SAMPLE EFFICIENT BEST-LLM IDENTIFICATION VIA ACTIVE PAIRWISE PREFERENCE ELICITATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Practitioners deploying Large Language Models (LLMs) routinely face the following problem: among the many candidate LLMs available, which one performs best on my task? Public leaderboards rank models on generic benchmarks that may not reflect a particular use case, while constructing a fully labeled task-specific evaluation set can be prohibitively expensive. This raises an active evaluation question: which model-pair-prompt comparisons should be annotated so that the best model can be identified with the fewest labels? In this paper, we study best-LLM identification from a fixed unlabeled pool of prompts. At each round, the learner selects a $(\text{model}_i, \text{model}_j, \text{prompt}_n)$ triple, elicits a pairwise preference from an annotator between the two candidate responses, and uses this feedback to decide which comparison to query next. We introduce TOPS (Top-focused Online Preference Sampling), an active acquisition strategy that maintains a Bradley-Terry posterior over model abilities and samples comparisons using acquisition scores that combine response dissimilarity with estimated reductions in uncertainty about the margins between the leader and its challengers. Across 17 settings, TOPS achieves the best average rank for cumulative regret, final regret, and the rank of the selected model. Using 500 queries, TOPS is able to identify the single best model from 20 candidates in 62% of runs.

1 INTRODUCTION

Large Language Models (LLMs) are actively being deployed for many real-world applications, including customer support (Brynjolfsson et al., 2025), clinical decision support (Korom et al., 2025; Shah et al., 2026), and tutoring (Wang et al., 2024a). Practitioners deploying LLMs for a specific application rarely train their own model. Instead, they face a model selection problem: *Given a handful of candidate LLMs from open and proprietary providers, which one performs the best for my particular use case?*

One possible solution is to deploy the top-performing model from generic benchmarks like HELM (Liang et al., 2022) or Chatbot Arena (Chiang et al., 2024). However, prior work has documented validity and external-generalization concerns with benchmarks (Salaudeen et al., 2025), suggesting that leaderboard performance may not reliably transfer to particular deployment settings (Alaa et al., 2025; Hardy et al., 2025; Eriksson et al., 2025; Zhang et al., 2026). To overcome these issues, practitioners can construct a task-specific evaluation set using prompts drawn from the intended deployment distribution. However, obtaining high-quality labels for such prompts can require substantial time and annotation cost.

In this work, we study the problem of best-LLM identification when the practitioner has access to an *unlabeled* pool of prompts drawn from their use case. We identify the best-performing LLM by eliciting pairwise preferences from an annotator, who compares two candidate model outputs for the same prompt and selects their preferred response. The use of pairwise preferences over pointwise scoring of each output in isolation is motivated by a long line of work showing that humans (and models) make relative judgments far more reliably than absolute ones (Stewart et al., 2005; Mantiuk et al., 2012; Liusie et al., 2024). For example, on datasets with ground truths, Remeli & Hardt (2026) find that pairwise rankings correlate with accuracy rankings and can outperform direct grading when the judge is weak.

Throughout this paper, we define "best" as relative to the specified annotator on the target prompt pool, and this need not coincide with an external ground-truth notion of correctness. Thus, our goal is

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.

to identify the model most preferred by the chosen evaluator for the intended use case, rather than an objectively best model in an absolute sense.

Given K candidate models and N prompts, a naive evaluation strategy would collect annotations for the full $K \times K \times N$ preference tensor. This is often infeasible: preference labels can be expensive, especially when they require domain expertise, and many comparisons are unnecessary for the deployment decision. If the goal is only to identify the single best model, then comparisons between clearly suboptimal models, or comparisons on prompts that do not distinguish the top candidates, provide little value. This motivates an *active* evaluation problem: how should we adaptively choose which model-pair-prompt triples to annotate so that we identify the best model using as few preference queries as possible?

In this work, we formalize the problem of active pairwise preference elicitation for best-LLM identification, where the goal is to identify the best-performing model given a fixed budget of T annotation queries. We propose **T**op-focused **O**nline **P**reference **S**ampling (TOPS), an active Bradley-Terry (Bradley & Terry, 1952) based algorithm that maintains uncertainty over model strengths. TOPS tracks an active set of plausible challengers to the current leader and samples model-pair-prompt triples using acquisition scores that estimate reductions in uncertainty about the leader-challenger margins. TOPS uses a pooled Bradley-Terry posterior to decide which model pairs are decision-relevant, and uses response dissimilarity as a lightweight heuristic for choosing prompts that are likely to distinguish those models. TOPS is motivated by prior work in dueling bandits, which provides a natural framework for learning from pairwise comparisons (Yue et al., 2012; Lekang & Lamperski, 2019). Crucially, compared to recent works in sample-efficient LLM evaluation (Chiang et al., 2024; Zhou et al., 2025), we do not aim to recover an accurate ranking over all models; our goal is only to identify the single best model. We evaluate TOPS on a variety of datasets with both realistic and deliberately pathological LLM judges as annotators. Across the 17 main evaluation settings, TOPS achieves the best average rank for cumulative regret, final regret, and the rank of the selected model. At $T = 500$, it identifies the winner in 62% of runs, compared with at most 46% for any baseline.

Our key contributions are as follows:

- We formalize best-LLM identification as an active pairwise preference elicitation problem over a fixed, unlabeled pool of prompts. Unlike standard benchmark evaluation, our goal is not to estimate a full ranking or a model's average performance on a generic test set, but to identify the best model for a particular use case while minimizing the number of preference annotations required.
- We propose TOPS, an acquisition rule that jointly selects two models and a prompt for best-model identification. TOPS samples triples using acquisition scores that combine response dissimilarity with estimated reductions in uncertainty about the margins between the leader and its challengers. We demonstrate that TOPS achieves the best average rank for cumulative regret, final regret, and the rank of the selected model across 17 evaluation settings.
- We provide theoretical justification for the design of TOPS by deriving regret bounds and finite-time identification guarantees under bounded prompt heterogeneity.

2 Related Work

Active Best-Arm Identification from Pairwise Comparisons Our work is related to best-arm identification from pairwise feedback, most notably the dueling-bandit literature. Classical dueling-bandit algorithms learn from comparisons between pairs of arms, using notions such as Condorcet, Copeland, or Borda winners, and select future comparisons to identify or compete with the best arm (Yue et al., 2012; Yue & Joachims, 2011; Zoghi et al., 2015; Dudík et al., 2015; Wu & Liu, 2016). These methods provide a natural foundation for our setting because feedback is pairwise and the goal is to identify a strong candidate under limited comparisons.

However, our setting differs in that we do not observe context-free comparisons between arms: each query is a triple consisting of two LLMs and one prompt from a fixed unlabeled pool. Thus, the learner must decide not only which two models to compare, but also on which prompt the comparison should be elicited. Regardless, TOPS is motivated by a series of algorithms from the dueling bandit literature which maintains a current best arm, and eliminates arms with low confidence of beating the leader (Yue et al., 2012; Yue & Joachims, 2011).

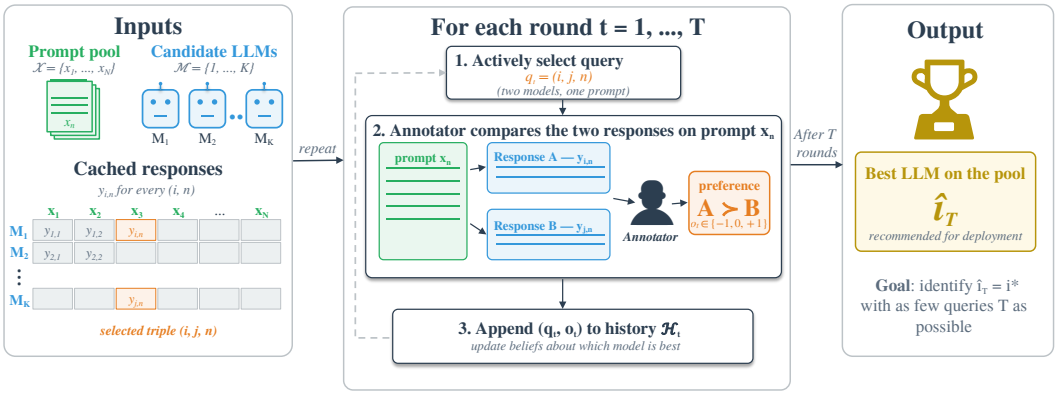


Figure 1: **Active LLM selection from a fixed prompt pool.** Given cached responses from K models on N prompts, the learner selects two models and one prompt, obtains an annotator's preference, and updates its estimates. After T queries, it recommends one model. The objective is to identify the pool-level Bradley-Terry winner using as few preference queries as possible.

Active Ranking A related line of work studies active ranking and best-item identification under Bradley-Terry preference models (Jamieson & Nowak, 2011; Falahatgar et al., 2017; Heckel et al., 2019; Li et al., 2026). Most relevant to this work is Bergström et al. (2024), which proposes an acquisition under a contextual logistic preference model by greedily acquiring comparisons which minimize expected uncertainty. However, in comparison with this line of work, our objective is top-focused. Rather than estimating a full preference matrix or recovering an accurate ranking over all models, we aim to identify only the best-performing model on the pool. This distinction is important because comparisons among clearly suboptimal models may be useful for full ranking, but provide little information about which model should be deployed.

Active Preference Learning for RLHF A parallel literature uses pairwise preferences to train reward models for reinforcement learning from human feedback (RLHF). In this setting, annotators compare candidate outputs, and the resulting preferences are used to fit a reward model that guides policy optimization (Christiano et al., 2017; Ouyang et al., 2022; Bai et al., 2022). Recent work has studied active query selection for preference learning, using uncertainty sampling, Bayesian information gain, or optimal-design criteria to decide which comparisons should be labeled (Houlsby et al., 2011; Mehta et al., 2023; Das et al., 2025; Muldrew et al., 2024).

Although these methods also use pairwise preference feedback, they typically optimize for a different downstream objective. Active RLHF uses preferences as training signal for a reward model and policy, so query selection often aims to cover the preference space broadly enough for the learned reward to generalize. In contrast, we use preferences for evaluation and selection: the candidate LLMs are fixed, and the goal is to identify which model is best on a given prompt pool.

Active Evaluation and Selection of LLMs The LLM evaluation literature has increasingly studied how to reduce the cost of evaluating many models on many prompts. Arena-style systems such as Chatbot Arena use pairwise preferences to estimate model rankings, and recent work has proposed more efficient pair-selection strategies for estimating leaderboard scores under Bradley-Terry-style models (Chiang et al., 2024; Liu et al.). However, these approaches optimize over model pairs only, leaving no mechanism to route the labeling budget toward prompts that better separate the top candidates. A complementary line of work focuses on the prompt axis, adaptively or statically selecting informative examples to reduce benchmark evaluation cost or to approximate full-benchmark rankings from a small prompt subset (Zhou et al., 2025; Shi et al., 2024; Polo et al., 2024; Vivek et al., 2024; Truong et al., 2025; Tolochinsky et al., 2026; Cordero Encinar et al., 2026; Fisch et al., 2026). These methods assume numeric per-prompt scores and an absolute-scoring objective rather than pairwise preferences. Most of them are passive (the subset is chosen once and reused across models).

3 PROBLEM SETUP

Setup Let $\mathcal{M} = \{1, \ldots, K\}$ denote the set of candidate LLMs and let $\mathcal{X} = \{x_1, \ldots, x_N\}$ denote a fixed unlabeled prompt pool of size N. In deployment settings, this pool may be modest in size and handcrafted for the target use case by domain experts, but exhaustively annotating all $O(K^2N)$ model-pair-prompt comparisons can still be prohibitively expensive. For every pair $(i, n) \in \mathcal{M} \times \mathcal{X}$, we have a model response $y_{i,n}$. We assume that candidate responses are already available or inexpensive to generate relative to expert annotation, and we will loosen this assumption empirically later.

Pairwise Preference Oracle For each prompt n we posit a latent per-prompt score vector $s_{\cdot,n} \in \mathbb{R}^K$ such that a user's preference between responses $y_{i,n}$ and $y_{j,n}$ follows a Bradley-Terry (BT) law:

$$\Pr(\, i \succ j \mid n \,) \;=\; \sigma\big(s_{i,n} - s_{j,n}\big), \qquad \sigma(z) = \frac{1}{1 + e^{-z}}.$$

We decompose the per-prompt score into a pool-level component and a zero-mean per-prompt residual,

$$s_{i,n} \;=\; w_i + \varepsilon_{i,n}, \qquad \mathbb{E}_n\big[\varepsilon_{i,n}\big] = 0 \text{ for every } i,$$

where $w \in \mathbb{R}^K$ is the pool-level BT parameter and ε captures prompt-specific heterogeneity. The prompt-agnostic special case $\varepsilon \equiv 0$ reduces to the classical context-free BT model; the regime of practical interest is $\mathrm{Var}_n(\varepsilon_{\cdot,n}) > 0$, under which different prompts rank the models differently.

Let $J \in \{-1, 0, +1\}^{K \times K \times N}$ denote the complete oracle tensor. Under the binary preference model, we draw one independent judgment for each unordered pair $i < j$ and prompt n. We set $J_{i,j,n} = +1$ if $i \succ j$ and $J_{i,j,n} = -1$ otherwise, and define $J_{j,i,n} = -J_{i,j,n}$ and $J_{i,i,n} = 0$. In experiments, annotators may also return a tie, represented by $J_{i,j,n} = 0$. Observing the full J is infeasible by assumption.

We define the pool-level best model as the argmax of the marginal BT maximum-likelihood estimator fit to the full J, when a finite maximizer exists:

$$w^* \;\in\; \arg \max_{u \in \mathbb{R}^K :\, \sum_i u_i = 0} \sum_{i<j} \sum_{n=1}^{N} \log \sigma\big(J_{i,j,n}\,(u_i - u_j)\big), \qquad i^\star \;=\; \arg\max_{i \in \mathcal{M}} w_i^*.$$

We write $r_i := w^*_{i^\star} - w^*_i \geq 0$ for the gap of model i to the best.

Sequential Protocol At each round $t = 1, \ldots, T$, the learner observes the history $\mathcal{H}_{t-1} = \{(q_\tau, o_\tau)\}_{\tau < t}$ and selects a query triple

$$q_t \;=\; (i_t, j_t, n_t) \;\in\; \mathcal{M} \times \mathcal{M} \times \mathcal{X}, \qquad i_t \neq j_t.$$

The oracle returns a preference $o_t \in \{-1, 0, +1\}$ with the same encoding as J. Under the binary preference model, each query yields a fresh outcome $o_t \in \{-1, +1\}$ with

$$\Pr(o_t = +1 \mid \mathcal{H}_{t-1}, q_t) \;=\; \sigma\big(s_{i_t,n_t} - s_{j_t,n_t}\big).$$

After T rounds the learner commits to a recommended model $\hat{\imath}_T \in \mathcal{M}$ based on the full history $\mathcal{H}_T$. We would like the learner to output $\hat{\imath}_T = i^\star$, or, at a minimum, $\hat{\imath}_T$ such that $r_{\hat{\imath}_T}$ is small. See Figure 1 for a visual overview.

4 METHODS

We now describe TOPS, our active acquisition strategy for the best-LLM identification problem described in Section 3. TOPS makes use of the following informal intuitions about this problem:

1. If two models i and j have identical or very similar outputs for a particular prompt n (i.e. $d(y_{i,n}, y_{j,n}) < \epsilon$ for some distance metric d), then querying the triple (i, j, n) is unlikely to provide useful information about their relative quality.
2. If the posterior indicates with high confidence that a model j is worse than the current leader, then any comparisons involving j are unlikely to affect the identity of the best model.

3. What ultimately determines the answer to the best-LLM identification problem is the sign of the decisive margins $w_{\hat{\imath}_{t-1}} - w_k$ for each remaining candidate k. So queries should not be scored by how much they reduce uncertainty in individual model strengths w_i, nor uniformly across every pairwise margin, but specifically by how much they shrink the posterior variance of these margins.

Dissimilarity Tensor Because the response cache $\{y_{i,n}\}$ is available offline, it can be summarized at zero annotation cost into a dissimilarity tensor

$$D \in \mathbb{R}_+^{K\times K\times N}, \qquad D_{ijn} = d\big(g(y_{i,n}),\, g(y_{j,n})\big), \tag{1}$$

where $g(\cdot)$ is a fixed sentence embedding and d a distance (we use the cosine distance). By construction, $D_{ijn} \to 0$ when $y_{i,n} \equiv y_{j,n}$, in which case the oracle's preference is uninformative.

Bradley-Terry Posterior In order to estimate confidence about differences in model strengths, we need a distribution over the BT parameters $w \in \mathbb{R}^K$. The BT model is invariant to a constant shift in w, so we fix $w_K = 0$ and place a zero-mean isotropic Gaussian prior $\mathcal{N}(0, I_{K-1})$ on the remaining coordinates. All estimates below are computed in these $K-1$ anchored coordinates and then centered so that $\sum_i w_i = 0$ (Appendix B). Let $\phi_{ij} := e_i - e_j$ denote the difference of standard basis vectors, with the abuse of notation that $\phi_\tau := \phi_{i_\tau, j_\tau}$. After $t-1$ queries with history $\mathcal{H}_{t-1} = \{(q_\tau, o_\tau)\}_{\tau<t}$, the BT estimate is then

$$\hat{w}_{t-1} \;=\; \arg\max_{w\in\mathbb{R}^K} \sum_{\tau<t} \log\sigma\big(o_\tau\, \phi_\tau^\top w\big) \;-\; \frac{1}{2}\,\|w\|_2^2.$$

To estimate the covariance of w, we use the Laplace approximation (Daxberger et al., 2021), which says that $w \mid \mathcal{H}_{t-1} \approx \mathcal{N}(\hat{w}_{t-1}, \Sigma_t)$, with

$$\Sigma_t^{-1} \;=\; \sum_{\tau<t} \alpha_{\tau,t}\, \phi_\tau \phi_\tau^\top + I, \qquad \alpha_{\tau,t} = \sigma(\phi_\tau^\top \hat{w}_{t-1})\,(1 - \sigma(\phi_\tau^\top \hat{w}_{t-1})). \tag{2}$$

As affine functions of a Gaussian are also Gaussian, we have that

$$\begin{aligned}(w_i - w_j) \mid \mathcal{H}_{t-1} &\approx \mathcal{N}\big(\hat{w}_{t-1,i} - \hat{w}_{t-1,j},\; \sigma_{ij,t}^2\big)\\ \sigma_{ij,t}^2 &= \Sigma_{t,ii} + \Sigma_{t,jj} - 2\,\Sigma_{t,ij} = \phi_{ij}^\top \Sigma_t \phi_{ij}\end{aligned} \tag{3}$$

In other words, every pairwise score gap has a Gaussian posterior whose mean and variance can be computed in closed form from $\hat{w}_{t-1}$ and Σ_t. See Appendix B for the full derivation of the above expressions.

Decisive Margin and Active Challenger Set Let $\hat{\imath}_{t-1} = \arg\max_i \hat{w}_{t-1,i}$ be the empirical leader. Best-LLM identification cares only about whether the *decisive margin*

$$m_{j,t} \;:=\; w_{\hat{\imath}_{t-1}} - w_j, \qquad j \neq \hat{\imath}_{t-1},$$

is positive. For each challenger, we assess whether this margin is positive using a one-sided posterior credible bound. We exploit this by maintaining an *active challenger set* $\mathcal{C}_t$ of models the posterior has not yet excluded, using the posterior derived in Eq. (3),

$$\mathcal{C}_t \;=\; \Big\{\, j \neq \hat{\imath}_{t-1} \;:\; \big(\hat{w}_{t-1,\hat{\imath}_{t-1}} - \hat{w}_{t-1,j}\big) \;-\; z_{\mathrm{cr}}\, \sigma_{\hat{\imath}_{t-1}j,\,t} \;\le\; 0 \,\Big\},$$

where z_{cr} is the standard normal quantile with $\Pr(Z > z_{\mathrm{cr}}) = \delta_{\mathrm{cr}}$ (we use $\delta_{\mathrm{cr}} = 0.05$). Here δ_{cr} is the posterior tail probability, and $1 - \delta_{\mathrm{cr}}$ is the one-sided credible level under the Gaussian approximation. $\mathcal{C}_t$ retains models whose lower $(1-\delta_{\mathrm{cr}})$-credible bound on the leader-vs-j margin is non-positive. The candidate pair set restricts attention to comparisons that can move the decisive margin:

$$\mathcal{P}_t \;=\; \big\{\, (\hat{\imath}_{t-1}, j) : j \in \mathcal{C}_t \,\big\} \;\cup\; \big\{\, (j,k) : j,k \in \mathcal{C}_t,\; j<k \,\big\}.$$

Variance Reduction Acquisition We now use the posterior derived in Eq. (2) to score candidate triples by how much they would reduce uncertainty about the decisive margins $w_{\hat{i}_{t-1}} - w_k$, $k \in \mathcal{C}_t$. Suppose at round t we are considering querying triple (i, j, n). Conditioning on a non-tie preference o_t amounts to adding a single rank-one term to the inverse covariance:

$$\Sigma_{t+1}^{-1} \approx \Sigma_t^{-1} \; + \; \alpha_{ij,t}\, \phi_{ij}\phi_{ij}^\top, \qquad \alpha_{ij,t} \; = \; p_{ij,t}(1 - p_{ij,t}),$$

where $p_{ij,t} = \sigma(\hat{w}_{t-1,i} - \hat{w}_{t-1,j})$ is the BT predicted probability that model i beats model j under the current MAP. The approximation comes from the fact that we evaluate $\alpha_{ij,t}$ at the current MAP $\hat{w}_{t-1}$ rather than the (yet-unknown) post-update MAP $\hat{w}_t$; this is standard for one-step active learning under generalized linear models (Lewi et al., 2007). Applying the Sherman-Morrison identity (Sherman, 1949) to invert, the corresponding approximate covariance update is

$$\Sigma_{t+1} \approx \Sigma_t - \frac{\alpha_{ij,t}\,(\Sigma_t\phi_{ij})(\Sigma_t\phi_{ij})^\top}{1 + \alpha_{ij,t}\,\phi_{ij}^\top\Sigma_t\phi_{ij}}. \tag{4}$$

For any challenger $k \in \mathcal{C}_t$, the decisive margin is $m_{k,t} := w_{\hat{i}_{t-1}} - w_k = \phi_{\hat{i}_{t-1}k}^\top w$. The posterior variance of this margin before querying (i, j, n) is $\phi_{\hat{i}_{t-1}k}^\top\Sigma_t\phi_{\hat{i}_{t-1}k}$ from Eq. (3). Combining with Eq. (4), the one-step reduction in variance in the decisive margin of k after acquiring (i, j, n) is then

$$\Delta V_t(k; i, j) := \phi_{\hat{i}_{t-1}k}^\top\Sigma_t\phi_{\hat{i}_{t-1}k} - \phi_{\hat{i}_{t-1}k}^\top\Sigma_{t+1}\phi_{\hat{i}_{t-1}k} \approx \frac{\alpha_{ij,t}\left(\phi_{\hat{i}_{t-1}k}^\top\Sigma_t\phi_{ij}\right)^2}{1 + \alpha_{ij,t}\,\phi_{ij}^\top\Sigma_t\phi_{ij}}. \tag{5}$$

Triple Selection We score a triple (i, j, n) by summing this variance reduction over all currently active challengers, so comparisons are prioritized according to their total expected effect on the decisive margins. Then, we weight by the dissimilarity tensor from Eq. (1):

$$A_t(i, j, n) = D_{ijn} \sum_{k \in \mathcal{C}_t} \frac{\alpha_{ij,t}\left(\phi_{\hat{i}_{t-1}k}^\top\Sigma_t\phi_{ij}\right)^2}{1 + \alpha_{ij,t}\,\phi_{ij}^\top\Sigma_t\phi_{ij}}. \tag{6}$$

The acquisition is evaluated only for pairs in $\mathcal{P}_t$ and for triples that have not already been queried. Thus, a triple receives a high score only when three conditions hold simultaneously: the two model outputs are sufficiently different on prompt n, the current BT estimate does not already strongly favor one model over the other (the $\alpha_{ij,t}$ term), and the comparison is informative about one or more leader-vs-challenger margins that still determine the identity of the best model.

To select a triple using these acquisition scores, we sample from the candidate set with probability proportional to the acquisition score:

$$\Pr(q_t = (i, j, n)) \propto A_t(i, j, n). \tag{7}$$

If $\mathcal{C}_t = \emptyset$ at any point, then no remaining challenger can plausibly overtake the current leader under the posterior, and the algorithm could terminate early by returning the current best estimated model. In our experiments, however, we evaluate all methods under a fixed annotation budget, so TOPS instead keeps the two non-leaders with the largest $\hat{w}_{t-1,i}$ as challengers and continues with the acquisition rule.

In practice, in order to obtain theoretical guarantees in Section 5, we also include a small exploration probability $p_{\text{exp}} = 0.05$. With probability p_{exp}, TOPS ignores the active set and explores: it draws a model pair uniformly at random from all $\binom{K}{2}$ pairs and then a prompt uniformly at random among those not yet queried for that pair. After observing the preference outcome, we add the new comparison to $\mathcal{H}_t$, refit the BT MAP estimate, recompute the Laplace covariance, and update the leader, active challenger set, and acquisition scores. At the end of the budget, TOPS returns $\hat{i}_T = \arg\max_i \hat{w}_{T,i}$.

An algorithm block can be found in Appendix A, and a toy example can be found in Figure 3.

5 THEORY

We study the heterogeneous binary preference model of Section 3.

Assumption 5.1 (Bounded Heterogeneous Bradley-Terry Model)**.** Let $K \geq 2$, $W > 0$, $\sum_i w_i = 0$, $\|w\|_\infty \leq W$, and $\mathbb{E}_n[\varepsilon_{i,n}] = 0$, with n uniform over the pool. The winner $i^* = \arg\max_i w_i$ is unique, with gaps $\Delta_j = w_{i^*} - w_j$.

Theorem 5.1 (Cumulative Regret of TOPS)**.** Under Assumption 5.1 and the analysis protocol in Appendix C.1, fix $\delta \in (0, 1)$, $\varepsilon > 0$, and exploration probability $p > 0$. Suppose every suboptimal model has gap $\Delta_j > \gamma$. For a universal constant $C > 0$, with probability at least $1-\delta$, simultaneously for all $T \geq 1$,

$$R_T^{\mathrm{lat}} := \sum_{t=1}^{T} \Delta_{\hat{\imath}_t} \leq 2W \min\{T, T_\varepsilon\} + \frac{Cm\beta_T^2}{(1-p)q_m\alpha_W^2} \sum_{j \in N_\varepsilon} \frac{\Delta_j}{g_j^2}$$

Here R_T^{lat} is the cumulative regret, which accumulates the score gaps of the recommendations $\hat{\imath}_t = \arg\max_i \hat{w}_{t,i}$. The first term bounds regret before some T_ε. After this time, the leader and credible challengers are restricted to the winner and $m = |N_\epsilon|$ near competitors. These competitors have gaps at most ε after subtracting the heterogeneity bias bound γ, giving $g_j = \Delta_j - \gamma$. The second term bounds subsequent regret.

The confidence factor β_T grows with $O(\sqrt{\log T})$, and α_W measures the sensitivity of preference probabilities to score differences. The quantity q_m lower-bounds the probability of an informative leader comparison on an active round with remaining credible challengers. After T_ε, this condition holds whenever the credible set is nonempty. Formal definitions and the confidence thresholds required for the guarantee can be found in Appendix C.1.

Since $\beta_T^2 = O(\log T)$, we obtain $O(\log T)$ regret for fixed instance parameters. The bound reflects two key components of the top-focused design of TOPS: challenger-set pruning and proportional sampling. Further, Theorem C.6 bounds the time to a correct certified recommendation. Note that both guarantees use the latent scores w, instead of the fitted full-pool scores w^* used in our experiments. Corollary D.0.1 gives a sufficient condition for the latent winner to agree with the full-pool BT winner. All proofs can be found in Appendix C.

6 Experiments

Candidate LLMs We use $K = 20$ candidate LLMs, spanning a range of local and API-based models. The full list can be found in Appendix E.1.

Datasets and Annotators We present results on five datasets. To mimic human annotators, we experiment with a range of realistic and deliberately pathological annotators, designed to mimic both ordinary annotator noise and systematic human biases such as length bias (Hu et al., 2023; Hosking et al., 2023). Importantly, these annotators need not be perfect. Our goal is not to identify the objectively best model under an external ground-truth notion of quality, but rather the model preferred by the given annotator on the target prompt pool. We do not utilize the ground truths in any of the datasets. We present results for the following combinations in the main paper, and descriptions of other annotators and their results can be found in Appendix E.3 and Appendix F.10 respectively.

1. `AlpacaEval` (AE) (Dubois et al., 2023): Open-ended instruction-following prompts. As AlpacaEval contains five distinct subsets from different data sources, we evaluate on each one separately: `helpful`, `koala`, `oasst`, `selfinstruct`, and `vicuna`. For each one, we directly prompt GPT-5.2 with the two responses for which response it prefers.
2. `WildBench` (WB) (Lin et al., 2025): Open-ended prompts from real users across 11 categories, 27% of which are multi-turn conversations. As with AlpacaEval, we evaluate each category separately. We prompt Qwen-2.5-72B with the two responses and the per-instance checklist using the official pairwise template, and ask which response it prefers.
3. `MedCase` (Wu et al., 2025): Clinical case-triage prompts with free-text diagnoses. Here, we use GPT-5.2 to play the role of the clinical annotator. We prompt for what it thinks the diagnosis is, and then we prompt it to judge whether each candidate model's free-text diagnosis matches this reference diagnosis. The induced pairwise preference is assigned to the model whose response matches GPT-5.2's diagnosis.

We additionally evaluate on two datasets with a single correct answer, `GSM8K` (Cobbe et al., 2021) and `MMLU-PRO` (Wang et al., 2024b). Their annotator setup is described in Appendix E.3 and their results are reported in Appendix F.10.

Baselines We compare TOPS against the following baselines adapted to our setting: (1) **Random**, which uniformly randomly selects a triple, (2) **RUCB** (Relative Upper Confidence Bound) (Zoghi et al., 2014), is a Condorcet-based dueling-bandit algorithm that maintains confidence intervals over pairwise win probabilities, (3) **DuelingEXP3** (Saha et al., 2021), an exponential-weighting algorithm for adversarial dueling bandits, (4) **Borda**, Thompson sampling (Lekang & Lamperski, 2019) for dueling bandits with Borda winners, (5) **BALD** (Bayesian active learning by disagreement) (Houlsby et al., 2011), which selects triples by expected posterior entropy reduction, (6) **FewerBattles** (Liu et al.), the D-optimality variant which maximizes the determinant of the Fisher matrix, and (7) **GURO** (Greedy Uncertainty Reduction for Ordering) (Bergström et al., 2024), which minimizes expected variance similar to TOPS but without the active challenger set or the prompt axis. We also compare against two pointwise active-evaluation baselines, UCB-E and UCB-E-LRF (Zhou et al., 2025), which query a 1-10 score for a single (model, prompt) cell instead of a pairwise preference (Appendix F.2). More details about baseline implementations can be found in Appendix E.2.

Evaluation Metric We evaluate the final regret (**FR**) of the chosen model $r_{\hat{i}_T} = w^*_{i^\star} - w^*_{\hat{i}_T}$. As a practitioner may not know the budget T ahead of time, we additionally evaluate the cumulative regret (**CR**), which measures the quality of model selected along the whole trajectory: $R_T = \sum_{t=1}^{T} \left(w^*_{i^\star} - w^*_{\hat{i}_t}\right)$. Finally, we show the **rank** of the final selected model among the 20 candidates (in the appendix). We run each algorithm 10 times with different random seeds, and report metrics as mean $\pm$ one standard error. Unless otherwise stated, we use budget $T = 500$ and pool size $N = 500$, or all available prompts if the dataset or subset contains fewer than 500. Dataset and subset sizes are listed in Appendix Table 4.

7 RESULTS

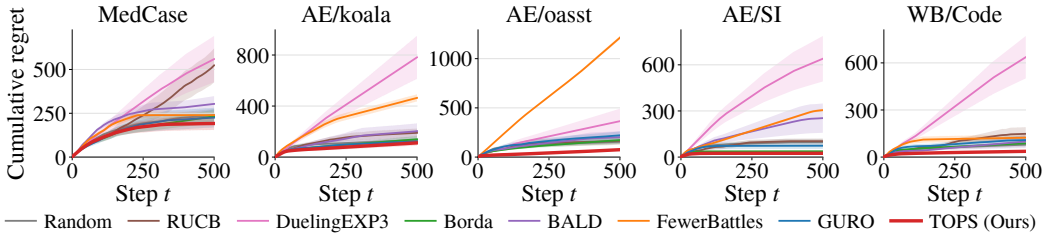


Figure 2: Cumulative regret as a function of t for five select settings. All 17 settings are shown in Appendix Figure 4. Bands show one standard error over 10 seeds.

Fixed-budget performance. Table 1 shows final and cumulative regret at $T = 500$ on `MedCase`, the five `AlpacaEval` sources, and selected `WildBench` categories. Appendix Table 15 reports all 17 settings and includes final ranks. Averaged over the 17 settings, TOPS has the best mean rank on all three metrics. For cumulative regret, it is top-2 on every setting and best on 14 of 17, and on final regret and final rank it is top-2 on 14 of 17 settings. Results for GSM8K and MMLU-Pro and for the other more pathological annotators are in Appendix F.10. Paired tests favor TOPS over all baselines on all three metrics ($p \leq 0.016$ after Holm correction, Appendix F.6). In our ablation experiments, we find that removing the active challenger set causes the largest and most consistent degradation (Appendix Table 5). TOPS also has lower cumulative regret than the pointwise baselines UCB-E and UCB-E-LRF (Zhou et al., 2025) (Appendix Table 7).

Performance over time. Figure 2 and Appendix Figure 4 show the cumulative regret as a function of time. We find that TOPS has an advantage across the whole budget, with the clearest advantage appearing around $t \approx 100 - 300$. At $T = 500$, TOPS identifies the fitted full-pool winner in 62% of runs, compared with at most 46% for any baseline (Appendix Table 11). In Appendix F.9, we report a small proof-of-concept experiment with a human annotator, in which TOPS identifies the best model with fewer human annotations than Borda and GURO.

Table 1: Final regret (FR) and cumulative regret (CR) at $T = 500$, mean $\pm$ 1 SEM over 10 seeds. Lower is better. Some Wildbench categories are omitted for space. The complete results, including final rank, are in Appendix Table 15. `Rank Avg.` averages method ranks over all six settings in Panel A, all 11 WildBench categories (`WB`), or all 17 settings (`All 17`), with SEM across settings. The best two values per column are bold, including ties. Ranks use displayed values, with average ranks for ties.

Panel A: GPT-5.2 annotator

Metric	Algorithm	MedCase	AE/HB	AE/koala	AE/oasst	AE/SI	AE/vicuna	Rank Avg.
FR (↓)	Random	0.122 ±0.074	0.120 ±0.068	**0.082 ±0.027**	0.133 ±0.048	0.002 ±0.001	0.291 ±0.098	**3.58 ±0.45**
	RUCB	1.597 ±0.333	0.224 ±0.113	0.114 ±0.025	**0.116 ±0.065**	**0.001 ±0.001**	0.305 ±0.103	4.25 ±1.05
	DuelingEXP3	0.831 ±0.281	1.229 ±0.252	1.484 ±0.372	0.617 ±0.270	0.779 ±0.321	1.444 ±0.417	7.67 ±0.21
	Borda	0.241 ±0.083	0.276 ±0.088	0.173 ±0.050	0.202 ±0.087	**0.001 ±0.000**	0.319 ±0.108	5.42 ±0.69
	BALD	0.192 ±0.139	0.190 ±0.097	0.371 ±0.183	**0.117 ±0.026**	0.120 ±0.079	0.166 ±0.085	4.50 ±0.76
	FewerBattles	**0.000 ±0.000**	**0.000 ±0.000**	0.577 ±0.000	2.530 ±0.066	0.063 ±0.059	0.278 ±0.113	4.58 ±1.19
	GURO	0.069 ±0.039	0.259 ±0.120	**0.081 ±0.027**	0.251 ±0.088	0.002 ±0.001	**0.125 ±0.084**	3.75 ±0.85
	TOPS (Ours)	**0.017 ±0.017**	**0.000 ±0.000**	0.148 ±0.016	0.151 ±0.017	**0.000 ±0.000**	**0.000 ±0.000**	**2.25 ±0.57**
CR (↓)	Random	238.7 ±36.4	194.7 ±19.1	132.5 ±23.9	**155.2 ±22.7**	97.7 ±21.6	266.0 ±55.4	3.83 ±0.54
	RUCB	523.5 ±96.7	277.3 ±37.7	190.0 ±33.1	164.8 ±39.5	103.5 ±16.3	207.0 ±41.6	5.00 ±0.58
	DuelingEXP3	559.1 ±132.3	697.1 ±121.7	781.8 ±167.6	364.9 ±126.8	639.8 ±148.1	944.6 ±199.7	7.83 ±0.17
	Borda	232.5 ±38.8	187.9 ±45.8	139.9 ±26.1	172.7 ±38.6	**37.7 ±5.6**	224.3 ±53.9	3.50 ±0.43
	BALD	303.5 ±44.5	302.9 ±49.6	203.5 ±61.4	202.2 ±60.1	253.4 ±95.4	316.9 ±87.9	6.17 ±0.31
	FewerBattles	239.2 ±19.6	**112.7 ±16.6**	463.9 ±27.9	1213.3 ±17.5	305.1 ±8.6	191.5 ±16.8	5.33 ±0.99
	GURO	**225.1 ±56.0**	197.5 ±30.7	**122.5 ±15.0**	221.4 ±39.1	73.7 ±20.8	**185.7 ±38.3**	**3.33 ±0.71**
	TOPS (Ours)	**191.7 ±36.6**	**69.6 ±13.1**	**111.3 ±19.5**	**74.9 ±11.1**	**24.5 ±4.9**	**65.3 ±20.7**	**1.00 ±0.00**

Panel B: Qwen-2.5-72B annotator, WildBench categories

Metric	Algorithm	WB/Info	WB/Code	WB/Creat	⋯	WB/Role	WB/Brain	WB/Advice	Rank Avg. WB	Rank Avg. All 17
FR (↓)	Random	0.101 ±0.051	0.056 ±0.013	0.227 ±0.052		0.183 ±0.068	0.247 ±0.102	0.233 ±0.106	4.00 ±0.38	**3.85 ±0.29**
	RUCB	0.313 ±0.101	0.066 ±0.012	0.493 ±0.073		0.268 ±0.049	0.259 ±0.093	0.096 ±0.096	5.45 ±0.37	5.03 ±0.44
	DuelingEXP3	1.314 ±0.401	1.139 ±0.332	0.735 ±0.162		1.289 ±0.172	0.845 ±0.332	1.138 ±0.262	7.91 ±0.09	7.82 ±0.10
	Borda	0.362 ±0.049	0.084 ±0.018	**0.104 ±0.056**		0.188 ±0.070	0.246 ±0.088	0.085 ±0.085	**3.55 ±0.58**	4.21 ±0.49
	BALD	0.320 ±0.111	0.145 ±0.115	0.193 ±0.086	⋯	0.189 ±0.070	**0.221 ±0.097**	0.137 ±0.070	4.36 ±0.43	4.41 ±0.37
	FewerBattles	**0.000 ±0.000**	0.053 ±0.010	0.901 ±0.000		0.390 ±0.065	0.334 ±0.000	**0.000 ±0.000**	5.05 ±0.79	4.88 ±0.64
	GURO	0.274 ±0.089	**0.046 ±0.015**	0.298 ±0.107		**0.106 ±0.035**	0.363 ±0.148	0.298 ±0.107	4.00 ±0.65	3.91 ±0.50
	TOPS (Ours)	**0.035 ±0.035**	**0.034 ±0.003**	**0.142 ±0.033**		**0.064 ±0.032**	**0.114 ±0.114**	**0.000 ±0.000**	**1.68 ±0.36**	**1.88 ±0.30**
CR (↓)	Random	166.2 ±33.3	**79.1 ±13.5**	158.4 ±23.0		170.1 ±29.8	245.5 ±27.6	222.0 ±40.5	3.55 ±0.39	3.65 ±0.31
	RUCB	229.0 ±49.9	148.1 ±42.3	225.2 ±29.0		203.5 ±14.6	234.0 ±19.3	225.5 ±43.3	5.64 ±0.41	5.41 ±0.33
	DuelingEXP3	781.5 ±182.3	636.4 ±135.2	497.6 ±91.1		618.7 ±69.9	467.8 ±151.4	689.4 ±96.1	7.91 ±0.09	7.88 ±0.08
	Borda	218.4 ±19.9	87.5 ±9.2	**133.4 ±24.2**		**136.8 ±26.7**	221.7 ±30.9	133.0 ±41.0	**2.64 ±0.24**	**2.94 ±0.23**
	BALD	268.8 ±52.0	99.3 ±41.7	158.8 ±39.1	⋯	220.2 ±35.6	255.6 ±37.4	210.9 ±50.2	5.27 ±0.41	5.59 ±0.30
	FewerBattles	**76.3 ±15.4**	123.7 ±19.8	631.8 ±20.0		281.3 ±5.9	**171.0 ±15.7**	**63.2 ±6.4**	5.09 ±0.77	5.18 ±0.59
	GURO	270.8 ±33.1	108.5 ±26.3	191.3 ±33.4		165.5 ±21.0	252.1 ±57.5	200.4 ±34.4	4.64 ±0.41	4.18 ±0.39
	TOPS (Ours)	**161.2 ±34.3**	**36.6 ±4.8**	**104.8 ±19.0**		**110.1 ±20.5**	**153.8 ±36.6**	**122.1 ±26.7**	**1.27 ±0.14**	**1.18 ±0.10**

Scaling N and K. Appendix Tables 8 and 9 report performance as we vary the pool size and number of candidate models. TOPS has the best average rank for cumulative regret at every tested pool size from $N = 100$ to $N = 1000$. With $K = 10$ and $K = 20$, TOPS has the best average rank on all three metrics, though simpler baselines perform better with $K = 5$.

8 Conclusion

In this work, we introduced TOPS, an active acquisition strategy for selecting the best LLM on a fixed unlabeled prompt pool from pairwise preferences.

Limitations First, other than the small proof-of-concept, we use LLM-based and synthetic annotators in place of human annotators to evaluate our algorithms. This enables cheap, reproducible comparisons across many datasets, budgets, and random seeds, but it does not fully capture the variability or inconsistency of real human annotators. A larger human-subjects study in a real deployment setting is one direction of future work. Second, our selection target is relative to the specified annotator. Systematic annotator bias can change this target, and thus the identified winner may not guarantee agreement with users in deployment. Finally, we assume candidate model responses can be generated and cached in advance, so that annotation cost is the main bottleneck. This is appropriate in settings where human preference labels are much more expensive than model inference.

AI USE STATEMENT

In this work, we used generative AI tools to assist with implementing and debugging experiment code, and for proving mathematical statements. For experiment code, authors reviewed the generated code for correctness and also checked the results for consistency. For mathematical proofs, authors verified the correctness of the proofs. We have not used generative AI tools for any other tasks. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

The methods in this paper are designed to make rigorous LLM evaluation cheaper, which we view as a net-positive contribution. Practitioners deploying models in high-stakes domains (medical triage, legal summarization, education) will be able to verify which model best fits their specific use case using budgets that were previously prohibitive, reducing the risk of choosing a poorly-suited model on the basis of generic public leaderboards. However, by reducing the cost of pairwise human-preference labeling, our methods could accelerate RLHF-style fine-tuning pipelines, including those used to train models for purposes the original annotators would not endorse. In addition, our offline cache-based pipeline normalizes the practice of evaluating models against fixed, curated prompt pools rather than the messier distribution of real user queries, and there may be still misspecification between the two. We caution against blindly deploying models selected by our algorithm without considering the validity of the underlying prompts.

REPRODUCIBILITY STATEMENT

The full algorithm can be found in Appendix A, with the derivations of its acquisition and covariance updates in Appendix B. The assumptions underlying our theoretical results are stated in Section 5, and the proofs are provided in Appendix C. Appendix E provides details on the baseline implementations, hyperparameters, datasets, and compute setup. The anonymized code is provided in the supplementary of this submission.

REFERENCES

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24, 2011.

Ahmed Alaa, Thomas Hartvigsen, Niloufar Golchini, Shiladitya Dutta, Frances Dean, Inioluwa Deborah Raji, and Travis Zack. Position: Medical large language model benchmarks should prioritize construct validity. In *Forty-second International Conference on Machine Learning Position Paper Track*, 2025.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

Herman Bergström, Emil Carlsson, Devdatt Dubhashi, and Fredrik D Johansson. Active preference learning for ordering items in-and out-of-sample. *Advances in Neural Information Processing Systems*, 37:71962–71997, 2024.

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

Erik Brynjolfsson, Danielle Li, and Lindsey Raymond. Generative ai at work. *The Quarterly Journal of Economics*, 140(2):889–942, 2025.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E Gonzalez, et al. Chatbot arena: An open platform for evaluating llms by human preference. *arXiv preprint arXiv:2403.04132*, 2024.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Paula Cordero Encinar, Taylan Cemgil, Arnaud Doucet, Virginia Aglietti, and Silvia Chiappa. BayesAME: Bayesian active model evaluation. *arXiv preprint arXiv:2607.27023*, 2026.

Nirjhar Das, Souradip Chakraborty, Aldo Pacchiano, and Sayak Ray Chowdhury. Active preference optimization for sample efficient rlhf. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 96–112. Springer, 2025.

Erik Daxberger, Agustinus Kristiadi, Alexander Immer, Runa Eschenhagen, Matthias Bauer, and Philipp Hennig. Laplace redux-effortless bayesian deep learning. *Advances in neural information processing systems*, 34:20089–20103, 2021.

Yann Dubois, Chen Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy S Liang, and Tatsunori B Hashimoto. Alpacafarm: A simulation framework for methods that learn from human feedback. *Advances in Neural Information Processing Systems*, 36:30039–30069, 2023.

Miroslav Dudík, Katja Hofmann, Robert E Schapire, Aleksandrs Slivkins, and Masrour Zoghi. Contextual dueling bandits. In *Conference on Learning Theory*, pp. 563–587. PMLR, 2015.

Maria Eriksson, Erasmo Purificato, Arman Noroozian, Joao Vinagre, Guillaume Chaslot, Emilia Gomez, and David Fernandez-Llorca. Can we trust ai benchmarks? an interdisciplinary review of current issues in ai evaluation. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 8, pp. 850–864, 2025.

Moein Falahatgar, Yi Hao, Alon Orlitsky, Venkatadheeraj Pichapati, and Vaishakh Ravindrakumar. Maxing and ranking with few assumptions. *Advances in Neural Information Processing Systems*, 30, 2017.

Adam Fisch, Daniel Deutsch, Joshua Maynez, Alekh Agarwal, Jonathan Berant, William Cohen, Amir Globerson, and Jacob Eisenstein. Collabeval: Statistically efficient collaborative model evaluation via matrix completion. *arXiv preprint arXiv:2607.05046*, 2026.

Amelia Hardy, Anka Reuel, Kiana Jafari Meimandi, Lisa Soder, Allie Griffith, Dylan M Asmar, Sanmi Koyejo, Michael S Bernstein, and Mykel John Kochenderfer. More than marketing? on the information value of ai benchmarks for practitioners. In *Proceedings of the 30th International Conference on Intelligent User Interfaces*, pp. 1032–1047, 2025.

Reinhard Heckel, Nihar B Shah, Kannan Ramchandran, and Martin J Wainwright. Active ranking from pairwise comparisons and when parametric assumptions do not help. 2019.

Tom Hosking, Phil Blunsom, and Max Bartolo. Human feedback is not gold standard. *arXiv preprint arXiv:2309.16349*, 2023.

Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. Bayesian active learning for classification and preference learning. *arXiv preprint arXiv:1112.5745*, 2011.

Yebowen Hu, Kaiqiang Song, Sangwoo Cho, Xiaoyang Wang, Hassan Foroosh, and Fei Liu. Decipherpref: Analyzing influential factors in human preference judgments via gpt-4. In *Proceedings of the 2023 Conference on Empirical Methods in natural language processing*, pp. 8344–8357, 2023.

Kevin G Jamieson and Robert Nowak. Active ranking using pairwise comparisons. *Advances in neural information processing systems*, 24, 2011.

Robert Korom, Sarah Kiptinness, Najib Adan, Kassim Said, Catherine Ithuli, Oliver Rotich, Boniface Kimani, Irene King'ori, Stellah Kamau, Elizabeth Atemba, et al. Ai-based clinical decision support for primary care: A real-world study. *arXiv preprint arXiv:2507.16947*, 2025.

Tyler Lekang and Andrew Lamperski. Simple algorithms for dueling bandits. *arXiv preprint arXiv:1906.07611*, 2019.

Jeremy Lewi, Robert Butera, and Liam Paninski. Efficient active learning with generalized linear models. In *Artificial Intelligence and Statistics*, pp. 267–274. PMLR, 2007.

Jiachun Li, David Simchi-Levi, and Will Wei Sun. Low rank for rank: Uncertainty-aware task-specific llm ranking under sparse pairwise comparisons. *arXiv preprint arXiv:2605.29395*, 2026.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*, 2022.

Bill Yuchen Lin, Yuntian Deng, Khyathi Chandu, Faeze Brahman, Abhilasha Ravichander, Valentina Pyatkin, Nouha Dziri, Ronan Le Bras, and Yejin Choi. WildBench: Benchmarking LLMs with challenging tasks from real users in the wild. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/771155abaae744e08576f1f3b4b7ac0d-Paper-Conference.pdf.

Zirui Liu, Xianquan Wang, Yan Zhuang, Jiatong Li, Qi Liu, Shuanghong Shen, Mingyue Cheng, and Shijin Wang. Fewer battles, more gain: An information-efficient framework for arena-based llm evaluation. In *The Fourteenth International Conference on Learning Representations*.

Adian Liusie, Potsawee Manakul, and Mark Gales. Llm comparative assessment: Zero-shot nlg evaluation through pairwise comparisons using large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 139–151, 2024.

Rafał K Mantiuk, Anna Tomaszewska, and Radosław Mantiuk. Comparison of four subjective methods for image quality assessment. In *Computer graphics forum*, volume 31, pp. 2478–2491. Wiley Online Library, 2012.

Viraj Mehta, Vikramjeet Das, Ojash Neopane, Yijia Dai, Ilija Bogunovic, Jeff Schneider, and Willie Neiswanger. Sample efficient reinforcement learning from human feedback via active exploration. 2023.

William Muldrew, Peter Hayes, Mingtian Zhang, and David Barber. Active preference learning for large language models. *arXiv preprint arXiv:2402.08114*, 2024.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

Felipe Maia Polo, Lucas Weber, Leshem Choshen, Yuekai Sun, Gongjun Xu, and Mikhail Yurochkin. tinybenchmarks: evaluating llms with fewer examples. *arXiv preprint arXiv:2402.14992*, 2024.

Mina Remeli and Moritz Hardt. Correct looks better: Pairwise comparisons reveal accuracy rankings. In *Proceedings of the Forty-Third International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2606.09409.

Aadirupa Saha, Tomer Koren, and Yishay Mansour. Adversarial dueling bandits. In *International Conference on Machine Learning*, pp. 9235–9244. PMLR, 2021.

Olawale Salaudeen, Anka Reuel, Ahmed Ahmed, Suhana Bedi, Zachary Robertson, Sudharsan Sundar, Ben Domingue, Angelina Wang, and Sanmi Koyejo. Measurement to meaning: A validity-centered framework for ai evaluation. *arXiv preprint arXiv:2505.10573*, 2025.

Nigam H Shah, Nerissa Ambers, Abby Pandya, Timothy Keyes, Juan M Banda, Srikar Nallan, Carlene Lugtu, Artem A Trotsyuk, Suhana Bedi, Alyssa Unell, et al. Adoption and use of llms at an academic medical center. *arXiv preprint arXiv:2602.00074*, 2026.

Jack Sherman. Adjustment of an inverse matrix corresponding to changes in the elements of a given column or a given row of the original matrix. *Annals of mathematical statistics*, 20(4):621, 1949.

Chengshuai Shi, Kun Yang, Zihan Chen, Jundong Li, Jing Yang, and Cong Shen. Efficient prompt optimization through the lens of best arm identification. *Advances in Neural Information Processing Systems*, 37:99646–99685, 2024.

Neil Stewart, Gordon DA Brown, and Nick Chater. Absolute identification by relative judgment. *Psychological review*, 112(4):881, 2005.

Elad Tolochinsky, Yaniv Tenzer, and Yaniv Romano. Valid best-model identification for llm evaluation via low-rank factorization. *arXiv preprint arXiv:2605.10405*, 2026.

Joel A Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, 12(4):389–434, 2012.

Sang T Truong, Yuheng Tu, Percy Liang, Bo Li, and Sanmi Koyejo. Reliable and efficient amortized model-based evaluation. In *International Conference on Machine Learning*, pp. 60238–60265. PMLR, 2025.

Rajan Vivek, Kawin Ethayarajh, Diyi Yang, and Douwe Kiela. Anchor points: Benchmarking models with much fewer examples. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1576–1601, 2024.

Rose E Wang, Ana T Ribeiro, Carly D Robinson, Susanna Loeb, and Dora Demszky. Tutor copilot: A human-ai approach for scaling real-time expertise. *arXiv preprint arXiv:2410.03017*, 2024a.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Advances in Neural Information Processing Systems*, 37: 95266–95290, 2024b.

Huasen Wu and Xin Liu. Double thompson sampling for dueling bandits. *Advances in neural information processing systems*, 29, 2016.

Kevin Wu, Eric Wu, Rahul Thapa, Kevin Wei, Angela Zhang, Arvind Suresh, Jacqueline J Tao, Min Woo Sun, Alejandro Lozano, and James Zou. Medcasereasoning: Evaluating and learning diagnostic reasoning from clinical case reports. *arXiv preprint arXiv:2505.11733*, 2025.

Yisong Yue and Thorsten Joachims. Beat the mean bandit. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pp. 241–248, 2011.

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.

Haoran Zhang, Hyewon Jeong, Olawale Elijah Salaudeen, Walter Gerych, Nigam Shah, and Marzyeh Ghassemi. Position: Benchmarks cannot establish deployment readiness of clinical ai. In *Forty-third International Conference on Machine Learning Position Paper Track*, 2026.

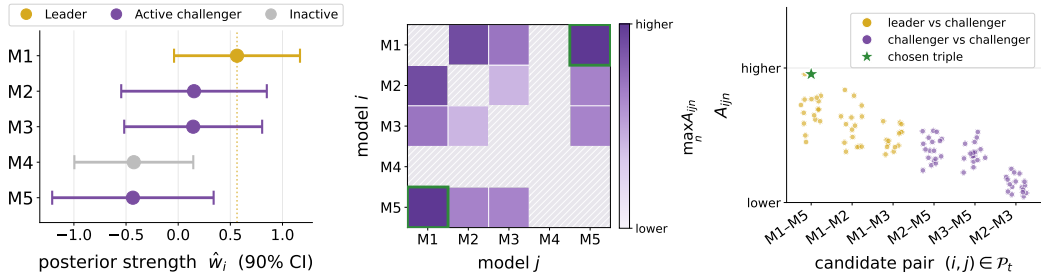
Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. *Advances in neural information processing systems*, 36:46595–46623, 2023.

Jin Zhou, Christian Belardi, Ruihan Wu, Travis Zhang, Carla Gomes, Wen Sun, and Kilian Weinberger. On speeding up language model evaluation. In *International Conference on Learning Representations*, volume 2025, pp. 65092–65111, 2025.

Masrour Zoghi, Shimon Whiteson, Remi Munos, and Maarten Rijke. Relative upper confidence bound for the k-armed dueling bandit problem. In *International conference on machine learning*, pp. 10–18. PMLR, 2014.

Masrour Zoghi, Zohar S Karnin, Shimon Whiteson, and Maarten De Rijke. Copeland dueling bandits. *Advances in neural information processing systems*, 28, 2015.

A TOPS ALGORITHM



(a) (b) (c)

Figure 3: A toy demonstration of our method. (a) We maintain a posterior over w (bars: 90% credible intervals on each w_i) and only consider as challengers the models whose lower credible bound on the margin to the leader is non-positive. (b) We evaluate the acquisition function A_{ijn} (Eq. 6) only on candidate pairs. (c) We sample a prompt-model-pair triple with probability proportional to A_{ijn}; here the triple that maximizes A_{ijn} is shown.

Algorithm 1 TOPS: Top-focused Online Preference Sampling

Require: Models $\{1, \dots, K\}$; prompts $\{1, \dots, N\}$; cached responses $\{y_{i,n}\}$; dissimilarity tensor $D \in \mathbb{R}^{K \times K \times N}$; budget T; prior variance $\sigma_0^2 = 1$; posterior tail probability δ_{cr}; exploration probability p_{exp}.

Initialize history $\mathcal{H}_0 \leftarrow \emptyset$.
for $t = 1, 2, \dots, T$ **do**
  $u \sim \text{Uniform}(0, 1)$
  if $u < p_{\text{exp}}$ **or** $\mathcal{H}_{t-1}$ contains no non-tie outcome **then**
    Sample a pair (i_t, j_t) uniformly over all $\binom{K}{2}$ pairs, then n_t uniformly among the prompts not yet queried for that pair.
  else
    Fit BT MAP $\hat{w}_{t-1}$ and Laplace covariance Σ_t from $\mathcal{H}_{t-1}$ using Eq. 2.
    $\hat{\imath}_{t-1} \leftarrow \arg\max_i \hat{w}_{t-1,i}$ ▷ empirical leader
    $\mathcal{C}_t \leftarrow \left\{ j \neq \hat{\imath}_{t-1} : (\hat{w}_{t-1,\hat{\imath}_{t-1}} - \hat{w}_{t-1,j}) - z_{\text{cr}}\, \sigma_{\hat{\imath}_{t-1} j,\, t} \leq 0 \right\}$ ▷ active challengers
    if $\mathcal{C}_t = \emptyset$ **then** $\mathcal{C}_t \leftarrow$ the two models $j \neq \hat{\imath}_{t-1}$ with the largest $\hat{w}_{t-1,j}$
    $\mathcal{P}_t \leftarrow \{(\hat{\imath}_{t-1}, j) : j \in \mathcal{C}_t\} \cup \{(j, k) : j, k \in \mathcal{C}_t,\ j < k\}$
    Compute $A_t(i, j, n)$ via Eq. 6 for each $(i, j) \in \mathcal{P}_t$ and each n such that (i, j, n) has not been queried in prior rounds.
    Sample (i_t, j_t, n_t) with probability proportional to $A_t(i_t, j_t, n_t)$ over the unused triples in $\mathcal{P}_t$.
  end if
  Query the annotator on $(y_{i_t,n_t},\, y_{j_t,n_t})$ to obtain preference $o_t \in \{-1, 0, +1\}$ (+1: first model wins, −1: second model wins, 0: tie). .
  $\mathcal{H}_t \leftarrow \mathcal{H}_{t-1} \cup \{(i_t, j_t, n_t, o_t)\}$.
end for
Fit final BT MAP $\hat{w}_T$ from $\mathcal{H}_T$.
return $\hat{\imath}_T = \arg\max_i \hat{w}_{T,i}$.

B DERIVATIONS OF EXPRESSIONS IN SECTION 4

B.1 EQUATION (3)

Here, we derive the Laplace covariance Σ_t in Eq. (2) and the closed-form margin distribution

$$(w_i - w_j) \mid \mathcal{H}_{t-1} \;\approx\; \mathcal{N}\big(\hat{w}_{t-1,i} - \hat{w}_{t-1,j},\ \phi_{ij}^\top \Sigma_t \phi_{ij}\big)$$

in Eq. (3), starting from the BT log-posterior in Section 4.

Setup. We sum only over non-tie observations. Recall that the BT likelihood for a single observation $(q_\tau = (i_\tau, j_\tau, n_\tau),\, o_\tau)$ with $o_\tau \in \{-1, +1\}$ and $\phi_\tau = e_{i_\tau} - e_{j_\tau}$ is

$$P(o_\tau \mid w) \;=\; \sigma\big(o_\tau\, \phi_\tau^\top w\big).$$

Combined with the isotropic prior $w \sim \mathcal{N}(0, I)$, the log-posterior at round t is

$$\ell_t(w) \;:=\; \log p(w \mid \mathcal{H}_{t-1}) \;=\; \sum_{\tau<t} \log \sigma\big(o_\tau\, \phi_\tau^\top w\big) \;-\; \tfrac{1}{2}\,\|w\|_2^2 \;+\; \text{const}, \tag{8}$$

and the MAP estimate $\hat{w}_{t-1}$ is its argmax.

Hessian of the log-posterior. The first term in Eq. (8) has gradient and Hessian

$$\nabla_w \log \sigma(o_\tau \phi_\tau^\top w) = o_\tau \big(1 - \sigma(o_\tau \phi_\tau^\top w)\big)\, \phi_\tau,$$
$$\nabla_w^2 \log \sigma(o_\tau \phi_\tau^\top w) = -\sigma(o_\tau \phi_\tau^\top w)\big(1 - \sigma(o_\tau \phi_\tau^\top w)\big)\, \phi_\tau \phi_\tau^\top.$$

Since $\sigma(-x) = 1 - \sigma(x)$ and $\sigma(x)(1-\sigma(x)) = \sigma(-x)(1-\sigma(-x))$, the Hessian factor is invariant in o_τ:

$$\sigma(o_\tau \phi_\tau^\top w)\big(1 - \sigma(o_\tau \phi_\tau^\top w)\big) \;=\; \sigma(\phi_\tau^\top w)\big(1 - \sigma(\phi_\tau^\top w)\big).$$

Adding the prior contribution $-\nabla_w^2 \tfrac{1}{2}\|w\|^2 = -I$ and negating, the negative Hessian of ℓ_t at any w is

$$-\nabla_w^2 \ell_t(w) \;=\; \sum_{\tau<t} \alpha_\tau(w)\, \phi_\tau \phi_\tau^\top \;+\; I, \quad \text{with} \quad \alpha_\tau(w) := \sigma(\phi_\tau^\top w)\big(1 - \sigma(\phi_\tau^\top w)\big). \tag{9}$$

This is positive definite (the first term is PSD and the prior contributes a strictly positive I), so ℓ_t is strictly concave and $\hat{w}_{t-1}$ is unique.

Laplace approximation. Around the MAP, a second-order Taylor expansion of the log-posterior gives

$$\ell_t(w) \;\approx\; \ell_t(\hat{w}_{t-1}) \;-\; \tfrac{1}{2}\,(w - \hat{w}_{t-1})^\top \big(-\nabla_w^2 \ell_t(\hat{w}_{t-1})\big)(w - \hat{w}_{t-1}).$$

Exponentiating identifies the leading Gaussian approximation

$$w \mid \mathcal{H}_{t-1} \;\approx\; \mathcal{N}(\hat{w}_{t-1},\, \Sigma_t), \qquad \Sigma_t^{-1} \;=\; -\nabla_w^2 \ell_t(\hat{w}_{t-1}).$$

Substituting Eq. (9) evaluated at $\hat{w}_{t-1}$ and writing $\alpha_{\tau,t} := \alpha_\tau(\hat{w}_{t-1})$ recovers Eq. (2):

$$\Sigma_t^{-1} \;=\; \sum_{\tau<t} \alpha_{\tau,t}\, \phi_\tau \phi_\tau^\top \;+\; I, \qquad \alpha_{\tau,t} \;=\; \sigma\big(\phi_\tau^\top \hat{w}_{t-1}\big)\big(1 - \sigma\big(\phi_\tau^\top \hat{w}_{t-1}\big)\big). \tag{10}$$

Margin distribution. The decisive quantity for best-arm identification is the score gap $w_i - w_j$, which is an affine function of w:

$$w_i - w_j \;=\; (e_i - e_j)^\top w \;=\; \phi_{ij}^\top w.$$

Affine maps preserve Gaussianity. Specifically, if $w \sim \mathcal{N}(\mu, \Sigma)$ and $a \in \mathbb{R}^K$ is a fixed vector, then $a^\top w \sim \mathcal{N}(a^\top \mu,\, a^\top \Sigma a)$. Applying this with $\mu = \hat{w}_{t-1}$, $\Sigma = \Sigma_t$, and $a = \phi_{ij}$ to the Laplace posterior gives

$$(w_i - w_j) \mid \mathcal{H}_{t-1} \;\approx\; \mathcal{N}\big(\phi_{ij}^\top \hat{w}_{t-1},\; \phi_{ij}^\top \Sigma_t\, \phi_{ij}\big). \tag{11}$$

The mean simplifies to $\phi_{ij}^\top \hat{w}_{t-1} = \hat{w}_{t-1,i} - \hat{w}_{t-1,j}$. The variance can be expanded coordinate-wise:

$$\begin{aligned}
\phi_{ij}^\top \Sigma_t\, \phi_{ij} &= (e_i - e_j)^\top \Sigma_t (e_i - e_j) \\
&= e_i^\top \Sigma_t e_i + e_j^\top \Sigma_t e_j - 2\, e_i^\top \Sigma_t e_j \\
&= \Sigma_{t,ii} + \Sigma_{t,jj} - 2\,\Sigma_{t,ij},
\end{aligned}$$

which matches Eq. (3).

B.2 EQUATION (5)

We now derive Eq. (5), the closed form for the one-step reduction in the posterior variance of the decisive margin $w_{i_{t-1}} - w_k$ after acquiring a hypothetical triple (i, j, n). Conditioning on a non-tie preference adds a single rank-one term to the inverse covariance,

$$\Sigma_{t+1}^{-1} \;\approx\; \Sigma_t^{-1} + \alpha_{ij,t}\,\phi_{ij}\phi_{ij}^\top,$$

where the approximation comes from evaluating $\alpha_{ij,t}$ at the current MAP $\hat{w}_{t-1}$ rather than the post-update MAP. Applying the Sherman–Morrison identity (Sherman, 1949) with $A = \Sigma_t^{-1}$, $u = \sqrt{\alpha_{ij,t}}\,\phi_{ij}$, $v = \sqrt{\alpha_{ij,t}}\,\phi_{ij}$ inverts this rank-one update in closed form:

$$\Sigma_{t+1} \;\approx\; \Sigma_t \;-\; \frac{\alpha_{ij,t}\,(\Sigma_t\phi_{ij})(\Sigma_t\phi_{ij})^\top}{1 + \alpha_{ij,t}\,\phi_{ij}^\top\Sigma_t\phi_{ij}}. \tag{12}$$

Sandwiching both sides with $\phi_{i_{t-1}k}^\top \cdot \phi_{i_{t-1}k}$ converts the matrix update into a scalar update on the margin variance:

$$\phi_{i_{t-1}k}^\top\Sigma_{t+1}\,\phi_{i_{t-1}k} \;\approx\; \phi_{i_{t-1}k}^\top\Sigma_t\,\phi_{i_{t-1}k} \;-\; \frac{\alpha_{ij,t}\,\phi_{i_{t-1}k}^\top(\Sigma_t\phi_{ij})(\Sigma_t\phi_{ij})^\top\phi_{i_{t-1}k}}{1 + \alpha_{ij,t}\,\phi_{ij}^\top\Sigma_t\phi_{ij}}.$$

The numerator simplifies because $(\Sigma_t\phi_{ij})^\top\phi_{i_{t-1}k} = \phi_{i_{t-1}k}^\top\Sigma_t\phi_{ij}$ is a scalar (using $\Sigma_t = \Sigma_t^\top$), so the bilinear form collapses to the square of a single inner product:

$$\phi_{i_{t-1}k}^\top(\Sigma_t\phi_{ij})(\Sigma_t\phi_{ij})^\top\phi_{i_{t-1}k} \;=\; \big(\phi_{i_{t-1}k}^\top\Sigma_t\phi_{ij}\big)^2.$$

Rearranging gives the variance reduction in Eq. (5):

$$\Delta V_t(k; i, j) \;:=\; \phi_{i_{t-1}k}^\top\Sigma_t\,\phi_{i_{t-1}k} \;-\; \phi_{i_{t-1}k}^\top\Sigma_{t+1}\,\phi_{i_{t-1}k} \;\approx\; \frac{\alpha_{ij,t}\,\big(\phi_{i_{t-1}k}^\top\Sigma_t\phi_{ij}\big)^2}{1 + \alpha_{ij,t}\,\phi_{ij}^\top\Sigma_t\phi_{ij}}. \quad \square$$

C PROOF OF THEOREMS

C.1 ANALYSIS SETUP

Throughout this appendix, we use Assumption 5.1, fix $\delta \in (0, 1)$, and write $\alpha = \alpha_W = \sigma'(2W)$. The confidence factor and credible multiplier are

$$\beta_t = \sqrt{\frac{2\log(3/\delta) + K\log(1 + 2\alpha t/K)}{\alpha}} + 2W\sqrt{K}, \qquad z_t = \frac{\beta_t}{2\sqrt{\alpha}} \tag{13}$$

After t labels, write $\phi_e = e_i - e_j$ for pair $e = \{i, j\}$ with $i < j$. Query s has vector ϕ_s and outcome y_s, indicating that its first model wins. We sample prompts with replacement and draw fresh outcomes independently of past outcomes conditional on the pair and prompt. The constrained MAP and covariance use a unit Gaussian prior on all K coordinates:

$$\begin{aligned} L_t(u) &= \tfrac{1}{2}\|u\|_2^2 + \sum_{s\le t}\{\log(1 + e^{\phi_s^\top u}) - y_s\phi_s^\top u\}, \qquad \hat{w}_t = \arg\min_{u\in[-W,W]^K} L_t(u), \\ \widehat{\Sigma}_t &= \left(I_K + \sum_{s\le t}\sigma'(\phi_s^\top\hat{w}_t)\phi_s\phi_s^\top\right)^{-1}, \qquad \widehat{\sigma}_e(t)^2 = \phi_e^\top\widehat{\Sigma}_t\phi_e \end{aligned} \tag{14}$$

Define

$$\widetilde{H}_t = I_K + \alpha\sum_{s\le t}\phi_s\phi_s^\top, \qquad \widetilde{v}_e(t) = \phi_e^\top\widetilde{H}_t^{-1}\phi_e, \qquad n_e(t) = \sum_{s\le t}\mathbf{1}\{e_s = e\}$$

For every u in the box, $|\phi_s^\top u| \le 2W$, so $\nabla^2 L_t(u) \succeq \widetilde{H}_t$.

Let $\ell_t = \hat{\imath}_t = \arg\max_i \hat{w}_{t,i}$, breaking ties by index, and set

$$\mathcal{C}_t = \{k \ne \ell_t : \hat{w}_{t,\ell_t} - \hat{w}_{t,k} \le z_t\widehat{\sigma}_{\ell_t k}(t)\}$$

For acquisition, $\mathcal{C}_t^{\mathrm{qry}} = \mathcal{C}_t$ unless it is empty, in which case we use the $\min\{2, K-1\}$ highest-scoring non-leaders. This fallback leaves $\mathcal{C}_t$ unchanged. Candidate pairs form the complete graph on $\{\ell_t\} \cup \mathcal{C}_t^{\mathrm{qry}}$.

An independent coin selects a uniform pair and prompt with probability $p = p_{\mathrm{exp}} \in (0,1)$. Otherwise, we use Equation (6) with query set $\mathcal{C}_t^{\mathrm{qry}}$:

$$\pi_t(e) \propto \bar{D}_e A_t(e) \quad (e \in \mathcal{P}_t), \qquad \bar{D}_{ij} = \sum_n D_{ijn} \tag{15}$$

Here $A_t(e)$ sums the variance-reduction score in Equation (5) over $\mathcal{C}_t^{\mathrm{qry}}$, evaluated at $(\hat{w}_t, \widehat{\Sigma}_t, \ell_t)$. The conditional prompt law is $q_{ij}^D(n) = D_{ijn}/\bar{D}_{ij}$. We assume fixed symmetric nonnegative D with $\chi_D := \min_{i<j} \bar{D}_{ij} / \max_{i<j} \bar{D}_{ij} > 0$, so every pair has positive prompt mass. The box constraint, all-coordinate prior, calibrated multiplier, and fresh binary sampling specify the analyzed variant. Section 4 uses the practical anchored fit, the fixed multiplier $z_{\mathrm{cr}} = 1.645$, and unused cached triples.

For the prompt residuals, define

$$\eta_{ijn} = \varepsilon_{i,n} - \varepsilon_{j,n}, \qquad \rho^2 = \max_{i \neq j} \mathbb{E}_n[\eta_{ijn}^2],$$

$$\kappa_D = \max_{i \neq j} \left| \sum_n q_{ij}^D(n) \eta_{ijn} \right|, \qquad \rho_D^2 = \max_{i \neq j} \sum_n q_{ij}^D(n) \eta_{ijn}^2,$$

$$\beta_D = \max\{\rho^2/20,\ \kappa_D/4 + \rho_D^2/20\}$$

For a fixed deterministic $\varepsilon > 0$, define

$$\gamma = \frac{(K-1)\beta_D}{\alpha}, \qquad g_j = \Delta_j - \gamma, \qquad N_\varepsilon = \{j \neq i^* : g_j \leq \varepsilon\}, \qquad m = |N_\varepsilon|$$

Thus $\gamma = 0$ for homogeneous prompts. The coverage constants are

$$q_0 = q_1 = 1, \qquad q_r = \frac{8\chi_D \alpha}{3(r+1)(2 + (r-1)^2/16)} \quad (r \geq 2) \tag{16}$$

Lemma C.4 proves the associated sampling guarantees. Let $a = t_{\mathrm{burn}}$ be the smallest positive integer such that

$$t \geq \frac{14(K-1)}{p} \log \frac{6Kt^2}{\delta} \qquad \text{for every integer } t \geq a$$

The localization time is

$$\varepsilon_t = \frac{2\beta_t}{\alpha} \sqrt{\frac{K-1}{pt}}, \qquad T_\varepsilon = \min\{t_0 \in \mathbb{N} : t_0 \geq a,\ \varepsilon_t \leq \varepsilon \text{ for every integer } t \geq t_0\} \tag{17}$$

Both times are finite because $\beta_t^2 = O(\log t)$. Neither ε nor γ is an algorithm input.

Let $\mathcal{F}_t$ contain the first t branches, pairs, prompts, and outcomes. Let $\mathcal{F}_t'$ additionally reveal the branch and pair of round $t+1$, before its prompt and outcome. Then $\mathcal{F}_t \subseteq \mathcal{F}_t' \subseteq \mathcal{F}_{t+1}$. Write $\bar{p}_s = \mathbb{E}[y_s \mid \mathcal{F}_{s-1}']$ and $b_s = \bar{p}_s - \sigma(\phi_s^\top w)$.

C.2 Shared Lemmas

We first record the exploration bound. Realize independent coins $c_s \sim \mathrm{Bernoulli}(p)$ and uniform pairs E_s in advance. Let Q have orthonormal columns spanning $\mathbf{1}^\perp$ and set $X_s = c_s Q^\top \phi_{E_s} \phi_{E_s}^\top Q$. These matrices satisfy $0 \preceq X_s \preceq 2I$ and $\mathbb{E}X_s = 2pI/(K-1)$. Matrix Chernoff (Tropp, 2012, Theorem 1.1), with deviation $1/2$, gives

$$\Pr\left(\lambda_{\min}\left(\sum_{s \leq t} X_s \right) < \frac{pt}{K-1} \right) \leq (K-1) e^{-pt/[7(K-1)]} \leq \frac{\delta}{6t^2} \qquad (t \geq a)$$

The full design includes these exploration observations. Inverting on $\mathbf{1}^\perp$ and taking a union bound over $t \geq a$ therefore gives an event E_{design} of probability at least $1 - \delta/3$ on which

$$\widetilde{v}_e(t) \leq \frac{2}{1 + \alpha pt/(K-1)} \leq \frac{4(K-1)}{\alpha pt} \qquad (t \geq a, \text{ all pairs } e) \tag{18}$$

Lemma C.1 (Perturbation of the Bradley-Terry Fit)**.** Let $L(u)$ be the objective in Equation (14) for arbitrary real labels. Let $\hat{u}$ minimize $L(u)$ and let $\tilde{u}$ minimize $L(u)+\sum_s d_s\phi_s^\top u$, both on $[-W,W]^K$. If $|d_s| \le b$, then

$$\max_{i,j} |(\hat{u}_i - \tilde{u}_i) - (\hat{u}_j - \tilde{u}_j)| \le \frac{(K-1)b}{\alpha}$$

Proof. Let $v = \hat{u} - \tilde{u}$, and let $n, \tilde{n}$ be box-normal vectors at the two minimizers. Subtracting the first-order conditions and integrating the Hessian along their connecting segment gives

$$v + \sum_s h_s \phi_s \phi_s^\top v + n - \tilde{n} = \sum_s d_s \phi_s, \qquad h_s \in [\alpha, 1/4]$$

If $v_i > 0$, then $\hat{u}_i > -W$ and $\tilde{u}_i < W$, so $n_i \ge 0 \ge \tilde{n}_i$. Sort the K coordinates of v and suppose an adjacent gap exceeds b/α. If its upper endpoint is positive, sum the equations over all vertices above this gap. Their prior terms have strictly positive sum, and their normal-vector terms have nonnegative sum. Internal comparison edges cancel. If n_{cut} observed edges cross the cut, their contribution exceeds bn_{cut} when this count is positive. The right side is at most bn_{cut}. When the cut has no observed edge, the positive prior term alone gives a contradiction. Replacing v by $-v$, swapping $n, \tilde{n}$, and negating each d_s handles the remaining case. Hence every adjacent gap is at most b/α, and summing the $K-1$ gaps proves the claim without a connectedness assumption. □

Lemma C.2 (Simultaneous Pairwise Confidence)**.** There is an event E_{noise} with probability at least $1-\delta/3$ on which, for every $t \ge 0$ and pair $e = \{i, j\}$,

$$|\phi_e^\top(\hat{w}_t - w)| \le \beta_t \sqrt{\widetilde{v}_e(t)} + \gamma \tag{19}$$

Write $E = E_{\mathrm{noise}} \cap E_{\mathrm{design}}$, so $\Pr(E) \ge 1 - 2\delta/3$.

Proof. Taylor expansion, $\sigma' \le 1/4$, and $|\sigma''| \le 1/(6\sqrt{3})$ give, for any prompt law q,

$$|\mathbb{E}_q \sigma(w_i - w_j + \eta_{ijn}) - \sigma(w_i - w_j)| \le \tfrac{1}{4}|\mathbb{E}_q \eta_{ijn}| + \tfrac{1}{20}\mathbb{E}_q \eta_{ijn}^2 \tag{20}$$

Thus $|b_s| \le \beta_D$ under either sampling branch, including fallback rounds. Uniform prompts have zero mean residual and bias at most $\rho^2/20$.

Put $S_t = \sum_{s \le t}(y_s - \bar{p}_s)\phi_s$. Under the interleaved filtration, ϕ_s is predictable and $y_s - \bar{p}_s$ is conditionally $1/2$-sub-Gaussian. The self-normalized bound of Abbasi-Yadkori et al. (2011, Theorem 1), with initial matrix $\alpha^{-1}I_K$ and the determinant–trace inequality, gives an event E_{noise} of probability at least $1-\delta/3$ on which $2\|S_t\|_{\widetilde{H}_t^{-1}} \le \beta_t - 2W\sqrt{K}$ for every t. Here $\mathrm{tr}(\sum_{s \le t} \phi_s \phi_s^\top) = 2t$.

Let $\tilde{w}_t$ fit the corrected labels $y_s - b_s$ with the same box and prior. Its objective has gradient $w - S_t$ at w and Hessian at least $\widetilde{H}_t$. Strong convexity and feasibility of w give

$$\|\tilde{w}_t - w\|_{\widetilde{H}_t} \le 2\|w - S_t\|_{\widetilde{H}_t^{-1}} \le \beta_t$$

Lemma C.1 bounds each contrast of $\hat{w}_t - \tilde{w}_t$ by γ. Cauchy–Schwarz proves the claim. □

Lemma C.3 (Active-Set Properties)**.** Assume $\Delta_{\min} := \min_{j \ne i^*} \Delta_j > \gamma$ and work on E_{noise}. The following hold simultaneously for all $t \ge 0$, with part (ii) additionally requiring E_{design}.

(i) $i^* \in \{\ell_t\} \cup \mathcal{C}_t$.

(ii) For $t \ge T_\varepsilon$, $(\{\ell_t\} \cup \mathcal{C}_t) \setminus \{i^*\} \subseteq N_\varepsilon$, and $|\mathcal{C}_t| \le m$.

(iii) If $\ell_t = i^*$ and $k \in \mathcal{C}_t$, then $n_{\{i^*,k\}}(t) < \beta_t^2/(\alpha^2 g_k^2)$.

(iv) If $\ell_t = j \ne i^*$ and $k \in \mathcal{C}_t$ satisfies $\widehat{\sigma}_{jk}(t) \ge \widehat{\sigma}_{ji^*}(t)/2$, then $n_{\{j,k\}}(t) < \beta_t^2/(\alpha^2 g_j^2)$.

Proof. The curvature bounds give $\widetilde{H}_t \preceq \widehat{\Sigma}_t^{-1} \preceq (4\alpha)^{-1}\widetilde{H}_t$, and $\widetilde{H}_t \succeq I_K + \alpha n_e(t)\phi_e\phi_e^\top$. Inversion and Sherman–Morrison give, for every history,

$$\begin{aligned} &4\alpha\widetilde{H}_t^{-1} \preceq \widehat{\Sigma}_t \preceq \widetilde{H}_t^{-1} \preceq I_K, \qquad \widehat{\sigma}_e(t)^2 \le 2, \\ &\widetilde{v}_e(t) \le \frac{2}{1 + 2\alpha n_e(t)}, \qquad n_e(t) < \frac{1}{\alpha \widetilde{v}_e(t)} \end{aligned} \tag{21}$$

Together with the multiplier definition, these imply

$$z_t\widehat{\sigma}_e(t) \geq \beta_t\sqrt{\widetilde{v}_e(t)}, \qquad z_t + \beta_t \leq \beta_t/\sqrt{\alpha}$$

For (i), if $\ell_t = j \neq i^*$, confidence gives

$$\hat{w}_{t,j} - \hat{w}_{t,i^*} \leq -g_j + \beta_t\sqrt{\widetilde{v}_{ji^*}(t)} < z_t\widehat{\sigma}_{ji^*}(t)$$

Thus $i^* \in \mathcal{C}_t$.

For (ii), a wrong leader j has $\Delta_j \leq \gamma + \beta_t\sqrt{\widetilde{v}_{ji^*}(t)}$. For a challenger $j \neq i^*$, leader optimality and membership give $\hat{w}_{t,i^*} - \hat{w}_{t,j} \leq \hat{w}_{t,\ell_t} - \hat{w}_{t,j} \leq z_t\widehat{\sigma}_{\ell_t j}(t)$. Confidence and the covariance bound therefore give

$$\Delta_j \leq \gamma + (z_t + \beta_t)\max_e \sqrt{\widetilde{v}_e(t)} \leq \gamma + \frac{2\beta_t}{\alpha}\sqrt{\frac{K-1}{pt}} = \gamma + \varepsilon_t$$

The same final bound applies to a wrong leader. For $t \geq T_\varepsilon$, these models belong to N_ε, proving the cardinality bound.

For (iii), the preceding challenger argument with $\ell_t = i^*$ gives $g_k \leq (z_t + \beta_t)\sqrt{\widetilde{v}_{i^*k}(t)}$. Hence $\widetilde{v}_{i^*k}(t) \geq \alpha g_k^2/\beta_t^2$, and Equation (21) bounds the count. For (iv), confidence and the wrong leader's optimality give $\widetilde{v}_{ji^*}(t) \geq g_j^2/\beta_t^2$. The covariance sandwich and the stated width condition imply

$$\widetilde{v}_{jk}(t) \geq \widehat{\sigma}_{jk}(t)^2 \geq \tfrac{1}{4}\widehat{\sigma}_{ji^*}(t)^2 \geq \alpha\widetilde{v}_{ji^*}(t) \geq \alpha g_j^2/\beta_t^2$$

Apply the count bound again. □

Lemma C.4 (Coverage of Proportional Sampling)**.** Fix a history with $r = |\mathcal{C}_t| \geq 1$. For any $k^\circ \in \mathcal{C}_t$, define $L = \{k \in \mathcal{C}_t : \widehat{\sigma}_{\ell_t k}(t) \geq \widehat{\sigma}_{\ell_t k^\circ}(t)/2\}$. The active pair law satisfies

$$\pi_t(\{\{\ell_t, k\} : k \in L\}) \geq q_r$$

Here q_r is defined in Equation (16); it is nonincreasing for integers $r \geq 1$. If $r = 1$, the only candidate pair has probability one.

Proof. The case $r = 1$ is immediate. For $r \geq 2$, put $a_k = \phi_{\ell_t k}$ and define the positive semidefinite Gram matrix $G_{kk'} = a_k^\top\widehat{\Sigma}_t a_{k'}$. The diagonal entries are $G_{kk} = \widehat{\sigma}_{\ell_t k}(t)^2 > 0$. Using $\sigma'(\phi_e^\top\hat{w}_t) \in [\alpha, 1/4]$ and $\widehat{\sigma}_e^2 \leq 2$ gives

$$\frac{2\alpha}{3}\|G_{:,k}\|^2 \leq A_t(\{\ell_t, k\}) \leq \tfrac{1}{4}\|G_{:,k}\|^2, \qquad A_t(\{k, k'\}) \leq \tfrac{1}{4}\|G_{:,k'} - G_{:,k}\|^2$$

In particular, leader-pair scores are positive, so the normalization in Equation (15) is nonzero. For $x \in \mathbb{R}^r$, $\sum_{k<k'}(x_k - x_{k'})^2 = r\|x\|^2 - (\mathbf{1}^\top x)^2 \leq r\|x\|^2$. Applying this identity to each row of G gives

$$\sum_{e\in\mathcal{P}_t} A_t(e) \leq \frac{r+1}{4}\|G\|_F^2$$

For any pair subset S, the prompt masses give $\pi_t(S) \geq \chi_D\sum_{e\in S} A_t(e)/\sum_{e\in\mathcal{P}_t} A_t(e)$.

To bound the mass of L, set $s = \widehat{\sigma}_{\ell_t k^\circ}(t)$ and $U = \sum_{k\in L}\|G_{:,k}\|^2$. Since $k^\circ \in L$, $U \geq s^4$. By symmetry, the squared Frobenius norm of G_{L,L^c} is at most U. The principal block G_{L^c,L^c} is positive semidefinite and

$$\|G_{L^c,L^c}\|_F \leq \operatorname{tr}(G_{L^c,L^c}) \leq (r-1)s^2/4$$

Splitting G into its columns in L and the two remaining blocks yields

$$\|G\|_F^2 \leq \left(2 + \frac{(r-1)^2}{16}\right)U$$

The lower bound on leader-pair scores now gives the required mass $8\chi_D\alpha/[3(r+1)(2 + (r-1)^2/16)] = q_r$. The formula decreases for $r \geq 2$, and $q_2 \leq 1 = q_1$. □

For $\Delta_{\min} > \gamma$, write $A_j(T) = \beta_T^2/(\alpha^2 g_j^2)$ and define

$$\begin{aligned} B_T(\varepsilon) &= \sum_{j \in N_\varepsilon} \lceil A_j(T) \rceil + \sum_{\{j,k\} \subseteq N_\varepsilon} \lceil \max\{A_j(T), A_k(T)\} \rceil \\ &\le m \sum_{j \in N_\varepsilon} A_j(T) + \frac{m(m+1)}{2} \end{aligned} \tag{22}$$

The second sum runs over unordered pairs. The inequality uses $\lceil x \rceil \le x + 1$ and $\max\{A_j, A_k\} \le A_j + A_k$.

Lemma C.5 (Witness Counting)**.** Suppose $\Delta_{\min} > \gamma$ and $m \ge 1$. There is an event $E_{\mathrm{wit}} \subseteq E$ with probability at least $1 - \delta$ on which, simultaneously for every integer $T \ge 1$ and $j \in N_\varepsilon$,

$$\#\{T_\varepsilon \le t \le T : \ell_t = j\} \le \frac{5 m A_j(T)}{(1-p) q_m}, \tag{23}$$

$$\#\{T_\varepsilon \le t < T : \mathcal{C}_t \ne \varnothing\} \le \frac{1.6}{(1-p) q_m} \left\{ B_T(\varepsilon) + \log \frac{6}{\delta} \right\} \tag{24}$$

Proof. For nonempty $\mathcal{C}_t$, choose $k_t^\circ = i^*$ if $i^* \in \mathcal{C}_t$, and the smallest-index challenger otherwise. Define the witness pairs

$$\mathcal{W}_t = \{\{\ell_t, k\} : k \in \mathcal{C}_t,\ \widehat{\sigma}_{\ell_t k}(t) \ge \widehat{\sigma}_{\ell_t k_t^\circ}(t)/2\}$$

Set $\mathcal{W}_t = \varnothing$ when $\mathcal{C}_t = \varnothing$. Let $I_t = \mathbf{1}\{t \ge T_\varepsilon,\ 1 \le |\mathcal{C}_t| \le m\}$, $I_t^j = I_t \mathbf{1}\{\ell_t = j\}$, and let H_{t+1} indicate an active query in $\mathcal{W}_t$. For either $X_t = I_t$ or $X_t = I_t^j$, $Z_{t+1} = X_t H_{t+1}$ satisfies $\mathbb{E}[Z_{t+1} \mid \mathcal{F}_t] \ge q X_t$ on every history, where $q = (1-p) q_m$, by Lemma C.4.

Since Z_{t+1} is binary, $\mathbb{E}[e^{-Z_{t+1}} \mid \mathcal{F}_t] \le e^{-(1-e^{-1}) q X_t}$. Thus $M_n = \exp\{(1-e^{-1}) q \sum_{t<n} X_t - \sum_{t<n} Z_{t+1}\}$ is a nonnegative supermartingale starting at one. Ville's inequality gives, with probability at least $1 - \delta'$, simultaneously for all n,

$$\sum_{t<n} X_t \le \frac{1.6}{q} \left\{ \sum_{t<n} Z_{t+1} + \log \frac{1}{\delta'} \right\}$$

All sums start at $t = 0$. Apply this with $\delta' = \delta/6$ for I_t and $\delta' = \delta/[6(K-1)]$ for each I_t^j. Intersecting with E costs at most $2\delta/3 + \delta/6 + \delta/6 = \delta$.

On this intersection, Lemma C.3(i)–(ii) makes these indicators equal to the round indicators in the statement. Every witness query after T_ε lies in $\{i^*\} \cup N_\varepsilon$ and increments its pair count. Part (iv) limits witnesses with leader j through state T to $\lceil A_j(T) \rceil$ for each of at most m partners. For all witnesses through state $T-1$, parts (iii)–(iv) give caps $\lceil A_k(T) \rceil$ on $\{i^*, k\}$ and $\lceil \max\{A_j(T), A_k(T)\} \rceil$ on $\{j,k\} \subseteq N_\varepsilon$. Hence the two witness totals are bounded by $m\lceil A_j(T) \rceil$ and $B_T(\varepsilon)$ respectively. Charge state T to query $T+1$ of the continuing process, which does not change the first T recommendations.

The counting inequality now proves Equation (24). For Equation (23), use $A_j(T) \ge 16K$ and $\log(6K/\delta) \le A_j(T)$, which follow from the formula for β_T, $g_j \le 2W$, and $\alpha^3 g_j^2 \le 4W^2 e^{-6W} < 1$. Thus $1.6\{m\lceil A_j(T) \rceil + \log(6K/\delta)\} \le 5 m A_j(T)$. □

C.3 Proof of Theorem 5.1

Proof. Rounds before T_ε contribute at most $2W \min\{T, T_\varepsilon\}$. If $m = 0$, Lemma C.3(ii) gives zero subsequent regret on E. Otherwise, on E_{wit}, only leaders in N_ε incur subsequent regret. Multiplying Equation (23) by Δ_j and summing gives the stated bound with $C = 5$. □

Using only confidence and Equation (18) gives at most $a + 4\beta_T^2(K-1)/(\alpha p g_j^2)$ wrong-leader rounds for each suboptimal j, yielding the all-model sum discussed in Section 5.

C.4 Stopping-Time Identification

The stopping variant follows the continuing process until

$$\tau = \inf\{t \geq 0 : \mathcal{C}_t = \varnothing\} \tag{25}$$

It returns and retains ℓ_τ.

Theorem C.6 (Identification Time of TOPS)**.** Under the protocol of Appendix C.1, assume $\Delta_{\min} > \gamma$ and fix $\varepsilon > 0$. With probability at least $1 - \delta$, the stopping rule returns $\ell_\tau = i^*$. If $m = 0$, then $\tau \leq T_\varepsilon$. If $m \geq 1$, then $\tau \leq T$ for every integer T satisfying

$$T \geq T_\varepsilon + 1 + \frac{1.6}{(1-p)q_m}\left\{B_T(\varepsilon) + \log\frac{6}{\delta}\right\} \tag{26}$$

Such a finite T exists. This guarantee and Theorem 5.1 hold jointly with probability at least $1 - \delta$ on the coupled process.

Proof. Lemma C.3(i) makes every finite stop correct on E_{noise}. When $m = 0$, part (ii) gives $\tau \leq T_\varepsilon$ on E. For $m \geq 1$, work on E_{wit}. If $\tau > T$, every state $T_\varepsilon \leq t < T$ has a nonempty credible set, so Equation (24) contradicts Equation (26). Since $B_T = O(\log T)$, a finite such T exists. The regret bound holds on the same event. □

This guarantee concerns the recommendation retained at certification. It does not assert correctness of the continuing leader at a prescribed final round.

D Different Notions of the Best Model

For a fixed evaluator and prompt pool, note that the model with the highest underlying score may be different from the model selected by fitting scores to observed preferences, see Table 2.

Table 2: Two notions of the best model.

Winner	Meaning	Role in this paper
Latent winner	The model with the highest underlying pool-level score w_i in the assumed preference model. These scores describe how preferences are generated and are not directly observed.	The theoretical sections of the paper: The target of the regret and stopping-time guarantees in Section 5.
Full-pool winner	The model with the highest BT score fitted to observed comparisons.	The full-pool target first defined in Section 3. The target used in experiments. Corollary D.0.1 gives a sufficient condition for agreement with the latent winner, when a finite fit exists.

D.1 Target Equivalence on the Realized Pool

The latent target i^* can differ from the full-pool BT target $i^\star$ in Section 3. We give a sufficient condition for agreement.

Corollary D.0.1 (Target Equivalence on the Realized Pool)**.** Under Assumption 5.1, construct J from one independent binary judgment for each unordered pair and prompt, with $J_{jin} = -J_{ijn}$. For every $\delta \in (0, 1)$, if

$$\Delta_{\min} > \frac{2}{\alpha_W}\left\{\frac{\rho^2}{20} + \sqrt{\frac{\log(K^2/\delta)}{2N}}\right\}$$

then, with probability at least $1 - \delta$ over J, every finite maximizer w^* of the unregularized BT objective in Section 3, when one exists, satisfies $i^\star = \arg\max_i w_i^* = i^*$.

Proof. Let $\hat{p}_{ij} := N^{-1} \sum_n \mathbf{1}\{J_{ijn} = +1\}$ and $\hat{R}_i := \sum_{j \neq i} \hat{p}_{ij}$. Hoeffding's inequality, a union bound over the $K(K-1)/2$ unordered pairs, and the uniform-prompt case of Equation (20) give, with probability at least $1 - \delta$,

$$|\hat{p}_{ij} - \sigma(w_i - w_j)| \leq \varepsilon_N := \frac{\rho^2}{20} + \sqrt{\frac{\log(K^2/\delta)}{2N}} \qquad (i \neq j)$$

Write $R_i = \sum_{j \neq i} \sigma(w_i - w_j)$ and add the self-term to obtain $R_i + 1/2 = \sum_j \sigma(w_i - w_j)$. Since $\sigma' \geq \alpha_W$ on $[-2W, 2W]$, the mean value theorem gives $R_{i^*} - R_k \geq K \alpha_W \Delta_k$. Thus the stated margin condition implies, for every $k \neq i^*$,

$$\hat{R}_{i^*} - \hat{R}_k \geq K \alpha_W \Delta_{\min} - 2(K-1)\varepsilon_N > 0$$

At a finite maximizer of the full-pool objective on $\sum_i u_i = 0$, the gradient coordinates are equal and sum to zero by shift invariance. The likelihood equations give $\hat{R}_i + 1/2 = \sum_j \sigma(w_i^* - w_j^*)$ for every i. Since $u \mapsto \sum_j \sigma(u - w_j^*)$ is strictly increasing, w^* and $\hat{R}$ have the same ordering, proving $i^\star = i^*$. □

E Experimental Details

E.1 Candidate LLMs

We use the following 20 candidate LLMs:

```
meta-llama/Llama-3.1-8B-Instruct
google/gemma-7b-it
mistralai/Mistral-7B-Instruct-v0.2
Qwen/Qwen2.5-7B-Instruct
Qwen/Qwen3-8B
CohereLabs/aya-expanse-8b
microsoft/Phi-4-mini-instruct
meta-llama/Llama-3.2-3B-Instruct
Qwen/Qwen2.5-14B-Instruct
Qwen/Qwen3-14B
google/gemma-3-12b-it
google/gemma-4-E4B-it
microsoft/phi-4
deepseek-ai/DeepSeek-R1-Distill-Qwen-14B
Qwen/Qwen3.5-9B
ibm-granite/granite-3.3-8b-instruct
gpt-4.1-mini
gpt-5.4-nano
gpt-5.4-mini
Kimi-K2.5
```

E.2 Implementation Details and Hyperparameters

For each baseline, we describe the implementation and hyperparameters used in our experiments. Unless stated otherwise, all triple-selection methods avoid replaying a triple (i, j, n) that has already been queried in earlier rounds.

Our method. We use prior variance $\sigma_0^2 = 1$ for TOPS. We run TOPS as described in Section 4, with $p_{\text{exp}} = 0.05$, $\delta_{\text{cr}} = 0.05$, and `Qwen/Qwen3-Embedding-0.6B` as the embedding model.

Baselines. The seven baselines we compare against span dueling-bandit, active learning, and Bradley–Terry experimental-design methods; we adapt each to our triple-selection (i, j, n) setting if the original method only chooses pairs.

- **Random.** Each round, sample $(i, j) \sim \mathrm{Unif}\binom{[K]}{2}$ uniformly and independently, then sample n uniformly from the prompts not yet queried for this pair.

- **RUCB** (Zoghi et al., 2014). Relative UCB for the Condorcet winner. Maintains pairwise win-rate UCBs, identifies the Condorcet candidate set, samples a champion, and picks the most-threatening challenger. Prompt n is sampled uniformly. We use $\alpha_{\mathrm{RUCB}} = 0.51$ (the smallest value with the regret guarantee of Zoghi et al. (2014)).

- **DuelingEXP3** (Saha et al., 2021). The D-EXP3 algorithm for adversarial dueling bandits. Maintains a softmax-mixture distribution q_t over arms to sample from. Prompt n is sampled uniformly. We use $\eta = (\log K/\sqrt{100K})^{2/3}$ and $\gamma = \min(\sqrt{\eta K}, 0.25)$.

- **Borda** (Lekang & Lamperski, 2019). Thompson sampling for Borda-winner identification. Maintains independent Beta posteriors on each pairwise win probability. Each round, with probability p_{Borda} samples a random pair; otherwise draws two independent samples from the full preference matrix. Prompt n is sampled uniformly. We use $p_{\mathrm{Borda}} = 0.1$ and uniform $\mathrm{Beta}(1, 1)$ on each pairwise win rate.

- **BALD** (Houlsby et al., 2011). Bayesian active learning by disagreement, applied to pairwise BT comparisons. The pair with the highest score is selected. Prompt n is sampled uniformly. We use prior variance $\sigma_0^2 = 1.0$.

- **FewerBattles** (Liu et al.). D-optimal Bradley-Terry pair selection, adapted to triples by scaling the pair score by $D[i, j, n]$. We use prior variance $\sigma_0^2 = 1.0$.

- **GURO** (Bergström et al., 2024). Greedy Uncertainty Reduction for Ordering: a pair-only variance-minimization rule that uses the same primitives as TOPS but on the unprojected square root form, without the active challenger set or top-focus restriction. We adapt this to triples by multiplying by $D[i, j, n]$. We use prior variance $\sigma_0^2 = 1.0$.

Pointwise baselines. We additionally compare against the two pointwise active-evaluation methods of Zhou et al. (2025) (Appendix F.2): **UCB-E**, best-arm identification over models by the upper confidence bound on the mean pointwise score, and **UCB-E-LRF**, which additionally maintains an ensemble of low-rank factorizations of the (model $\times$ prompt) score matrix and picks the cell with the largest ensemble disagreement within the chosen arm. The Ground truth remains the full-matrix pairwise Bradley-Terry oracle of Section 3, so results are comparable with Table 1. UCB-E uses the adaptive exploration coefficient $a = c\,(T - K)/\hat{H}_1$ with $c = 25/36$. UCB-E-LRF uses an ensemble of 8 rank-2 factorizations fit on 80% bagged subsamples, and disagreement weight $\eta = 1$.

E.3 Full List of Datasets and Judges

In addition to the datasets of Section 6, we evaluate on two verifiable datasets. `GSM8K` (Cobbe et al., 2021): grade-school math word problems with numerical answers. Here, we use Qwen-2.5-72B to play the role of the annotator. We first prompt for what it thinks the answer is. The induced pairwise preference is assigned to the model whose response matches its answer. If both or neither response matches, the preference is a tie, which is treated as uninformative. `MMLU-PRO` (Wang et al., 2024b): multiple-choice STEM and humanities questions with 10 answer options. We use the same prompting setup as `GSM8K`. Table 3 lists all (judge, dataset) combination we evaluate.

Table 3: Full list of dataset and judge combinations that we evaluate methods on. AlpacaEval contains five source subsets: `helpful`, `koala`, `oasst`, `selfinstruct`, and `vicuna`, and WildBench is evaluated separately on each of its 11 categories. Bolded entries correspond to those shown in the main paper.

Judge	Dataset	Description
`local`	GSM8K MMLU-Pro MedCase	Queries Qwen-2.5-72B for its own answer and scores each response against it (a tie if both or neither match). Simulates a somewhat flawed human annotator.
	AlpacaEval	Queries Qwen-2.5-72B for which output it prefers.
	WildBench	Queries Qwen-2.5-72B for which output it prefers.
`api`	**MedCase** **AlpacaEval**	Queries GPT-5.2. Same role as `local` but with a stronger simulated annotator.
`length_biased`	GSM8K MMLU-Pro MedCase AlpacaEval WildBench	A trivial judge that simply selects the answer with the most tokens 90% of the time, and a random choice 10% of the time. A simulation for potential biases in human annotators (Hu et al., 2023).

Table 4: Numbers of prompts in the prepared datasets and subsets used in our experiments, before sampling the evaluation pool. AlpacaEval (AE) sources and WildBench (WB) categories are evaluated separately. By default, we sample 500 prompts without replacement from larger datasets and use all prompts from smaller datasets or subsets. All methods and seeds share the same pool within each setting. Experiments that vary pool size state their values separately.

Dataset / subset	Source / category	Dataset size
`GSM8K`		8,792
`MMLU-Pro`		12,032
`MedCase`		14,489
`AE/HB`	helpful_base	129
`AE/koala`	koala	156
`AE/oasst`	oasst	188
`AE/SI`	selfinstruct	252
`AE/vicuna`	vicuna	80
`WB/Info`	Information seeking	182
`WB/Code`	Coding and debugging	170
`WB/Creat`	Creative writing	146
`WB/Reas`	Reasoning	133
`WB/Plan`	Planning	116
`WB/Math`	Math	87
`WB/Edit`	Editing	82
`WB/Data`	Data analysis	33
`WB/Role`	Role playing	30
`WB/Brain`	Brainstorming	24
`WB/Advice`	Advice seeking	21

E.4 Compute Setup

We run our experiments on a shared Slurm cluster. All experiments used one NVIDIA A100 with 80 GB VRAM, 10 CPU cores of Intel Xeon Ice Lake Platinum 8368, and 80 GB RAM.

E.5 Prompts

This subsection lists the exact prompt templates used in our pipeline. Curly-braced placeholders (e.g. `{question}`, `{answer_a}`) are replaced at runtime with the corresponding content. All templates can be found in `llm_selection/constants.py`.

Response generation. For each candidate LLM and each prompt in the pool, we generate a single response using the dataset-specific template below.

GSM8K:

```
Please solve this math problem step by step. At the end, finish your
solution with "The answer is [number]":

{question}
```

MMLU-Pro:

```
Answer the following multiple choice question. Think step by step,
then finish your solution with "The answer is (X)" where X is one of
A, B, C, D, E, F, G, H, I, J.

{question}
```

MedCase:

```
Please solve this medical diagnosis case problem. At the end, finish
your solution with "The diagnosis is [diagnosis]":

{question}
```

Verifiable annotators For `GSM8K` and `MMLU`, the response is parsed and string-matched against the annotator's answer (numerical answer / letter A-J), so no judge prompt is needed. For `MedCase`, candidate and oracle diagnoses are free-text strings; we use a judge with the following template:

MedCase yes/no diagnosis match:

```
Given the true disease and a candidate disease diagnosis, answer with
"Yes" if the candidate diagnosis is accurate for the true disease,
otherwise answer "No".

True disease: {true_disease}
Candidate diagnosis: {candidate_diagnosis}
```

Pairwise annotators. For each pairwise dataset, both candidate responses $y_{i,n}$ and $y_{j,n}$ are shown to a judge LLM with the templates below.

AlpacaEval pairwise:

```
Select the output (a) or (b) that best matches the given instruction.
Choose your preferred output, which can be subjective. Your answer
should ONLY contain: Output (a) or Output (b). Here's an example:

# Example:
## Instruction:
Give a description of the following job: "ophthalmologist"

## Output (a):
An ophthalmologist is a medical doctor who specializes in the
diagnosis and treatment of eye diseases and conditions.

## Output (b):
An ophthalmologist is a medical doctor who pokes and prods at your
eyes while asking you to read letters from a chart.

## Which is best, Output (a) or Output (b)?
Output (a)
```

```
Here the answer is Output (a) because it provides a comprehensive and
accurate description of the job of an ophthalmologist. In contrast,
output (b) is more of a joke.

# Task:
Now is the real task, do not explain your answer, just say Output (a)
or Output (b).

## Instruction:
{question}

## Output (a):
{reference_answer}

## Output (b):
{model_response}

## Which is best, Output (a) or Output (b)?
```

For `WildBench`, we use the following adaptation of the WildBench v2 pairwise template (Lin et al., 2025). A++ and A+ count as a win for response A, B++ and B+ as a win for response B, and A=B as a tie.

WildBench pairwise:

```
# Instruction

You are an expert evaluator. Read the conversation history and the
current user query, examine the two candidate responses (A and B), and
judge which response better addresses the user using the
instance-specific checklist below as your primary guidance.

# Conversation between User and AI

## History

<|begin_of_history|>
{history}
<|end_of_history|>

## Current User Query

<|begin_of_query|>
{question}
<|end_of_query|>

## Response A

<|begin_of_response_A|>
{answer_a}
<|end_of_response_A|>

## Response B

<|begin_of_response_B|>
{answer_b}
<|end_of_response_B|>

# Evaluation

## Checklist

<|begin_of_checklist|>
{checklist}
<|end_of_checklist|>
```

```
Use this checklist as the principal lens for your evaluation; do not
restrict yourself to it -- also consider correctness, helpfulness,
completeness, instruction adherence, and safety. Avoid position bias,
length bias, and style bias.

## Rules

Compare the two responses based on the user query, the conversation
history, and the checklist. Decide which response is better according
to the following labels:

- "A++" : A is much better than B (≥ 5 clear advantages, whether large
or small).
- "A+"  : A is slightly better than B (1-4 advantages).
- "A=B" : tie -- no significant differences.
- "B+"  : B is slightly better than A (1-4 advantages).
- "B++" : B is much better than A (≥ 5 clear advantages).

## Output Format

First write a short analysis of each response, then summarize where
they are equivalent, where A is better, and where B is better, and
finally render the verdict. Return your answer ONLY as a JSON object
with this schema (fill in each placeholder; the "choice" must be one
of the five labels above):

{
    "analysis of A": "[concise analysis of Response A]",
    "analysis of B": "[concise analysis of Response B]",
    "reason of A=B": "[aspects where A and B perform equally]",
    "reason of A>B": "[aspects where A is better than B]",
    "reason of B>A": "[aspects where B is better than A]",
    "choice": "[A++ or A+ or A=B or B+ or B++]"
}
```

Pathological annotators. The `length_biased` annotator does not query an LLM: it returns $+1$ for whichever response has more whitespace-tokenized words 90% of the time, and a uniform random verdict the remaining 10%.

E.6 Dataset Licenses

`GSM8K` and `MMLU` are released under the MIT license. `AlpacaEval` is released under the Apache 2.0 license. `MedCase` is released under the CC-BY 4.0 license. `WildBench` is released under the CC BY license.

F Additional Experimental Results

F.1 Ablations

Table 5: Ablation of TOPS components, $T = 500$, 10 seeds, on MedCase and the five AlpacaEval sources with the GPT-5.2 annotator, and GSM8K and MMLU-Pro with the Qwen-2.5-72B annotator. Pool sizes follow Section 6. Each $-x$ row removes a single component from TOPS: $-$**fullcov** zeroes the off-diagonal entries of the Laplace covariance Σ_t before computing the active set and the acquisition; $-$**dissim** drops the dissimilarity tensor ($D \equiv 1$, uniformly random prompt within a pair); $-$**activeset** disables credible-interval pruning of the challenger set; $-$**topfocus** scores all $\binom{K}{2}$ pairs (no leader-vs-challenger restriction); $-$**explore** sets the exploration probability p_{exp} to 0. **FR**: final regret; **CR**: cumulative regret; **Rank**: true rank of the chosen model (1 = best). Lower is better. Cells show mean $\pm$ 1 SEM. `Rank Avg.`: mean per-setting rank against the 5 competing variants, averaged across these eight settings ($\pm$ 1 SEM across settings). Top-2 per column (within each metric) in bold. Ranks and bolding use the displayed values; tied values receive their average rank, and every cell tied at the second-smallest displayed value is bold.

Metric	Algorithm	MedCase	AE/HB	AE/koala	AE/oasst	AE/SI	AE/vicuna	GSM8K	MMLU-Pro	Rank Avg.
FR (↓)	TOPS	**0.017** ±0.017	**0.000** ±0.000	0.148 ±0.016	0.151 ±0.017	**0.000** ±0.000	**0.000** ±0.000	0.514 ±0.131	**0.071** ±0.017	**3.00** ±0.41
	−fullcov	**0.017** ±0.017	**0.000** ±0.000	0.148 ±0.016	0.167 ±0.000	**0.000** ±0.000	**0.000** ±0.000	**0.360** ±0.121	**0.076** ±0.021	**3.00** ±0.37
	−dissim	**0.000** ±0.000	**0.000** ±0.000	0.165 ±0.000	0.167 ±0.000	**0.000** ±0.000	**0.000** ±0.000	0.510 ±0.160	0.144 ±0.056	3.69 ±0.57
	−activeset	0.235 ±0.126	0.063 ±0.063	**0.124** ±0.057	**0.067** ±0.027	0.003 ±0.001	0.111 ±0.074	0.590 ±0.171	0.374 ±0.105	4.88 ±0.74
	−topfocus	0.118 ±0.026	**0.000** ±0.000	**0.082** ±0.027	**0.100** ±0.027	0.002 ±0.001	**0.000** ±0.000	0.414 ±0.175	0.085 ±0.021	3.12 ±0.48
	−explore	0.103 ±0.069	**0.000** ±0.000	0.148 ±0.016	0.167 ±0.000	**0.000** ±0.000	**0.000** ±0.000	**0.267** ±0.137	0.108 ±0.019	3.31 ±0.43
CR (↓)	TOPS	191.7 ±36.6	69.6 ±13.1	**111.3** ±19.5	74.9 ±11.1	**24.5** ±4.9	65.3 ±20.7	**355.0** ±40.9	124.5 ±12.5	2.75 ±0.41
	−fullcov	**141.6** ±41.1	73.2 ±15.9	121.4 ±24.9	101.2 ±11.4	37.0 ±7.9	**44.9** ±7.6	**342.7** ±56.7	**113.9** ±23.6	**2.62** ±0.56
	−dissim	**153.5** ±30.2	**49.1** ±10.6	**107.4** ±9.7	126.8 ±14.7	52.2 ±11.6	70.5 ±9.2	457.2 ±61.1	139.9 ±14.5	3.38 ±0.65
	−activeset	250.7 ±58.6	134.1 ±30.9	182.8 ±37.4	**54.9** ±14.4	84.1 ±22.8	133.8 ±31.3	467.8 ±69.9	223.6 ±34.3	5.38 ±0.62
	−topfocus	183.1 ±27.6	112.8 ±13.9	122.5 ±31.4	**66.4** ±11.5	41.8 ±12.8	82.9 ±23.8	462.9 ±90.5	164.0 ±22.3	4.38 ±0.37
	−explore	181.4 ±42.0	**60.3** ±11.9	111.9 ±19.1	91.8 ±8.3	**28.3** ±5.5	**62.6** ±15.8	384.8 ±62.3	**94.3** ±10.5	**2.50** ±0.33
Rank (↓)	TOPS	**1.10** ±0.10	**1.00** ±0.00	2.80 ±0.20	1.90 ±0.10	**1.00** ±0.00	**1.00** ±0.00	3.90 ±0.98	**2.20** ±0.29	**3.00** ±0.41
	−fullcov	**1.10** ±0.10	**1.00** ±0.00	2.80 ±0.20	2.00 ±0.00	**1.00** ±0.00	**1.00** ±0.00	**2.70** ±0.65	**2.30** ±0.40	**3.00** ±0.37
	−dissim	**1.00** ±0.00	**1.00** ±0.00	3.00 ±0.00	2.00 ±0.00	**1.00** ±0.00	**1.00** ±0.00	3.80 ±0.94	2.90 ±0.62	3.56 ±0.55
	−activeset	2.10 ±0.46	1.20 ±0.20	**2.10** ±0.38	**1.40** ±0.16	2.20 ±0.33	1.20 ±0.13	4.30 ±1.04	6.00 ±0.98	4.88 ±0.74
	−topfocus	1.70 ±0.15	**1.00** ±0.00	**2.00** ±0.33	**1.60** ±0.16	2.00 ±0.33	**1.00** ±0.00	3.10 ±1.02	2.70 ±0.47	3.12 ±0.48
	−explore	1.50 ±0.31	**1.00** ±0.00	2.80 ±0.20	2.00 ±0.00	**1.00** ±0.00	**1.00** ±0.00	**2.40** ±0.75	3.10 ±0.43	3.44 ±0.48

Table 6: Prompt heterogeneity and final-regret differences for the eight settings in Table 5. We compute $\hat{\rho}^2 = \max_{i \neq j} \mathbb{E}_n[(\hat{\varepsilon}_{i,n} - \hat{\varepsilon}_{j,n})^2]$, where $\hat{\varepsilon}_{i,n}$ is model i's Borda score on prompt n minus its average over prompts. As a reference, we shuffle prompt labels independently within each model pair. This preserves each pair's total wins, losses, and ties but breaks their alignment across prompts. The reported heterogeneity is $\hat{\rho}^2$ minus the average of this measure over 20 such shuffles. ΔFR is the mean final regret of TOPS minus that of its $-$dissim ablation, over 10 seeds at $T = 500$. Negative values indicate that dissimilarity helps.

Setting	Heterogeneity	ΔFR
MedCase	0.0857	+0.0168
AE/HB	0.1097	0.0000
AE/koala	0.1124	-0.0165
AE/oasst	0.1093	-0.0167
AE/SI	0.1078	0.0000
AE/vicuna	0.1280	0.0000
GSM8K	0.0877	+0.0046
MMLU-Pro	0.1118	-0.0726

Table 6 relates measured prompt heterogeneity to the final-regret difference between TOPS and its $-$dissim ablation. Across the eight settings, Spearman's rank correlation is $\rho_s = -0.610$ ($p = 0.121$, two-sided permutation test), indicating that the dissimilarity tensor helps when heterogeneity is higher.

F.2 POINTWISE ACTIVE-EVALUATION BASELINES

Table 7: Pointwise active-evaluation baselines (UCB-E and UCB-E-LRF, Zhou et al. (2025)) against the pairwise methods, $T = 500$, 10 seeds, on the five AlpacaEval sources with the GPT-5.2 annotator. Each pointwise query observes a 1–10 judge score for one (model, prompt) cell; ground truth is the pairwise Bradley-Terry oracle of Section 3 for every method, so all numbers are comparable. Subset sizes are listed in Table 4. **FR**: final regret; **CR**: cumulative regret; **Rank**: true rank of the chosen model (1 = best). Lower is better. Cells: mean ± 1 SEM. `Rank Avg.`: mean per-setting rank against the 9 competing methods, averaged across the five sources (± 1 SEM across settings). Top-2 per column (within each metric) in bold. Ranks and bolding use the displayed values; tied values receive their average rank, and every cell tied at the second-smallest displayed value is bold.

Metric	Algorithm	`AE/HB` (api)	`AE/koala` (api)	`AE/oasst` (api)	`AE/SI` (api)	`AE/vicuna` (api)	`Rank Avg.`
FR (↓)	Random	0.120 ±0.068	**0.082 ±0.027**	0.133 ±0.048	0.002 ±0.001	0.291 ±0.098	3.90 ±0.81
	RUCB	0.224 ±0.113	0.114 ±0.025	**0.116 ±0.065**	**0.001 ±0.001**	0.305 ±0.103	**3.70 ±0.86**
	DuelingEXP3	1.229 ±0.252	1.484 ±0.372	0.617 ±0.270	0.779 ±0.321	1.444 ±0.417	9.80 ±0.20
	Borda	0.276 ±0.088	0.173 ±0.050	0.202 ±0.087	**0.001 ±0.000**	0.319 ±0.108	6.10 ±0.81
	BALD	0.190 ±0.097	0.371 ±0.183	**0.117 ±0.026**	0.120 ±0.079	0.166 ±0.085	5.20 ±1.39
	FewerBattles	**0.000 ±0.000**	0.577 ±0.000	2.530 ±0.066	0.063 ±0.059	0.278 ±0.113	6.50 ±1.61
	GURO	0.259 ±0.120	**0.081 ±0.027**	0.251 ±0.088	0.002 ±0.001	**0.125 ±0.084**	4.50 ±1.24
	TOPS (Ours)	**0.000 ±0.000**	0.148 ±0.016	0.151 ±0.017	**0.000 ±0.000**	**0.000 ±0.000**	**2.50 ±0.84**
	UCB-E	0.285 ±0.000	0.162 ±0.000	0.347 ±0.076	**0.001 ±0.000**	0.694 ±0.000	6.90 ±1.03
	UCB-E-LRF	0.256 ±0.028	0.130 ±0.022	0.447 ±0.050	**0.001 ±0.000**	0.680 ±0.014	5.90 ±0.95
CR (↓)	Random	194.7 ±19.1	132.5 ±23.9	**155.2 ±22.7**	97.7 ±21.6	266.0 ±55.4	3.80 ±0.66
	RUCB	277.3 ±37.7	190.0 ±33.1	164.8 ±39.5	103.5 ±16.3	207.0 ±41.6	5.20 ±0.86
	DuelingEXP3	697.1 ±121.7	781.8 ±167.6	364.9 ±126.8	639.8 ±148.1	944.6 ±199.7	9.80 ±0.20
	Borda	187.9 ±45.8	139.9 ±26.1	172.7 ±38.6	**37.7 ±5.6**	224.3 ±53.9	**3.60 ±0.51**
	BALD	302.9 ±49.6	203.5 ±61.4	202.2 ±60.1	253.4 ±95.4	316.9 ±87.9	7.20 ±0.66
	FewerBattles	**112.7 ±16.6**	463.9 ±27.9	1213.3 ±17.5	305.1 ±8.6	191.5 ±16.8	6.60 ±1.69
	GURO	197.5 ±30.7	**122.5 ±15.0**	221.4 ±39.1	73.7 ±20.8	**185.7 ±38.3**	**3.60 ±0.81**
	TOPS (Ours)	**69.6 ±13.1**	**111.3 ±19.5**	**74.9 ±11.1**	**24.5 ±4.9**	**65.3 ±20.7**	**1.00 ±0.00**
	UCB-E	211.3 ±11.8	165.1 ±16.8	231.5 ±29.2	104.5 ±30.7	415.4 ±31.4	6.40 ±0.51
	UCB-E-LRF	221.0 ±17.5	233.5 ±42.5	270.1 ±24.0	131.0 ±12.6	483.5 ±50.0	7.80 ±0.37
Rank (↓)	Random	1.40 ±0.22	**1.70 ±0.26**	**1.70 ±0.21**	2.30 ±0.21	1.60 ±0.22	**3.70 ±0.89**
	RUCB	1.70 ±0.33	2.00 ±0.26	**1.50 ±0.27**	1.80 ±0.25	1.70 ±0.26	**3.70 ±0.77**
	DuelingEXP3	4.00 ±0.42	7.50 ±1.69	3.10 ±0.81	4.00 ±1.18	4.60 ±1.15	9.80 ±0.20
	Borda	1.90 ±0.28	2.70 ±0.30	1.90 ±0.31	**1.60 ±0.22**	1.80 ±0.29	5.30 ±0.96
	BALD	1.60 ±0.31	2.90 ±0.50	**1.70 ±0.15**	2.40 ±0.40	**1.30 ±0.15**	5.00 ±1.25
	FewerBattles	**1.00 ±0.00**	4.00 ±0.00	11.00 ±0.67	3.10 ±0.10	1.80 ±0.33	7.20 ±1.54
	GURO	1.80 ±0.36	**1.60 ±0.22**	2.10 ±0.31	2.10 ±0.28	**1.30 ±0.21**	4.30 ±1.07
	TOPS (Ours)	**1.00 ±0.00**	2.80 ±0.20	1.90 ±0.10	**1.00 ±0.00**	**1.00 ±0.00**	**3.00 ±1.19**
	UCB-E	2.00 ±0.00	2.00 ±0.00	2.40 ±0.31	2.00 ±0.00	3.00 ±0.00	6.80 ±1.01
	UCB-E-LRF	1.90 ±0.10	1.80 ±0.13	2.80 ±0.20	2.00 ±0.00	2.90 ±0.10	6.20 ±1.03

Table 7 compares the pairwise methods with UCB-E and UCB-E-LRF on the five AlpacaEval sources with the GPT-5.2 annotator. The reason neither pointwise method wins is a target mismatch rather than an inefficiency of the algorithms: given the full score matrix, the model with the highest mean pointwise score need not be the model preferred under pairwise comparisons (Zheng et al., 2023; Liusie et al., 2024).

F.3 Cumulative Regret Versus Time for All Settings

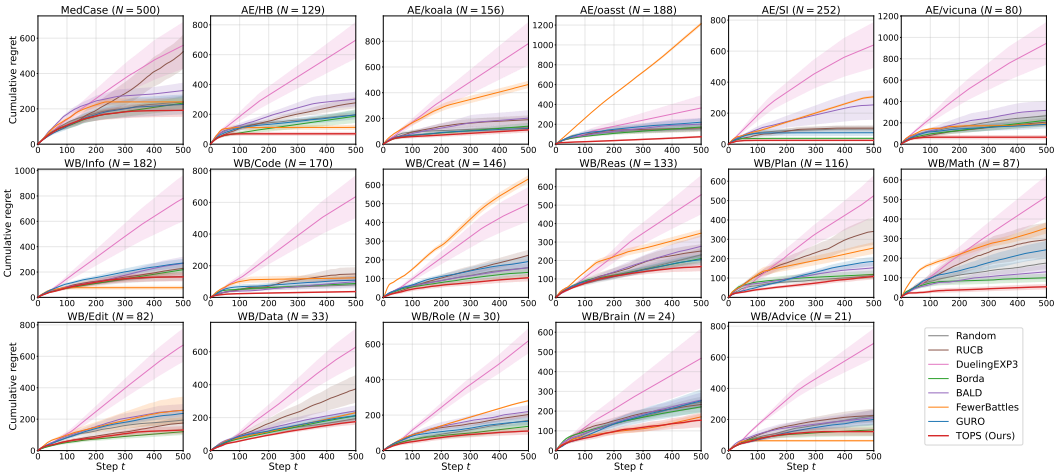


Figure 4: Cumulative regret as a function of t on all 17 settings of Table 15. Bands are one standard error around the mean over 10 seeds.

F.4 Varying Pool Size

Table 8: `Rank Avg.` as a function of pool size N, computed over the seven verifiable (judge, dataset) settings of Table 16 whose pool-size ladder reaches $N = 1000$: `local` $\times$ {`GSM8K`, `MedCase`, `MMLU`}, `api` $\times$ `MedCase`, and `length_biased` $\times$ {`GSM8K`, `MedCase`, `MMLU`}; $T = 500$, 10 seeds. Each cell is a method's mean per-setting rank against the 7 competing methods, averaged across the 7 settings ($\pm$ 1 SEM across settings). **FR**: final regret; **CR**: cumulative regret; **Rank**: true rank of the chosen model (1 = best). Lower is better. Top-2 per column (within each metric) in bold. Ranks and bolding use the displayed values; tied values receive their average rank, and every cell tied at the second-smallest displayed value is bold.

Metric	Algorithm	N=100	N=250	N=500	N=1000
FR (↓)	Random	4.14 ±0.70	4.50 ±0.29	3.36 ±0.28	4.43 ±0.48
	RUCB	6.43 ±0.87	6.21 ±0.86	6.07 ±0.92	6.00 ±0.96
	DuelingEXP3	7.43 ±0.20	7.57 ±0.20	7.29 ±0.29	7.43 ±0.20
	Borda	3.50 ±0.42	4.29 ±0.52	5.00 ±0.44	3.71 ±0.58
	BALD	5.29 ±0.64	5.00 ±0.90	5.14 ±0.86	5.14 ±0.77
	FewerBattles	**2.57 ±0.64**	**2.21 ±0.53**	**2.71 ±0.99**	**2.29 ±0.74**
	GURO	4.57 ±0.52	4.43 ±0.65	4.43 ±0.43	4.71 ±0.57
	TOPS (Ours)	**2.07 ±0.67**	**1.79 ±0.15**	**2.00 ±0.52**	**2.29 ±0.51**
CR (↓)	Random	5.43 ±0.43	4.71 ±0.29	4.14 ±0.34	3.86 ±0.40
	RUCB	7.29 ±0.36	6.86 ±0.34	6.71 ±0.52	6.43 ±0.72
	DuelingEXP3	7.43 ±0.20	7.71 ±0.18	7.57 ±0.20	7.57 ±0.20
	Borda	2.71 ±0.18	3.14 ±0.55	**3.14 ±0.26**	**3.57 ±0.78**
	BALD	4.71 ±0.47	5.14 ±0.80	5.00 ±0.82	4.86 ±0.46
	FewerBattles	**2.43 ±0.65**	**2.57 ±0.30**	4.00 ±0.87	4.57 ±0.95
	GURO	4.71 ±0.29	4.57 ±0.53	4.29 ±0.61	3.71 ±0.61
	TOPS (Ours)	**1.29 ±0.18**	**1.29 ±0.29**	**1.14 ±0.14**	**1.43 ±0.30**
Rank (↓)	Random	4.50 ±0.56	4.43 ±0.46	3.21 ±0.26	4.43 ±0.49
	RUCB	6.43 ±0.87	5.93 ±0.92	6.07 ±0.92	6.14 ±0.88
	DuelingEXP3	7.43 ±0.20	7.57 ±0.20	7.29 ±0.29	7.29 ±0.29
	Borda	**3.14 ±0.39**	4.43 ±0.56	4.86 ±0.24	4.00 ±0.61
	BALD	4.79 ±0.80	5.00 ±0.78	5.14 ±0.86	5.00 ±0.76
	FewerBattles	**3.14 ±0.93**	**2.50 ±0.70**	**2.57 ±0.92**	**2.43 ±0.86**
	GURO	4.29 ±0.41	4.43 ±0.58	4.71 ±0.51	4.21 ±0.53
	TOPS (Ours)	**2.29 ±0.67**	**1.71 ±0.15**	**2.14 ±0.66**	**2.50 ±0.77**

F.5 VARYING THE NUMBER OF CANDIDATE MODELS

Table 9: `Rank Avg.` as a function of the number of candidate models K, on the 16 open-ended sub-pools of Table 15 (5 AlpacaEval sources with the GPT-5.2 annotator, 11 WildBench categories with the Qwen-2.5-72B annotator) at $T = 500$ with 10 seeds. For $K < 20$, each setting keeps the top-K finalists under its own full-pool Bradley-Terry ranking and the ground truth is recomputed on the restricted problem; $K = 20$ is the Table 15 setting. Cells: mean per-setting rank against the 7 competing methods, averaged across the 16 settings ($\pm$ 1 SEM across settings). **FR**: final regret; **CR**: cumulative regret; **Rank**: true rank of the chosen model (1 = best). Lower is better. Top-2 per column (within each metric) in bold. Ranks and bolding use the displayed values; tied values receive their average rank.

Metric	Algorithm	K=5	K=10	K=20
FR (↓)	Random	**3.66** ±**0.44**	4.22 ±0.34	**3.84** ±**0.31**
	RUCB	5.28 ±0.54	4.81 ±0.56	4.84 ±0.43
	DuelingEXP3	7.81 ±0.10	7.75 ±0.17	7.88 ±0.09
	Borda	3.75 ±0.36	4.22 ±0.50	4.09 ±0.50
	BALD	4.22 ±0.46	4.97 ±0.39	4.38 ±0.40
	FewerBattles	**3.06** ±**0.52**	**3.28** ±**0.66**	5.12 ±0.63
	GURO	4.25 ±0.40	3.53 ±0.48	3.97 ±0.53
	TOPS (Ours)	3.97 ±0.54	**3.22** ±**0.44**	**1.88** ±**0.32**
CR (↓)	Random	**3.38** ±**0.44**	3.75 ±0.39	3.62 ±0.33
	RUCB	4.97 ±0.57	5.12 ±0.54	5.31 ±0.34
	DuelingEXP3	7.31 ±0.28	7.62 ±0.20	7.88 ±0.09
	Borda	**3.25** ±**0.47**	**3.56** ±**0.46**	**2.94** ±**0.25**
	BALD	4.69 ±0.49	4.75 ±0.45	5.56 ±0.32
	FewerBattles	4.38 ±0.75	4.62 ±0.76	5.19 ±0.63
	GURO	4.31 ±0.44	3.81 ±0.39	4.31 ±0.38
	TOPS (Ours)	3.72 ±0.46	**2.75** ±**0.39**	**1.19** ±**0.10**
Rank (↓)	Random	**3.69** ±**0.43**	4.56 ±0.36	4.19 ±0.36
	RUCB	5.06 ±0.55	4.69 ±0.56	4.91 ±0.39
	DuelingEXP3	7.75 ±0.14	7.53 ±0.29	7.88 ±0.09
	Borda	3.84 ±0.37	4.19 ±0.49	4.06 ±0.49
	BALD	4.19 ±0.48	4.56 ±0.37	4.12 ±0.37
	FewerBattles	**3.31** ±**0.61**	3.59 ±0.72	4.81 ±0.68
	GURO	4.06 ±0.31	**3.53** ±**0.48**	**4.00** ±**0.49**
	TOPS (Ours)	4.09 ±0.53	**3.34** ±**0.48**	**2.03** ±**0.41**

Table 9 varies the number of candidate models K on the 16 open-ended settings of Table 15. For $K < 20$ each setting keeps only the top-K finalists under its own full-pool Bradley-Terry ranking, which mirrors the deployment scenario in which practitioners narrow down to a few finalists before running preference evaluation, and is also the hardest regime, since the finalists are close in quality by construction. With $K = 5$, the active challenger set has little to prune, so TOPS does not perform significantly better than baselines. The advantage of TOPS appears at $K = 10$, where it has the best `Rank Avg.` on all three metrics, and is largest at $K = 20$.

F.6 Statistical Significance and Identification Curves

Table 10: Paired comparison of TOPS against each baseline over the 17 settings of Table 15 ($T = 500$, 10 seeds). *Settings*: number of settings (of 17) on which TOPS has the better / worse mean (ties omitted). p: two-sided exact seed-level sign-flip test on the mean standardized per-seed difference across settings (settings share seeds, so the 10 seeds are the exchangeable units; the smallest attainable p is $2^{-9} \approx 0.002$; the test is exact under the null that the two methods' labels are exchangeable within each seed). Ties (differences below 10^{-9}, arising from separately fitted oracles) are not counted. Bold: significant at 0.05 after a Holm correction over the 7 baselines within each metric.

	FR		CR		Rank	
TOPS vs	settings	p	settings	p	settings	p
Random	14/3	**0.012**	17/0	**0.0020**	14/3	**0.016**
RUCB	15/2	**0.0020**	17/0	**0.0020**	14/3	**0.0020**
DuelingEXP3	17/0	**0.0020**	17/0	**0.0020**	17/0	**0.0020**
Borda	15/2	**0.0020**	16/1	**0.0020**	13/3	**0.0020**
BALD	15/2	**0.0020**	17/0	**0.0020**	15/2	**0.0020**
FewerBattles	13/2	**0.0020**	15/2	**0.0020**	12/2	**0.0020**
GURO	14/3	**0.0020**	17/0	**0.0020**	14/2	**0.0039**

Table 10 tests the differences in Table 15 with paired per-seed statistics. Because the 17 settings share the same 10 seeds, we use the seed as the exchangeable unit: for each seed we average the standardized per-setting difference between TOPS and a baseline across settings, and test the resulting 10 values with an exact sign-flip test over all 2^{10} sign patterns. TOPS is significantly better than every baseline on every metric after a Holm correction (all $p \leq 0.016$), and has the better mean on at least 12 of the 17 settings in every comparison.

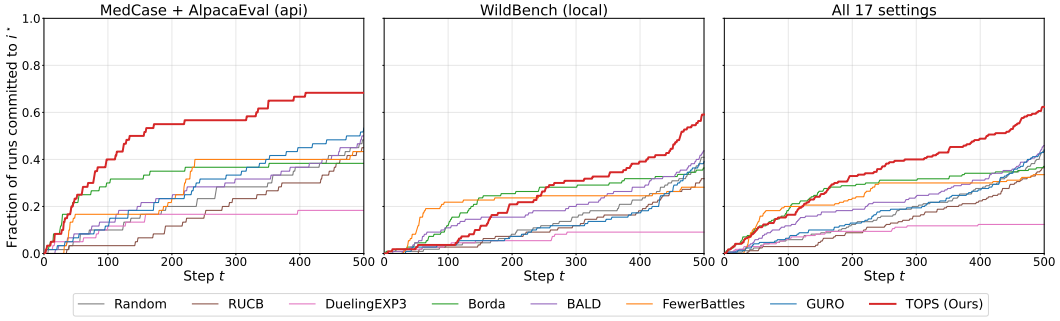


Figure 5: Fraction of runs that have committed to $i^\star$ as a function of t, averaged over the settings of each group (a run has committed at t if its predicted top-1 equals $i^\star$ at t and at every later step; $T = 500$, 10 seeds per setting).

Table 11: Percentage of runs committed to the fitted full-pool winner by t over the 17 settings of Table 15, with 10 seeds per setting and final budget $T = 500$. A run has committed by t if its recommendation is correct at t and remains correct at every later step through T. Thus, the $t = 500$ column is the final identification rate. Cells show the mean percentage across settings [95% percentile bootstrap CI with 20,000 samples]. Intervals describe variation across runs on these fixed settings, not uncertainty over new tasks. Higher is better. The highest mean in each column is bold.

Algorithm	$t = 100$	$t = 250$	$t = 500$
Random	5.9 [2.4, 9.4]	15.9 [10.6, 21.2]	43.5 [32.9, 55.3]
RUCB	2.9 [0.6, 5.3]	11.2 [7.6, 15.3]	36.5 [30.0, 42.9]
DuelingEXP3	6.5 [2.4, 10.6]	9.4 [4.7, 14.7]	12.4 [7.6, 17.1]
Borda	18.2 [11.2, 25.3]	31.2 [23.5, 38.8]	37.1 [28.8, 44.7]
BALD	11.8 [7.6, 17.1]	21.8 [16.5, 27.6]	45.9 [40.0, 52.4]
FewerBattles	**20.0** [17.6, 22.4]	30.0 [28.2, 31.8]	33.5 [31.2, 35.9]
GURO	6.5 [2.9, 10.0]	18.2 [14.7, 21.8]	44.1 [39.4, 49.4]
TOPS (Ours)	16.5 [13.5, 20.0]	**37.1** [31.8, 42.9]	**62.4** [58.2, 65.9]

Figure 5 and Table 11 report the sample complexity of identification. We say that a run has committed at step t if its predicted top-1 equals $i^\star$ at t and at every later step; we use this rather than the first correct step because runs can be correct transiently and later switch away for good. By $t = 100$, Borda and FewerBattles have committed on slightly more runs than TOPS (0.18 and 0.20 against 0.16), reflecting their greedier early behaviour; from $t \approx 200$ on, TOPS leads, and at $T = 500$ it has committed on 62% of runs against at most 46% for any baseline.

F.7 RUNTIME

Table 12: Total per-step wall-clock cost on `AlpacaEval` with the GPT-5.2 annotator ($K = 20$, $N = 500$, $T = 500$, 10 seeds, 4 CPU cores). Values are mean $\pm$ 1 SEM across seeds, in milliseconds, and include acquisition, update, and ranking. One-off setup is excluded.

Algorithm	Total (ms)
Random	3.354 $\pm$0.071
RUCB	0.198 $\pm$0.005
DuelingEXP3	0.099 $\pm$0.002
Borda	0.293 $\pm$0.007
BALD	5.954 $\pm$0.092
FewerBattles	7.163 $\pm$0.144
GURO	10.452 $\pm$0.187
TOPS (Ours)	7.229 $\pm$0.117

F.8 Partial Response Caches

Table 13: TOPS with a dissimilarity tensor estimated from a fraction of the response cells. For each prompt, responses of only c of the $K = 20$ models are observed (cell budget c/K, independent random mask per seed); the unobserved entries of D are imputed with an additive model $D_{ijn} \approx \mu + a_{ij} + b_n$ fit on the observed entries, and TOPS is run with the imputed tensor. *full* D uses the complete cache. $T = 500$, 10 seeds; cells: mean $\pm$ 1 SEM. We use the GPT-5.2 annotator on the five AlpacaEval sources at their natural pool sizes (Table 4), with $c \in \{2, 4, 10\}$ only.

Metric	D	AE/HB	AE/koala	AE/oasst	AE/SI	AE/vicuna
FR (↓)	full D	0.000 ±0.000	0.148 ±0.016	0.151 ±0.017	0.000 ±0.000	0.000 ±0.000
	c=2 (10%)	0.000 ±0.000	0.165 ±0.000	0.167 ±0.000	0.000 ±0.000	0.000 ±0.000
	c=4 (20%)	0.000 ±0.000	0.165 ±0.000	0.167 ±0.000	0.000 ±0.000	0.000 ±0.000
	c=10 (50%)	0.063 ±0.063	0.165 ±0.000	0.151 ±0.017	0.000 ±0.000	0.000 ±0.000
CR (↓)	full D	69.6 ±13.1	111.3 ±19.5	74.9 ±11.1	24.5 ±4.9	65.3 ±20.7
	c=2 (10%)	73.5 ±14.9	102.3 ±7.8	105.2 ±5.7	37.0 ±8.3	58.3 ±10.4
	c=4 (20%)	59.7 ±12.5	93.8 ±9.8	108.4 ±11.6	39.8 ±8.6	52.8 ±10.8
	c=10 (50%)	113.0 ±36.4	107.1 ±10.0	97.4 ±17.8	30.8 ±7.1	49.7 ±12.0

TOPS requires the full $K \times N$ response cache only through the dissimilarity tensor D. Table 13 tests whether D can be estimated from a fraction of the cells. For each prompt we observe the responses of only c of the $K = 20$ models (a cell budget of c/K, with an independent random mask per seed), impute the unobserved entries of D with an additive model $D_{ijn} \approx \mu + a_{ij} + b_n$ fit on the observed entries, and run TOPS with the imputed tensor. We find only small degradations in performance for some datasets when D is not fully observed.

F.9 Human Annotation Study

We have conducted a small proof-of-concept experiment with a human annotator (an author of the paper), on the `AlpacaEval` self-instruct subset ($N = 252$), for $K = 8$ models. The primary challenge of a human study for this problem setup is that ground truth for the true "best" model is not available. Definitively establishing this would mean labeling the whole $K \times K \times N$ matrix. Here, we approximate it by first labeling a hold-out set of 200 pairwise comparisons (selected i.i.d. from the matrix), and learning a Bradley-Terry model. The fit places Kimi-K2.5 first with a bootstrap top-1 probability of 0.997. This stage is far from exhaustive, but it identifies a winner with high confidence.

Next, we run three independent active labeling studies on the same dataset, using TOPS, Borda, and GURO. In each trial, we stop when the BT model confirms a winner (i.e., the 95% lower bound excludes all other models), mimicking a real deployment case where one continues to label until they are able to confirm the best model. Table 14 shows the number of labels needed to confirm the winner. All three methods eventually selected Kimi-K2.5.

Table 14: Labels to certification in the proof-of-concept human annotation study on `AlpacaEval` self-instruct ($N = 252$, $K = 8$). Each method stops when the BT model's 95% lower bound excludes all other models. Results are from a single trial with no replicates. Fewer labels are better, and the lowest count is bolded.

Method	Labels to certification
TOPS	**38**
Borda	45
GURO	51

F.10 All Judges and Datasets

Table 15 gives the complete version of Table 1, including final ranks. Table 16 also includes the additional datasets and annotators.

Table 15: Complete results for the 17 main evaluation settings at $T = 500$, mean $\pm$ 1 SEM over 10 seeds. **FR**: final regret; **CR**: cumulative regret; **Rank**: true rank of the selected model. Lower is better. Dataset sizes are listed in Table 4. `Rank Avg.` averages method ranks across the settings in each panel, with SEM across settings; `All 17` combines both panels. The best two values per column are bold, including ties. Ranks use displayed values, with average ranks for ties. Panel B reports all 11 WildBench categories with SEMs below the means. The `WB` column averages method ranks over these categories.

Panel A: GPT-5.2 annotator

Metric	Algorithm	MedCase	AE/HB	AE/koala	AE/oasst	AE/SI	AE/vicuna	Rank Avg.
FR (↓)	Random	0.122 ±0.074	0.120 ±0.068	**0.082 ±0.027**	0.133 ±0.048	0.002 ±0.001	0.291 ±0.098	**3.58 ±0.45**
	RUCB	1.597 ±0.333	0.224 ±0.113	0.114 ±0.025	**0.116 ±0.065**	**0.001 ±0.001**	0.305 ±0.103	4.25 ±1.05
	DuelingEXP3	0.831 ±0.281	1.229 ±0.252	1.484 ±0.372	0.617 ±0.270	0.779 ±0.321	1.444 ±0.417	7.67 ±0.21
	Borda	0.241 ±0.083	0.276 ±0.088	0.173 ±0.050	0.202 ±0.087	**0.001 ±0.000**	0.319 ±0.108	5.42 ±0.69
	BALD	0.192 ±0.139	0.190 ±0.097	0.371 ±0.183	**0.117 ±0.026**	0.120 ±0.079	0.166 ±0.085	4.50 ±0.76
	FewerBattles	**0.000 ±0.000**	**0.000 ±0.000**	0.577 ±0.000	2.530 ±0.066	0.063 ±0.059	0.278 ±0.113	4.58 ±1.19
	GURO	0.069 ±0.039	0.259 ±0.120	**0.081 ±0.027**	0.251 ±0.088	0.002 ±0.001	**0.125 ±0.084**	3.75 ±0.85
	TOPS (Ours)	**0.017 ±0.017**	**0.000 ±0.000**	0.148 ±0.016	0.151 ±0.017	**0.000 ±0.000**	**0.000 ±0.000**	**2.25 ±0.57**
CR (↓)	Random	238.7 ±36.4	194.7 ±19.1	132.5 ±23.9	**155.2 ±22.7**	97.7 ±21.6	266.0 ±55.4	3.83 ±0.54
	RUCB	523.5 ±96.7	277.3 ±37.7	190.0 ±33.1	164.8 ±39.5	103.5 ±16.3	207.0 ±41.6	5.00 ±0.58
	DuelingEXP3	559.1 ±132.3	697.1 ±121.7	781.8 ±167.6	364.9 ±126.8	639.8 ±148.1	944.6 ±199.7	7.83 ±0.17
	Borda	232.5 ±38.8	187.9 ±45.8	139.9 ±26.1	172.7 ±38.6	**37.7 ±5.6**	224.3 ±53.9	3.50 ±0.43
	BALD	303.5 ±44.5	302.9 ±49.6	203.5 ±61.4	202.2 ±60.1	253.4 ±95.4	316.9 ±87.9	6.17 ±0.31
	FewerBattles	239.2 ±19.6	**112.7 ±16.6**	463.9 ±27.9	1213.3 ±17.5	305.1 ±8.6	191.5 ±16.8	5.33 ±0.99
	GURO	**225.1 ±56.0**	197.5 ±30.7	**122.5 ±15.0**	221.4 ±39.1	73.7 ±20.8	**185.7 ±38.3**	**3.33 ±0.71**
	TOPS (Ours)	**191.7 ±36.6**	**69.6 ±13.1**	**111.3 ±19.5**	**74.9 ±11.1**	**24.5 ±4.9**	**65.3 ±20.7**	**1.00 ±0.00**
Rank (↓)	Random	1.60 ±0.34	1.40 ±0.22	**1.70 ±0.26**	**1.70 ±0.21**	2.30 ±0.21	1.60 ±0.22	**3.42 ±0.45**
	RUCB	9.70 ±2.02	1.70 ±0.33	2.00 ±0.26	**1.50 ±0.27**	1.80 ±0.25	1.70 ±0.26	4.17 ±0.98
	DuelingEXP3	5.60 ±1.62	4.00 ±0.42	7.50 ±1.69	3.10 ±0.81	4.00 ±1.18	4.60 ±1.15	7.67 ±0.21
	Borda	2.20 ±0.36	1.90 ±0.28	2.70 ±0.30	1.90 ±0.31	**1.60 ±0.22**	1.80 ±0.29	5.00 ±0.76
	BALD	1.90 ±0.59	1.60 ±0.31	2.90 ±0.50	**1.70 ±0.15**	2.40 ±0.40	**1.30 ±0.15**	4.33 ±0.65
	FewerBattles	**1.00 ±0.00**	**1.00 ±0.00**	4.00 ±0.00	11.00 ±0.67	3.10 ±0.10	1.80 ±0.33	5.17 ±1.26
	GURO	1.40 ±0.22	1.80 ±0.36	**1.60 ±0.22**	2.10 ±0.31	2.10 ±0.28	**1.30 ±0.21**	3.75 ±0.81
	TOPS (Ours)	**1.10 ±0.10**	**1.00 ±0.00**	2.80 ±0.20	1.90 ±0.10	**1.00 ±0.00**	**1.00 ±0.00**	**2.50 ±0.73**

Panel B: Qwen-2.5-72B annotator, all WildBench categories

Metric	Algorithm	WB/ Info	WB/ Code	WB/ Creat	WB/ Reas	WB/ Plan	WB/ Math	WB/ Edit	WB/ Data	WB/ Role	WB/ Brain	WB/ Advice	Rank Avg. WB	Rank Avg. All 17
FR (↓)	Random	0.101 ±0.051	0.056 ±0.013	0.227 ±0.052	0.397 ±0.103	0.131 ±0.048	0.177 ±0.092	0.196 ±0.074	**0.225 ±0.078**	0.183 ±0.068	0.247 ±0.102	0.233 ±0.106	4.00 ±0.38	**3.85 ±0.29**
	RUCB	0.313 ±0.101	0.066 ±0.012	0.493 ±0.073	0.215 ±0.072	0.193 ±0.082	0.275 ±0.129	0.327 ±0.154	0.615 ±0.204	0.268 ±0.049	0.259 ±0.093	0.096 ±0.096	5.45 ±0.37	5.03 ±0.44
	DuelingEXP3	1.314 ±0.401	1.139 ±0.332	0.735 ±0.162	1.043 ±0.281	1.106 ±0.384	0.979 ±0.259	1.155 ±0.257	1.120 ±0.275	1.289 ±0.172	0.845 ±0.332	1.138 ±0.262	7.91 ±0.09	7.82 ±0.10
	Borda	0.362 ±0.049	0.084 ±0.018	**0.104 ±0.056**	0.303 ±0.066	**0.070 ±0.028**	**0.048 ±0.014**	0.152 ±0.037	0.307 ±0.063	0.188 ±0.070	0.246 ±0.088	0.085 ±0.085	**3.55 ±0.58**	4.21 ±0.49
	BALD	0.320 ±0.111	0.145 ±0.115	0.193 ±0.086	0.395 ±0.104	0.162 ±0.057	0.123 ±0.097	0.155 ±0.058	0.273 ±0.047	0.189 ±0.070	**0.221 ±0.097**	0.137 ±0.070	4.36 ±0.43	4.41 ±0.37
	FewerBattles	**0.000 ±0.000**	0.053 ±0.010	0.901 ±0.000	0.421 ±0.000	0.327 ±0.048	0.487 ±0.068	**0.143 ±0.143**	0.393 ±0.000	0.390 ±0.065	0.334 ±0.000	**0.000 ±0.000**	5.05 ±0.79	4.88 ±0.64
	GURO	0.274 ±0.089	**0.046 ±0.015**	0.298 ±0.107	**0.136 ±0.069**	**0.116 ±0.052**	0.253 ±0.124	0.198 ±0.082	0.271 ±0.068	**0.106 ±0.035**	0.363 ±0.148	0.298 ±0.107	4.00 ±0.65	3.91 ±0.50
	TOPS (Ours)	**0.035 ±0.035**	**0.034 ±0.003**	**0.142 ±0.033**	**0.177 ±0.073**	0.191 ±0.065	**0.034 ±0.014**	**0.100 ±0.052**	**0.156 ±0.065**	**0.064 ±0.032**	**0.114 ±0.114**	**0.000 ±0.000**	**1.68 ±0.36**	**1.88 ±0.30**

Continued on next page

Table 15 (continued)

Panel B: Qwen-2.5-72B annotator, all WildBench categories (continued)

Metric	Algorithm	WB/ Info	WB/ Code	WB/ Creat	WB/ Reas	WB/ Plan	WB/ Math	WB/ Edit	WB/ Data	WB/ Role	WB/ Brain	WB/ Advice	Rank Avg. WB	Rank Avg. All 17
CR (↓)	Random	166.2 ±33.3	**79.1 ±13.5**	158.4 ±23.0	230.4 ±36.4	**112.8 ±13.0**	175.0 ±31.8	194.0 ±21.3	**193.5 ±22.4**	170.1 ±29.8	245.5 ±27.6	222.0 ±40.5	3.55 ±0.39	3.65 ±0.31
	RUCB	229.0 ±49.9	148.1 ±42.3	225.2 ±29.0	251.3 ±36.8	340.8 ±70.2	295.1 ±87.7	175.5 ±36.9	374.6 ±83.6	203.5 ±14.6	234.0 ±19.3	225.5 ±43.3	5.64 ±0.41	5.41 ±0.33
	DuelingEXP3	781.5 ±182.3	636.4 ±135.2	497.6 ±91.1	556.2 ±106.1	525.0 ±106.1	514.7 ±108.9	670.8 ±106.7	627.9 ±113.4	618.7 ±69.9	467.8 ±151.4	689.4 ±96.1	7.91 ±0.09	7.88 ±0.08
	Borda	218.4 ±19.9	87.5 ±9.2	**133.4 ±24.2**	210.5 ±23.7	118.4 ±20.5	**99.9 ±20.7**	**120.3 ±21.2**	211.2 ±24.6	**136.8 ±26.7**	221.7 ±30.9	133.0 ±41.0	**2.64 ±0.24**	**2.94 ±0.23**
	BALD	268.8 ±52.0	99.3 ±41.7	158.8 ±39.1	276.9 ±60.9	153.0 ±27.3	130.5 ±37.7	256.5 ±39.2	240.9 ±33.1	220.2 ±35.6	255.6 ±37.4	210.9 ±50.2	5.27 ±0.41	5.59 ±0.30
	FewerBattles	**76.3 ±15.4**	123.7 ±19.8	631.8 ±20.0	348.2 ±20.5	255.2 ±30.4	354.3 ±28.3	256.2 ±86.9	230.0 ±4.7	281.3 ±5.9	**171.0 ±15.7**	**63.2 ±6.4**	5.09 ±0.77	5.18 ±0.59
	GURO	270.8 ±33.1	108.5 ±26.3	191.3 ±33.4	**204.4 ±28.3**	185.7 ±23.6	243.0 ±45.4	236.5 ±24.4	212.0 ±31.7	165.5 ±21.0	252.1 ±57.5	200.4 ±34.4	4.64 ±0.41	4.18 ±0.39
	TOPS (Ours)	**161.2 ±34.3**	**36.6 ±4.8**	**104.8 ±19.0**	**167.2 ±24.6**	**108.0 ±15.2**	**54.0 ±13.8**	**131.2 ±24.8**	**175.6 ±19.3**	**110.1 ±20.5**	**153.8 ±36.6**	**122.1 ±26.7**	**1.27 ±0.14**	**1.18 ±0.10**
Rank (↓)	Random	1.30 ±0.15	3.70 ±0.62	2.90 ±0.48	3.80 ±0.85	3.30 ±0.50	3.10 ±0.64	2.30 ±0.52	3.70 ±0.94	2.10 ±0.46	2.50 ±0.65	2.00 ±0.52	4.59 ±0.42	4.18 ±0.34
	RUCB	2.40 ±0.65	4.20 ±0.57	5.10 ±0.60	2.10 ±0.41	3.80 ±0.83	4.00 ±1.13	3.20 ±1.02	6.20 ±1.63	2.60 ±0.43	2.40 ±0.60	1.50 ±0.50	5.59 ±0.29	5.09 ±0.41
	DuelingEXP3	7.50 ±1.78	10.20 ±1.82	6.90 ±1.24	7.80 ±1.70	9.10 ±2.25	8.40 ±1.63	8.80 ±1.82	9.50 ±1.82	11.40 ±1.61	5.60 ±1.75	7.30 ±1.51	7.91 ±0.09	7.82 ±0.10
	Borda	2.70 ±0.33	4.70 ±0.62	**1.80 ±0.49**	2.80 ±0.44	**2.70 ±0.37**	**2.50 ±0.43**	2.00 ±0.26	4.00 ±0.77	2.20 ±0.49	2.20 ±0.51	1.40 ±0.40	**3.73 ±0.59**	4.18 ±0.48
	BALD	2.90 ±0.69	3.20 ±0.96	2.50 ±0.69	3.30 ±0.87	3.40 ±0.73	2.70 ±0.76	2.00 ±0.37	**3.50 ±0.60**	2.40 ±0.58	2.20 ±0.63	1.50 ±0.27	4.09 ±0.44	4.18 ±0.35
	FewerBattles	**1.00 ±0.00**	3.20 ±0.49	8.00 ±0.00	**2.00 ±0.00**	5.40 ±0.31	4.70 ±0.21	**1.80 ±0.80**	6.00 ±0.00	4.20 ±0.53	**2.00 ±0.00**	**1.00 ±0.00**	4.27 ±0.82	4.59 ±0.68
	GURO	2.30 ±0.45	**3.00 ±0.56**	3.50 ±0.90	**2.00 ±0.52**	**3.20 ±0.70**	3.40 ±0.90	2.30 ±0.56	3.60 ±0.83	**1.50 ±0.17**	3.10 ±0.95	2.10 ±0.41	4.05 ±0.60	**3.94 ±0.47**
	TOPS (Ours)	**1.20 ±0.20**	**2.40 ±0.22**	**2.10 ±0.28**	**2.00 ±0.45**	4.00 ±0.89	**2.10 ±0.41**	**1.60 ±0.31**	**2.80 ±0.80**	**1.30 ±0.15**	**1.70 ±0.70**	**1.00 ±0.00**	**1.77 ±0.44**	**2.03 ±0.38**

Table 16: Comparison across 23 (judge, dataset or subset) settings at $T = 500$ with 10 seeds. We report each of the five AlpacaEval sources with the GPT-5.2 (`api`) annotator and each of the eleven WildBench categories with the Qwen-2.5-72B (`local`) annotator separately, at their natural pool sizes. Dataset sizes and subset abbreviations are listed in Table 4. **FR**: final regret; **CR**: cumulative regret; **Rank**: true rank of the chosen model (1 = best). Lower is better for all three. Cells show mean $\pm$ 1 SEM. Top-2 per setting are bold within each metric, including all ties at the second-smallest displayed value.

Judge	Dataset / subset	Method	FR (↓)	CR (↓)	Rank (↓)
local	GSM8K	Random	0.822 ±0.273	554.7 ±73.2	5.70 ±1.71
		RUCB	4.280 ±0.610	1508.9 ±180.2	17.70 ±0.91
		DuelingEXP3	1.828 ±0.507	840.9 ±216.5	10.20 ±1.87
		Borda	0.838 ±0.224	458.5 ±73.3	5.90 ±1.52
		BALD	0.959 ±0.259	504.7 ±45.7	6.30 ±1.60
		FewerBattles	**0.372 ±0.079**	**322.9 ±22.4**	**2.40 ±0.40**
		GURO	0.734 ±0.188	436.9 ±63.4	4.90 ±1.18
		TOPS (Ours)	**0.514 ±0.131**	**355.0 ±40.9**	**3.90 ±0.98**
local	MMLU	Random	0.188 ±0.064	179.1 ±29.9	3.70 ±0.82
		RUCB	1.265 ±0.310	584.2 ±64.6	12.10 ±2.17
		DuelingEXP3	0.645 ±0.188	349.9 ±74.2	8.20 ±1.38
		Borda	0.177 ±0.032	160.0 ±16.0	4.30 ±0.47
		BALD	**0.158 ±0.053**	**159.4 ±14.9**	**3.20 ±0.59**
		FewerBattles	0.802 ±0.073	314.1 ±8.9	9.40 ±0.60
		GURO	0.234 ±0.075	208.2 ±33.6	4.30 ±0.84
		TOPS (Ours)	**0.071 ±0.017**	**124.5 ±12.5**	**2.20 ±0.29**
local	MedCase	Random	0.145 ±0.061	156.5 ±24.8	3.00 ±0.63
		RUCB	0.949 ±0.162	384.0 ±42.1	12.90 ±2.04
		DuelingEXP3	0.513 ±0.085	276.4 ±22.6	6.30 ±1.16
		Borda	0.147 ±0.094	149.1 ±28.9	3.20 ±1.25
		BALD	**0.138 ±0.061**	**145.9 ±33.6**	**2.70 ±0.58**
		FewerBattles	0.382 ±0.013	186.2 ±8.9	3.90 ±0.46
		GURO	0.372 ±0.111	218.3 ±36.6	5.30 ±1.32
		TOPS (Ours)	**0.083 ±0.053**	**119.5 ±25.4**	**1.80 ±0.51**
local	WB/Info	Random	0.101 ±0.051	166.2 ±33.3	1.30 ±0.15
		RUCB	0.313 ±0.101	229.0 ±49.9	2.40 ±0.65
		DuelingEXP3	1.314 ±0.401	781.5 ±182.3	7.50 ±1.78
		Borda	0.362 ±0.049	218.4 ±19.9	2.70 ±0.33
		BALD	0.320 ±0.111	268.8 ±52.0	2.90 ±0.69
		FewerBattles	**0.000 ±0.000**	**76.3 ±15.4**	**1.00 ±0.00**
		GURO	0.274 ±0.089	270.8 ±33.1	2.30 ±0.45
		TOPS (Ours)	**0.035 ±0.035**	**161.2 ±34.3**	**1.20 ±0.20**
local	WB/Code	Random	0.056 ±0.013	**79.1 ±13.5**	3.70 ±0.62
		RUCB	0.066 ±0.012	148.1 ±42.3	4.20 ±0.57
		DuelingEXP3	1.139 ±0.332	636.4 ±135.2	10.20 ±1.82
		Borda	0.084 ±0.018	87.5 ±9.2	4.70 ±0.62
		BALD	0.145 ±0.115	99.3 ±41.7	3.20 ±0.96
		FewerBattles	0.053 ±0.010	123.7 ±19.8	3.20 ±0.49
		GURO	**0.046 ±0.015**	108.5 ±26.3	**3.00 ±0.56**
		TOPS (Ours)	**0.034 ±0.003**	**36.6 ±4.8**	**2.40 ±0.22**
local	WB/Creat	Random	0.227 ±0.052	158.4 ±23.0	2.90 ±0.48
		RUCB	0.493 ±0.073	225.2 ±29.0	5.10 ±0.60
		DuelingEXP3	0.735 ±0.162	497.6 ±91.1	6.90 ±1.24
		Borda	**0.104 ±0.056**	**133.4 ±24.2**	**1.80 ±0.49**
		BALD	0.193 ±0.086	158.8 ±39.1	2.50 ±0.69
		FewerBattles	0.901 ±0.000	631.8 ±20.0	8.00 ±0.00
		GURO	0.298 ±0.107	191.3 ±33.4	3.50 ±0.90
		TOPS (Ours)	**0.142 ±0.033**	**104.8 ±19.0**	**2.10 ±0.28**

(continued on next page)

(continued from previous page)

Judge	Dataset / subset	Method	FR (↓)	CR (↓)	Rank (↓)
local	WB/Reas	Random	0.397 ±0.103	230.4 ±36.4	3.80 ±0.85
		RUCB	0.215 ±0.072	251.3 ±36.8	2.10 ±0.41
		DuelingEXP3	1.043 ±0.281	556.2 ±106.1	7.80 ±1.70
		Borda	0.303 ±0.066	210.5 ±23.7	2.80 ±0.44
		BALD	0.395 ±0.104	276.9 ±60.9	3.30 ±0.87
		FewerBattles	0.421 ±0.000	348.2 ±20.5	**2.00 ±0.00**
		GURO	**0.136 ±0.069**	**204.4 ±28.3**	**2.00 ±0.52**
		TOPS (Ours)	**0.177 ±0.073**	**167.2 ±24.6**	**2.00 ±0.45**
local	WB/Plan	Random	0.131 ±0.048	**112.8 ±13.0**	3.30 ±0.50
		RUCB	0.193 ±0.082	340.8 ±70.2	3.80 ±0.83
		DuelingEXP3	1.106 ±0.384	525.0 ±106.1	9.10 ±2.25
		Borda	**0.070 ±0.028**	118.4 ±20.5	**2.70 ±0.37**
		BALD	0.162 ±0.057	153.0 ±27.3	3.40 ±0.73
		FewerBattles	0.327 ±0.048	255.2 ±30.4	5.40 ±0.31
		GURO	**0.116 ±0.052**	185.7 ±23.6	**3.20 ±0.70**
		TOPS (Ours)	0.191 ±0.065	**108.0 ±15.2**	4.00 ±0.89
local	WB/Math	Random	0.177 ±0.092	175.0 ±31.8	3.10 ±0.64
		RUCB	0.275 ±0.129	295.1 ±87.7	4.00 ±1.13
		DuelingEXP3	0.979 ±0.259	514.7 ±108.9	8.40 ±1.63
		Borda	**0.048 ±0.014**	**99.9 ±20.7**	**2.50 ±0.43**
		BALD	0.123 ±0.097	130.5 ±37.7	2.70 ±0.76
		FewerBattles	0.487 ±0.068	354.3 ±28.3	4.70 ±0.21
		GURO	0.253 ±0.124	243.0 ±45.4	3.40 ±0.90
		TOPS (Ours)	**0.034 ±0.014**	**54.0 ±13.8**	**2.10 ±0.41**
local	WB/Edit	Random	0.196 ±0.074	194.0 ±21.3	2.30 ±0.52
		RUCB	0.327 ±0.154	175.5 ±36.9	3.20 ±1.02
		DuelingEXP3	1.155 ±0.257	670.8 ±106.7	8.80 ±1.82
		Borda	0.152 ±0.037	**120.3 ±21.2**	2.00 ±0.26
		BALD	0.155 ±0.058	256.5 ±39.2	2.00 ±0.37
		FewerBattles	**0.143 ±0.143**	256.2 ±86.9	**1.80 ±0.80**
		GURO	0.198 ±0.082	236.5 ±24.4	2.30 ±0.56
		TOPS (Ours)	**0.100 ±0.052**	**131.2 ±24.8**	**1.60 ±0.31**
local	WB/Data	Random	**0.225 ±0.078**	**193.5 ±22.4**	3.70 ±0.94
		RUCB	0.615 ±0.204	374.6 ±83.6	6.20 ±1.63
		DuelingEXP3	1.120 ±0.275	627.9 ±113.4	9.50 ±1.82
		Borda	0.307 ±0.063	211.2 ±24.6	4.00 ±0.77
		BALD	0.273 ±0.047	240.9 ±33.1	**3.50 ±0.60**
		FewerBattles	0.393 ±0.000	230.0 ±4.7	6.00 ±0.00
		GURO	0.271 ±0.068	212.0 ±31.7	3.60 ±0.83
		TOPS (Ours)	**0.156 ±0.065**	**175.6 ±19.3**	**2.80 ±0.80**
local	WB/Role	Random	0.183 ±0.068	170.1 ±29.8	2.10 ±0.46
		RUCB	0.268 ±0.049	203.5 ±14.6	2.60 ±0.43
		DuelingEXP3	1.289 ±0.172	618.7 ±69.9	11.40 ±1.61
		Borda	0.188 ±0.070	**136.8 ±26.7**	2.20 ±0.49
		BALD	0.189 ±0.070	220.2 ±35.6	2.40 ±0.58
		FewerBattles	0.390 ±0.065	281.3 ±5.9	4.20 ±0.53
		GURO	**0.106 ±0.035**	165.5 ±21.0	**1.50 ±0.17**
		TOPS (Ours)	**0.064 ±0.032**	**110.1 ±20.5**	**1.30 ±0.15**
local	WB/Brain	Random	0.247 ±0.102	245.5 ±27.6	2.50 ±0.65
		RUCB	0.259 ±0.093	234.0 ±19.3	2.40 ±0.60
		DuelingEXP3	0.845 ±0.332	467.8 ±151.4	5.60 ±1.75
		Borda	0.246 ±0.088	221.7 ±30.9	2.20 ±0.51
		BALD	**0.221 ±0.097**	255.6 ±37.4	2.20 ±0.63
		FewerBattles	0.334 ±0.000	**171.0 ±15.7**	**2.00 ±0.00**
		GURO	0.363 ±0.148	252.1 ±57.5	3.10 ±0.95
		TOPS (Ours)	**0.114 ±0.114**	**153.8 ±36.6**	**1.70 ±0.70**

(continued on next page)

(continued from previous page)

Judge	Dataset / subset	Method	FR (↓)	CR (↓)	Rank (↓)
local	WB/Advice	Random	0.233 ±0.106	222.0 ±40.5	2.00 ±0.52
		RUCB	0.096 ±0.096	225.5 ±43.3	1.50 ±0.50
		DuelingEXP3	1.138 ±0.262	689.4 ±96.1	7.30 ±1.51
		Borda	0.085 ±0.085	133.0 ±41.0	1.40 ±0.40
		BALD	0.137 ±0.070	210.9 ±50.2	1.50 ±0.27
		FewerBattles	**0.000 ±0.000**	**63.2 ±6.4**	**1.00 ±0.00**
		GURO	0.298 ±0.107	200.4 ±34.4	2.10 ±0.41
		TOPS (Ours)	**0.000 ±0.000**	**122.1 ±26.7**	**1.00 ±0.00**
api	MedCase	Random	0.122 ±0.074	238.7 ±36.4	1.60 ±0.34
		RUCB	1.597 ±0.333	523.5 ±96.7	9.70 ±2.02
		DuelingEXP3	0.831 ±0.281	559.1 ±132.3	5.60 ±1.62
		Borda	0.241 ±0.083	232.5 ±38.8	2.20 ±0.36
		BALD	0.192 ±0.139	303.5 ±44.5	1.90 ±0.59
		FewerBattles	**0.000 ±0.000**	239.2 ±19.6	**1.00 ±0.00**
		GURO	0.069 ±0.039	**225.1 ±56.0**	1.40 ±0.22
		TOPS (Ours)	**0.017 ±0.017**	**191.7 ±36.6**	**1.10 ±0.10**
api	AE/HB	Random	0.120 ±0.068	194.7 ±19.1	1.40 ±0.22
		RUCB	0.224 ±0.113	277.3 ±37.7	1.70 ±0.33
		DuelingEXP3	1.229 ±0.252	697.1 ±121.7	4.00 ±0.42
		Borda	0.276 ±0.088	187.9 ±45.8	1.90 ±0.28
		BALD	0.190 ±0.097	302.9 ±49.6	1.60 ±0.31
		FewerBattles	**0.000 ±0.000**	**112.7 ±16.6**	**1.00 ±0.00**
		GURO	0.259 ±0.120	197.5 ±30.7	1.80 ±0.36
		TOPS (Ours)	**0.000 ±0.000**	**69.6 ±13.1**	**1.00 ±0.00**
api	AE/koala	Random	**0.082 ±0.027**	132.5 ±23.9	**1.70 ±0.26**
		RUCB	0.114 ±0.025	190.0 ±33.1	2.00 ±0.26
		DuelingEXP3	1.484 ±0.372	781.8 ±167.6	7.50 ±1.69
		Borda	0.173 ±0.050	139.9 ±26.1	2.70 ±0.30
		BALD	0.371 ±0.183	203.5 ±61.4	2.90 ±0.50
		FewerBattles	0.577 ±0.000	463.9 ±27.9	4.00 ±0.00
		GURO	**0.081 ±0.027**	**122.5 ±15.0**	**1.60 ±0.22**
		TOPS (Ours)	0.148 ±0.016	**111.3 ±19.5**	2.80 ±0.20
api	AE/oasst	Random	0.133 ±0.048	**155.2 ±22.7**	**1.70 ±0.21**
		RUCB	**0.116 ±0.065**	164.8 ±39.5	**1.50 ±0.27**
		DuelingEXP3	0.617 ±0.270	364.9 ±126.8	3.10 ±0.81
		Borda	0.202 ±0.087	172.7 ±38.6	1.90 ±0.31
		BALD	**0.117 ±0.026**	202.2 ±60.1	**1.70 ±0.15**
		FewerBattles	2.530 ±0.066	1213.3 ±17.5	11.00 ±0.67
		GURO	0.251 ±0.088	221.4 ±39.1	2.10 ±0.31
		TOPS (Ours)	0.151 ±0.017	**74.9 ±11.1**	1.90 ±0.10
api	AE/SI	Random	0.002 ±0.001	97.7 ±21.6	2.30 ±0.21
		RUCB	**0.001 ±0.001**	103.5 ±16.3	1.80 ±0.25
		DuelingEXP3	0.779 ±0.321	639.8 ±148.1	4.00 ±1.18
		Borda	**0.001 ±0.000**	**37.7 ±5.6**	**1.60 ±0.22**
		BALD	0.120 ±0.079	253.4 ±95.4	2.40 ±0.40
		FewerBattles	0.063 ±0.059	305.1 ±8.6	3.10 ±0.10
		GURO	0.002 ±0.001	73.7 ±20.8	2.10 ±0.28
		TOPS (Ours)	**0.000 ±0.000**	**24.5 ±4.9**	**1.00 ±0.00**
api	AE/vicuna	Random	0.291 ±0.098	266.0 ±55.4	1.60 ±0.22
		RUCB	0.305 ±0.103	207.0 ±41.6	1.70 ±0.26
		DuelingEXP3	1.444 ±0.417	944.6 ±199.7	4.60 ±1.15
		Borda	0.319 ±0.108	224.3 ±53.9	1.80 ±0.29
		BALD	0.166 ±0.085	316.9 ±87.9	**1.30 ±0.15**
		FewerBattles	0.278 ±0.113	191.5 ±16.8	1.80 ±0.33
		GURO	**0.125 ±0.084**	**185.7 ±38.3**	**1.30 ±0.21**
		TOPS (Ours)	**0.000 ±0.000**	**65.3 ±20.7**	**1.00 ±0.00**

(continued on next page)

(continued from previous page)

Judge	Dataset / subset	Method	FR (↓)	CR (↓)	Rank (↓)
length_biased	GSM8K	Random	0.061 ±0.061	246.9 ±40.1	1.10 ±0.10
		RUCB	0.122 ±0.081	261.8 ±67.6	1.20 ±0.13
		DuelingEXP3	1.413 ±0.408	838.4 ±189.0	3.90 ±1.16
		Borda	0.183 ±0.093	146.0 ±46.4	1.30 ±0.15
		BALD	0.525 ±0.309	515.6 ±159.8	1.90 ±0.60
		FewerBattles	**0.000 ±0.000**	**52.7 ±3.1**	**1.00 ±0.00**
		GURO	0.244 ±0.100	296.6 ±34.5	1.40 ±0.16
		TOPS (Ours)	**0.000 ±0.000**	**29.5 ±8.3**	**1.00 ±0.00**
length_biased	MMLU	Random	**0.007 ±0.004**	108.0 ±25.6	**1.30 ±0.15**
		RUCB	0.012 ±0.004	110.0 ±16.8	1.50 ±0.17
		DuelingEXP3	1.277 ±0.351	679.4 ±155.4	5.40 ±1.33
		Borda	0.096 ±0.083	**93.2 ±40.2**	1.70 ±0.21
		BALD	0.216 ±0.197	198.2 ±90.5	2.40 ±0.64
		FewerBattles	**0.000 ±0.000**	286.0 ±30.3	**1.00 ±0.00**
		GURO	0.017 ±0.004	102.0 ±33.8	1.70 ±0.15
		TOPS (Ours)	0.024 ±0.000	**60.1 ±12.6**	2.00 ±0.00
length_biased	MedCase	Random	0.097 ±0.097	232.0 ±56.7	1.10 ±0.10
		RUCB	0.097 ±0.097	334.2 ±83.8	1.10 ±0.10
		DuelingEXP3	1.695 ±0.386	1012.4 ±159.9	4.20 ±0.93
		Borda	0.390 ±0.159	265.3 ±78.7	1.40 ±0.16
		BALD	1.042 ±0.392	608.9 ±141.4	3.00 ±0.83
		FewerBattles	**0.000 ±0.000**	**85.6 ±7.6**	**1.00 ±0.00**
		GURO	0.274 ±0.192	303.8 ±76.2	1.50 ±0.40
		TOPS (Ours)	**0.000 ±0.000**	**71.2 ±27.1**	**1.00 ±0.00**