s)}{\sum_{u \in \mathcal{S}_x^{\text{strong}}} q_x^{(0)}(u)}, & s \in \mathcal{S}_x^{\text{strong}}, \\ 0.9 \dfrac{q_x^{(0)}(s)a(s)}{\sum_{u \in \mathcal{S}_x \setminus \mathcal{S}_x^{\text{strong}}} q_x^{(0)}(u)a(u)}, & s \notin \mathcal{S}_x^{\text{strong}}. \end{cases}$$

All candidates retain positive proposal probability, and the distributions are held fixed across methods.

Pairwise judge. `gpt-5.6-luna`, with reasoning effort set to `none`, serves as both the sampling judge and final evaluator. Given the supplied SQLite reference, it selects the more factually correct, complete, and relevant answer. Both sampling and evaluation use the two A/B display orders.

Checkpoint pairwise evaluation. Each method receives $T = 5000$ updates. At each checkpoint and for each directed matchup, each replication performs 150 fresh evaluation trials. In a trial, a prompt X is sampled from the fixed deployment distribution f, and answers $A \sim (\widehat{\pi}_T^L)_X$ and $B \sim (\widehat{\pi}_T^R)_X$ are sampled independently from the two saved checkpoint policies. If $A = B$, the trial score is $Z = 1/2$ without an API call. Otherwise, the evaluator judges the pair twice, once in each A/B display order. We set $Z = 1$ only if both judgments choose the left method's answer, $Z = 0$ only if both choose the right method's answer, and $Z = 1/2$ if the two orders disagree. The replication win ratio is $150^{-1} \sum_{i=1}^{150} Z_i$; the plotted point averages the three replication ratios. Hence 0.5 denotes parity and a value above 0.5 favors the first-named method. Table 3 summarizes the three displayed matchups at the final checkpoint.

Table 3: SQLite experiment at $T = 5000$ ($B = 10000$). Means and descriptive 95% bootstrap intervals use three replications. Ratios above 0.5 favor the first-named method.

Matchup	Mean ratio	95% interval
RWPS+CTM vs. RWPS+ROUND-ROBIN	0.553	$[0.533, 0.583]$
RWPS+CTM vs. UNIFORM+ROUND-ROBIN	0.710	$[0.673, 0.770]$
RWPS+ROUND-ROBIN vs. UNIFORM+ROUND-ROBIN	0.728	$[0.680, 0.780]$

Scope. The experiment evaluates policies over fixed, AI-generated and AI-reviewed candidate sets, without independent domain-expert validation. The same model supplies sampling judgments and final evaluation, so counterbalancing addresses answer-order effects but does not provide independent validation or verify the theoretical preference-model assumptions. The results concern pairwise preference under the specified deployment and proposal distributions, rather than the KL-regularized objective gap.