SAMPLE EFFICIENT RANKING OF LANGUAGE MODELS USING PAIRWISE COMPARISONS

Anonymous authors
Paper under double-blind review

ABSTRACT

The frontier of large language model (LLM) research is advancing rapidly, with new models being released every few weeks. In this fast-paced environment, rigorous evaluation is crucial to accurately assess a model's quality. Yet evaluation prompts vary in how much evidence they provide about the relative quality of a model pair. In this work, we study how to reduce evaluation cost by adaptively selecting both model pairs and evaluation prompts while controlling comparison errors. Our contribution is twofold: First, we formulate pairwise LLM evaluation as a sequential hypothesis-testing problem and develop an e-process-based procedure. We further derive a sampling distribution for actively selecting evaluation prompts. Second, we construct a complete LLM ranking by using this pairwise evaluation procedure as a black-box comparison function within a sorting algorithm. To further reduce evaluation cost, we adaptively allocate the error budget across pairwise comparisons. Empirically, across both synthetic and real-world setups, we observe that our approach can efficiently rank LLMs using only $\sim$7-22% of the total evaluation budget.

1 INTRODUCTION

Evaluating language models on open-ended tasks often relies on human feedback, which is costly to collect. Platforms such as Chatbot Arena (Chiang et al., 2024) rank models by asking users to compare pairs of anonymous model responses to their prompts. This process is costly: human annotations are expensive, many prompts provide little information for distinguishing models, and the emergence of long-context tasks has made generating model responses substantially more expensive. To reduce these costs, prior work has studied active ranking techniques (Jamieson & Nowak, 2011; Maystre & Grossglauser, 2017; Heckel et al., 2019), which adaptively select the item pairs to compare in order to minimize the total number of comparisons. These techniques typically assume that evaluating an item pair yields an immediate noisy comparison outcome. For LLM evaluation, however, a comparison outcome is not directly observed: it must be elicited by selecting a prompt, generating responses from both models, and collecting a human judgment. Therefore, efficient LLM ranking requires adapting not only over model pairs, but also over the prompts used to compare them.

In this work, we bridge this gap with an algorithm that adaptively selects both the pairs of LLMs to compare and the prompts on which to evaluate them. We achieve this by proposing a sample efficient pairwise evaluation algorithm and leverage this algorithm within a sorting routine to construct a full ranking of models. First, we formulate pairwise LLM evaluation as a sequential hypothesis-testing problem and use the *e-process* framework (Shafer, 2021). An e-process sequentially accumulates evidence against a null hypothesis through betting, yielding an anytime-valid stopping rule with Type-I error controlled at level α. For each model pair, we construct two e-processes, one for each LLM, to gather evidence that the corresponding model is the superior one. We then derive an active sampling distribution that maximizes the e-process's expected growth rate. Its structure prioritizes prompts on which human annotators are likely to confidently express a preference between the two models. Because human preferences are not known in advance, we introduce a practical LLM-driven strategy for approximating this distribution without explicitly modeling them. Second, we use a comparison-based sorting algorithm to turn the pairwise procedure into a full-ranking algorithm. Each comparison invokes the pairwise evaluator as a black box and returns the superior model with a tunable correctness probability of $1 - \alpha$. Finally, to further reduce evaluation cost, we optimize an

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.

end-to-end objective, which jointly balances ranking quality and sample complexity to adaptively allocate the error budget across pairwise comparisons within sorting algorithms.

Our proposed ranking algorithm produces high-quality rankings using only a fraction of the evaluations required by existing approaches. Empirically, we show that our ranking algorithm produces accurate results using only $\sim$7-22% of the total evaluation budget. We further demonstrate that prioritizing discriminative prompts substantially accelerates evidence accumulation during pairwise evaluation. When combined with a comparison-based sorting algorithm, our approach focuses evaluations on the most relevant model pairs while allocating the error budget efficiently. To summarize, our contributions are as follows:

- We frame pairwise evaluation of language models as a sequential hypothesis testing problem and formulate an e-process for this setup. We derive an active sampling distribution that maximizes the growth rate of the e-process, enabling selection of discriminative prompts.
- We use a comparison-based sorting algorithm to construct LLM rankings from our pairwise evaluation procedure. We further adaptively allocate the error budget across comparisons by optimizing an objective that balances ranking quality and sample complexity.
- We provide rigorous theoretical bounds on the expected sample complexity for both the pairwise evaluation and ranking problem.
- Empirically, we show that our proposed approach is able to construct high quality rankings using only $\sim$7-22% of the evaluation budget across several datasets.

2 Background

In this section, we provide an outline of the active testing and active ranking literature and discuss its connection to our LLM ranking setup. We then introduce sequential hypothesis testing, a key component of our proposed pairwise evaluation framework.

Active Testing. Active testing (Kossen et al., 2021; Berrada et al., 2026; Sawade et al., 2010) focuses on adaptively selecting examples to label in order to efficiently estimate the performance of a single model on a given benchmark. Unlike active learning (Houlsby et al., 2011; Farquhar et al., 2021), which aims to reduce uncertainty in model predictions, active testing treats the model as fixed and prioritizes selecting examples that reduce uncertainty about the performance estimate itself (e.g., the variance of the risk estimator (Sawade et al., 2010)). Recent work has improved the proposal distribution for active estimation by leveraging predictions from other LLMs (Wu et al., 2026), while a separate line of work estimates model performance using synthetic labels from LLM verifiers (Zhou et al., 2025b; Angelopoulos et al., 2025; Cowen-Breen et al., 2026). We argue that most active testing techniques are not ideal for pairwise model evaluation, as the primary objective is to reliably identify the better model rather than accurately estimate the win rate. Sawade et al. (2012) introduced early work on active testing with pairwise predictions, prioritizing examples where two classification models' predictions differ. However, extending this approach to modern LLMs is non-trivial, as quantifying the difference between two model responses is itself expensive, and for complex tasks, even generating the responses can be costly. To address this issue, we formulate pairwise evaluation as a hypothesis testing problem and introduce an active selection strategy that accelerates convergence.

Active Ranking. Active ranking techniques can be broadly classified into two categories: methods that rank models using a fixed benchmark (Zhou et al., 2025a; Zouhar et al., 2026) and methods that rank models using pairwise evaluation data (Gu et al., 2026). In our work, we focus on active ranking methods using pairwise comparisons, which reduce annotation costs by adaptively selecting informative pairs. In one of the first works in this setup, Jamieson & Nowak (2011) perform ranking in a d-dimensional Euclidean space by exploiting low-dimensional geometric structure to reduce comparison complexity. Ailon (2011) considers a more general setting and obtains query-efficient approximate rankings under potentially non-transitive preferences. Sorting-based approaches include Maystre & Grossglauser (2017), who analyze quicksort under Bradley-Terry assumptions, and Falahatgar et al. (2017), who combine confidence-based sequential comparisons with merge sort and noisy binary search to obtain PAC ranking guarantees. Heckel et al. (2019) propose an active ranking algorithm that doesn't require parametric assumptions about win probabilities, which keeps track of individual win rates and stops when the confidence interval of the collected data are

sufficiently separated. These methods primarily optimize *which items to compare*, treating observations for each pair as supplied by a noisy comparison oracle. In contrast, our work *jointly optimizes both LLM pair selection and prompt selection* for the corresponding pairwise evaluation, improving ranking efficiency along both axes.

Sequential Hypothesis Testing. Sequential hypothesis testing has a rich theoretical history rooted in Wald's Sequential Probability Ratio Test (SPRT) (Wald, 1992), which tests a simple null $H_0 : \theta = \theta_0$ against a simple alternative $H_1 : \theta = \theta_1$ by accumulating a likelihood ratio statistic $\Lambda_t = \prod_{i=1}^{t} \frac{p_{\theta_1}(X_i)}{p_{\theta_0}(X_i)}$ and stopping at the first time τ this ratio crosses a threshold. Wald & Wolfowitz (1948) established that the SPRT minimizes expected sample size among all tests with prescribed Type I and Type II error rates (α, β). This result established that sequential methods are superior to fixed-sample tests, which require fixing the sample size in advance. Follow-up works (Robbins, 1970; Lai, 1988), extended sequential tests to composite hypotheses and studied asymptotic optimality.

A related line of work reframes sequential testing within a betting or gambling framework. The central object is an *e-value*: a nonnegative random variable E satisfying the bound, $\mathbb{E}_{H_0}[E] \leq 1$, under the null hypothesis. E variable quantifies evidence against the null hypothesis, a larger E indicates stronger evidence (Vovk & Wang, 2021; Grünwald et al., 2024). A canonical instantiation is the *wealth process* shown below:

$$E_t = E_{t-1} e_t(X_t), \text{ where } e_t(X_t) = 1 + \lambda_t(X_t - \mu_0) \tag{1}$$

where $\lambda_t \in \mathcal{F}_{t-1}$ is a bet placed using only past observations, μ_0 is the null mean, and $E_0 = 1$. Under the null hypothesis H_0, $\{E_t\}_{t\geq 0}$ is a nonnegative martingale with $\mathbb{E}_{H_0}[E_t] \leq 1$. Using Ville's inequality (Ville, 1939), we can reject the null hypothesis when $E_t \geq 1/\alpha$ and still control the Type I error at α uniformly over all stopping times. Intuitively, this implies that under the null hypothesis, it is highly unlikely for wealth to grow to a large amount; therefore, we can reject the null hypothesis when $E_t \geq 1/\alpha$. Unlike p-values, products of e-values $\prod_{t=1}^{\tau} e_t(X_t)$ remain valid at any data-dependent τ, making them composable by construction. The optimal bet maximizing expected log-wealth, $\lambda^* = \arg\max_\lambda \mathbb{E}_{H_1}[\log E_t]$, recovers the SPRT under simple hypotheses and extends naturally to composite settings (Shafer, 2021; Waudby-Smith & Ramdas, 2024). Unifying these ideas, an *e-process* $\{E_t\}_{t\geq 0}$ (Shekhar & Ramdas, 2024) requires $\sup_{\tau\in\mathcal{T}} \mathbb{E}_{H_0}[E_\tau] \leq 1$ for *all* stopping times $\tau \in \mathcal{T}$, strictly generalizing test martingales to composite and nonparametric nulls. Recently, e-processes have been applied to LLM evaluation in two concurrent works that address different problem settings. Gu et al. (2026) combine e-processes with i.i.d. prompt sampling to produce an anytime-valid partial ordering of LLMs, whereas we target a full ranking. Hsu & Shekhar (2026) adaptively select prompts to maximize the growth rate of an e-process, but evaluate a single model on a fixed benchmark rather than comparing multiple models. In contrast, our contribution lies in jointly optimizing prompt selection via active sampling and error allocation via adaptive α within a comparison-based sorting algorithm.

3 Efficient Ranking Using Pairwise Comparisons

We introduce a ranking algorithm that makes two decisions: which model pairs to compare and which prompt to use within each comparison. A comparison-based sorting algorithm makes the first decision. A pairwise evaluator makes the second, collecting human judgments until the evidence supports a clear winner. We first construct this evaluator and bound its sample complexity, then use these guarantees to guide the ranking algorithm.

Problem Setup. Consider n models $\{M_1, M_2, \ldots, M_n\}$. For each model pair (M_i, M_j), let $\mathcal{X}_{ij} = \left\{x_1^{ij}, x_2^{ij}, \ldots\right\}$ be a finite prompt population with a fixed **target distribution** $p^{ij}(x) > 0$. A human evaluation is a binary label $Y_{ij} \in \{0, 1\}$, where $Y_{ij} = 1$ indicates a preference for M_j and vice-versa. Write $\pi_{ij}(x)$ for the probability of this outcome on prompt x. The **target win rate**, and the preference gap magnitude are:

$$\mu_{ij} = \sum_{x\in\mathcal{X}_{ij}} p^{ij}(x)\pi_{ij}(x), \qquad \Delta_{ij} = |2\mu_{ij} - 1|. \tag{2}$$

Thus, M_j is preferred when $\mu_{ij} > 1/2$. The target distribution specifies which comparison we want. For example, two equally weighted prompts with $\pi(x)$: 0.9 and 0.6 give a target win rate of 0.75.

3.1 Efficient Pairwise Evaluation

Our goal is to select prompts and based on that sample human judgments. In this section, we fix a model pair (M_i, M_j), and omit its subscripts. At round t, let $\mathcal{F}_{t-1}$ contain the information available before the next judgment. Using this information, choose a **sampling distribution** q_t (which we specify later) and draw a prompt $X_t \sim q_t$. Throughout, we require $q_t(x) > 0$ wherever $p(x) > 0$. To account for noisy human judgment, we assume that each input x is associated with a latent preference probability, $\pi(x) \in [0, 1]$, and the human labels drawn from:

$$\mathbb{P}\big(Y_t = 1 \mid \mathcal{F}_{t-1}, X_t = x\big) = \pi(x). \tag{3}$$

This lets prompt selection to be history-dependent but human labels are conditionally independent.

E-processes and stopping. Since the winner is initially unknown, we test two null hypotheses simultaneously: $H_0^+ : \mu \le 1/2$ (M_i is better than M_j) and $H_0^- : \mu \ge 1/2$ (M_j is better than M_i). We do not consider the possibility of a tie as it is unlikely in real-world settings (the boundary case $\mu = 1/2$ is the only point at which both nulls hold, and is handled separately in Appendix A.1). We construct two e-processes, E_t^+ and E_t^-, that accumulate evidence against H_0^+ and H_0^-, respectively:

$$\begin{aligned} e_t^+(y) &= 1 + \lambda_t(y - 1/2), & E_t^+ &= E_{t-1}^+ \frac{p(X_t)}{q_t(X_t)} e_t^+(Y_t), \\ e_t^-(y) &= 1 - \lambda_t(y - 1/2), & E_t^- &= E_{t-1}^- \frac{p(X_t)}{q_t(X_t)} e_t^-(Y_t), \end{aligned} \tag{4}$$

where $\lambda_t = 4(\nu_t - 1/2)$ (following Breiman (1961)) and the bet $\nu_t \in [1/2, 1)$ controls the strength of each label's contribution. Both processes start from $E_0^+ = E_0^- = 1$, and it is easy to see that E_t^+ grows as we collect evidence for model M_j (when $y = 1$) and E_t^- grows when we collect evidence for M_i (when $y = 0$). We first allow any such bet; next section specifies how to choose it from $\mathcal{F}_{t-1}$. Both processes use the same prompt, judgment, and realized sampling probability. For a given error tolerance $\alpha \in (0, 1)$, we stop when either process crosses the threshold:

$$\tau_\alpha = \inf\{t \ge 1 : \max(E_t^+, E_t^-) \ge 1/\alpha\}. \tag{5}$$

The pairwise algorithm returns M_j if E_t^+ crosses first and M_i otherwise.

Lemma 3.1. *Under equation 3, E_t^+ is a nonnegative supermartingale under H_0^+, and E_t^- is a nonnegative supermartingale under H_0^-. For any $\mu \neq .5$, the probability of stopping and returning the inferior model is at most α, regardless of prediction accuracy of q_t.*

In fact, this lemma shows that E^+ and E^- are valid e-processes and using Ville's inequality we can bound the crossing probability by α (Ville, 1939; Ramdas & Wang, 2025) (see Appendix A.1). At a tie where $\mu = .5$, one can use $2/\alpha$ threshold to control the probability of either declaration by α.

Choosing prompts. To choose prompts, we examine how quickly each process grows on a log-scale. We first treat $\pi(x)$ as known; and then replace it by predictions to obtain an implementable rule. Following the notion of e-power (Ramdas & Wang, 2025), define the contribution of a prompt as the local e-power for each prompt x and each direction $s \in \{+, -\}$: $h_t^s(x) = \mathbb{E}_{Y_t}[\log e_t^s(Y_t) \mid \mathcal{F}_{t-1}, X_t = x]$. Then including the importance weighting, the expected log increment under proposal q is

$$G_t^s(q) = \sum_x q(x) h_t^s(x) - D_{\mathrm{KL}}(q \| p), \qquad \text{for } s \in \{+, -\}. \tag{6}$$

The first term rewards evidence; the second accounts for the cost of changing the prompt distribution. Here D_{KL} denotes KL divergence, including between Bernoulli distributions when its arguments are probabilities. The **optimal proposal** and its rate are

$$q_t^{s,\star}(x) = \frac{p(x) e^{h_t^s(x)}}{\sum_z p(z) e^{h_t^s(z)}}, \qquad G_t^{s,\star} = \log \sum_x p(x) e^{h_t^s(x)}. \tag{7}$$

Indeed, $G_t^s(q) = G_t^{s,\star} - D_{\mathrm{KL}}(q \| q_t^{s,\star})$. This identity identifies both the optimizer and the exact loss from using a different proposal. For our factors it gives

$$q_t^{+,\star}(x) \propto p(x) \left(\frac{\nu_t}{1 - \nu_t}\right)^{\pi(x)}, \qquad q_t^{-,\star}(x) \propto p(x) \left(\frac{1 - \nu_t}{\nu_t}\right)^{\pi(x)}. \tag{8}$$

The positive proposal favors prompts likely to favor M_j; the negative proposal favors the opposite direction. Their optimality concerns one-step log growth at the current bet. Appendix A.2 gives the derivation and its local-e-power interpretation.

The two proposals answer different questions: where to seek evidence for M_j versus for M_i. Since the winner is unknown, we combine them in proportion to their accumulated evidence:

$$\widetilde{q}_t(x) = w_{t-1} q_t^{+,\star}(x) + (1 - w_{t-1}) q_t^{-,\star}(x), \qquad w_t = E_t^+ / (E_t^+ + E_t^-). \tag{9}$$

Initially $w_0 = 1/2$; thereafter, the proposal associated with the larger process receives the larger weight. We call this *the oracle mixture* when it uses the true $\pi(x)$.

In practice, the preference probabilities $\pi(x)$ are unknown and estimating them is challenging. We explore various online learning based predictors $\hat{\pi}(x) = f(x)$ using past observations (Sec. 4.2) but found an LLM-based predictor to be most effective. An LLM predicts $\hat{\pi}(x)$ from prompt texts and previous comparisons; we also consider a predictor trained on prompt embeddings. These predictions guide prompt selection, while the human judgment supplies the label used to update evidence. Replacing π in Eq. 8 gives normalized proposals $\widehat{q}_t^+$ and $\widehat{q}_t^-$. We use

$$q_t = \kappa p + (1 - \kappa)\big[w_{t-1}\widehat{q}_t^+ + (1 - w_{t-1})\widehat{q}_t^-\big], \qquad \kappa \in (0, 1]. \tag{10}$$

This final q_t is the proposal used both to draw the prompt and to compute the importance weight.

Choosing the bet size. The matched target-sampling bet is $\nu^\star = \max\{\mu, 1 - \mu\}$, obtained by maximizing the expected Bernoulli log factor (Breiman, 1961). We estimate it from past judgments:

$$\widehat{\mu}_{t-1} = \mathrm{clip}_{[0,1]}\left(\frac{1}{t-1}\sum_{u<t}\frac{p(X_u)}{q_u(X_u)}Y_u\right), \quad t \geq 2, \tag{11}$$

$$\nu_t = \min\{\bar{\nu}, \max\{\widehat{\mu}_{t-1}, 1 - \widehat{\mu}_{t-1}\}\}, \qquad 1/2 < \bar{\nu} < 1. \tag{12}$$

A prespecified $\nu_1 \in [1/2, \bar{\nu}]$ initializes the procedure. The cap keeps both factors positive. These steps define $\texttt{PairEval}(M_i, M_j, \alpha)$: choose the bet and proposal, draw a prompt and judgment, update both processes, and test Eq. 5.

Pairwise sample complexity. We next relate the evidence rate to the number of human evaluations. Without loss of generality assume $\mu > .5$. To isolate the cost of learning the directions, we first analyze the oracle version of the completed procedure: $\pi(x)$ are known, the bet is fixed $\nu_t = \mu$. The component proposals are therefore fixed, and only their mixture weights adapt, resulting in the correct proposal's growth rate:

$$G^\star = \log \sum_x p(x)(2\mu)^{\pi(x)}\{2(1-\mu)\}^{1-\pi(x)}. \tag{13}$$

The resulting bound connects the local growth objective to expected evaluation cost for this benchmark; the learned rule requires additional growth control.

Theorem 3.1 (Sample complexity)**.** *Suppose $\mu \in (1/2, 1)$ and at least one prompt has $0 < \pi(x) < 1$. For the oracle mixture with bet μ, a finite constant C, independent of $\alpha, \beta \in (0, 1)$, satisfies*

$$\mathbb{E}[\tau_\alpha] \leq \frac{\log(1/\alpha)}{G^\star} + C, \qquad \mathbb{E}[\tau_\alpha] \sim \frac{\log(1/\alpha)}{G^\star} \quad \text{as } \alpha \downarrow 0. \tag{14}$$

Moreover, with probability at least $1 - \beta$,

$$\tau_\alpha \leq \frac{\log(1/\alpha)}{G^\star} + C\left\{\sqrt{\log(1/\alpha)\log(2/\beta)} + \log(2/\beta) + 1\right\}. \tag{15}$$

Learning the direction therefore leaves the leading sample cost unchanged. The proof has a simple structure: importance weights cancel in E_t^+/E_t^-; the expected weight on the wrong proposal decreases geometrically; and its cumulative KL loss is finite. The interior-preference assumption supplies the strict drift needed for this argument. Appendices A.4–A.5 give the proof. The gain over target sampling comes from prompt heterogeneity: $G^\star \geq D_{\mathrm{KL}}(\mu \| 1/2)$, exactly when $\pi(x)$ varies across prompts. Thus the oracle mixture has a smaller leading expected cost in heterogeneous populations (Appendix A.7).

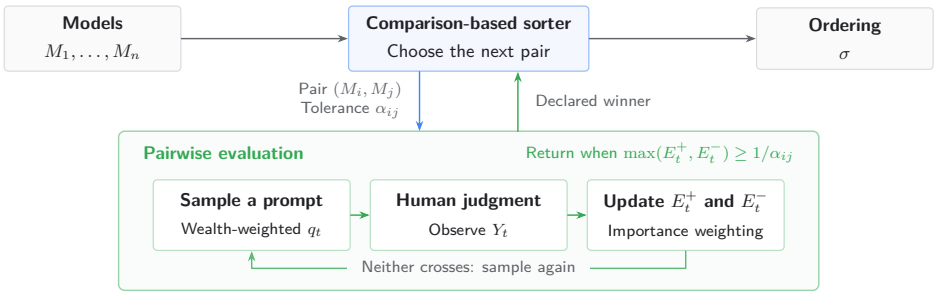


Figure 1: **Ranking with adaptive pairwise evaluation**. A comparison-based sorter requests model pairs. For each pair, the evaluator adaptively samples prompts and updates two e-processes.

When does prompt selection help? Our analysis also explains when imperfect prediction of π can reduce evaluation cost. Suppose that the predictions differ from the true preference probabilities by at most ϵ for every prompt and at every round. With the matched bet $\nu = \mu$, replacing the true probabilities by these predictions reduces the expected log growth of each directional proposal by at most $\frac{\epsilon^2}{2}\log^2(\frac{\mu}{1-\mu})$. This bound quantifies the cost of prediction error. In particular, when this penalty is smaller than the oracle gain, $G^\star - D_{\mathrm{KL}}(\mu \,\|\, 1/2)$, prompt selection can reduce the leading expected sample cost. Appendix A.8-A.9 state the precise condition and extends the analysis to ν_t estimated from past observations. Regardless of prediction accuracy, the comparison-error guarantee in Lemma 3.1 remains valid.

In the following section, we treat the proposed pairwise evaluation algorithm as a black box, accessed via a single function call $\texttt{PairEval}(M_i, M_j, \alpha_{ij})$. This function takes as input a model pair (M_i, M_j) and a target error rate α_{ij}, and returns the superior model with probability at least $1 - \alpha_{ij}$.

3.2 The Ranking Algorithm

We now use the pairwise evaluator to construct an ordering of n LLMs. Pairwise preferences may contain cycles, as illustrated in Figures 3 and 4, so no linear ordering may agree with every pairwise winner. So before describing the algorithm, we need to formalize the ranking objective. For a permutation σ, where larger $\sigma(i)$ means higher rank for M_i, we define the ranking quality as:

$$Q(\sigma) = \sum_{1 \leq i < j \leq n} \mu_{ij} \mathbf{1}[\sigma(j) > \sigma(i)] + (1 - \mu_{ij}) \mathbf{1}[\sigma(i) > \sigma(j)]. \tag{16}$$

Each term measures the expected agreement between ordering σ and the pair's human preference. To improve the sample efficiency of ranking, we reduce both (a) the number of calls to $\texttt{PairEval}(\cdot, \cdot, \cdot)$ and (b) the number of samples required within each call. For (a), we use any general comparison based sorting algorithm. For (b), we adapt the error rate α_{ij} to the pair (M_i, M_j) being compared.

Using a comparison-based sorter. Let $\mathcal{A}$ be a comparison-based ranking algorithm, e.g., mergesort or quicksort. At call k, it requests a pair $e_k = (i_k, j_k)$ and receives the result of $\texttt{PairEval}(M_{i_k}, M_{j_k}, \alpha_k)$. The tolerance α_k is chosen before the call. Each evaluation starts from unit evidence and uses fresh judgments satisfying Eq. 3. $\mathcal{A}$ sees the returned winner, not the evaluator's internal sample path, and uses these winners to choose subsequent comparisons (see Fig. 1).

This interface separates the statistical evaluation from the sorting procedure. Sorting-based active ranking has been studied previously (Maystre & Grossglauser, 2017; Falahatgar et al., 2017); here the evaluator also controls how prompts are selected within each comparison. Next, we derive a principled way to allocate these pair-specific error rates without compromising the ranking quality.

Allocating error across comparisons. We allocate error tolerances across comparisons based on their estimated sampling costs. Let $\mathcal{E}$ be the set of unordered pairs that the ranker may request. If the ranker requests a pair that has already been compared, we use the previously declared winner. Thus, each pair incurs an error tolerance only once and requires no additional labels on subsequent requests (Appendix B extends the analysis to repeated comparisons). When a pair is first requested, obtain estimates of $\widehat{G}_{ij}$ and $\widehat{\Delta}_{ij}$. These estimates may come from a small pilot sample. They are used only to allocate error budget; the validity of the pairwise test does not depend on their accuracy.

The sample complexity bound in equation 14 motivates using $\log(1/\alpha_{ij})/\widehat{G}_{ij}$ as the cost of evaluating pair (i, j). Since the pairs ultimately requested by the ranking algorithm are not known in advance, we minimize the sum over all the candidate pairs. This favors larger tolerances for comparisons that are expensive to resolve. We also account for the smaller direct effect of errors on nearly tied pairs. We use $\widehat{\Delta}_{ij}$ to estimate the magnitude of the change in a pair's contribution to ranking quality when the order of (i, j) is reversed. This motivates the estimated direct-error budget $\sum_{\mathcal{E}} \widehat{\Delta}_{ij}\alpha_{ij} \leq \varepsilon$. Together with the global error budget $\sum_{\mathcal{E}} \alpha_{ij} \leq \delta$ and the bounds $0 < \alpha_{ij} \leq \bar{\alpha}$, this defines a convex cost model whose exact optimizer is given by Eq. 17.

$$\alpha_{ij} = \min\left\{\bar{\alpha}, \frac{1}{\widehat{G}_{ij}(\xi\widehat{\Delta}_{ij} + \zeta)}\right\}, \qquad (i, j) \in \mathcal{E}. \tag{17}$$

The nonnegative multipliers ξ, ζ are chosen to satisfy the budgets and complementary slackness; Appendix B.3 gives the optimization problems and the derivation.

Sample complexity of ranking. We now translate the pairwise guarantees into a guarantee for the full ranking. Let M denote an upper bound on the number of pairwise comparisons requested by the ranking algorithm in one run. For example, merge sort uses $M = O(n \log n)$ comparisons. Also, let $\widehat{\sigma}_{\mathcal{A}}$ denote the ranking returned by Algorithm 1. We compare it with the ordering $\sigma^{\star}_{\mathcal{A}}$ that the same ranking algorithm would return if every requested pair were answered by its true pairwise winner. Thus, $\sigma^{\star}_{\mathcal{A}}$ separates errors due to pairwise evaluation from the behavior of the ranking algorithm.

Theorem 3.2 (Sample complexity of ranking)**.** *Let $\mathcal{E}_{\text{eval}} \subseteq \mathcal{E}$ be the set of pairs evaluated during a run. Suppose that each pairwise test satisfies the assumptions of Theorem 3.1, and the tolerances satisfy $\sum_{(i,j)\in\mathcal{E}} \alpha_{ij} \leq \delta$. Algorithm 1 then satisfies $\mathbb{P}(\widehat{\sigma}_{\mathcal{A}} \neq \sigma^{\star}_{\mathcal{A}}) \leq \delta$. Moreover, let N denote the total number of judgments collected during ranking, and M denote the number of pairwise comparisons requested by the ranking algorithm then*

$$\mathbb{E}[N] \leq M \max_{(i,j)\in\mathcal{E}} \left\{\frac{\log(1/\alpha_{ij})}{G_{ij}} + C_{ij}\right\}, \tag{18}$$

where G_{ij} and C_{ij} are the pairwise quantities in Theorem 3.1.

The theorem bounds the total cost by the number of comparison requests times the expected largest per-pair cost bound encountered during the run. A tighter bound sums costs along the realized path; Appendix B.1 states and proves it, together with Theorem 3.2.

4 Experiments

In this section, we describe our experimental setup and evaluate the proposed pairwise evaluation and sample-efficient ranking algorithm across a range of settings.

4.1 Setup

We perform experiments using pairwise LLM evaluation data available in three datasets: (a) LMArena Human Preference dataset (`arena-human-preference-140k`), which has 140K human labels across 1200+ model pairs, (b) Preference Proxy Evaluations (PPE) (Frick et al., 2024), a collection of 16K labeled human preference pairs from Chatbot Arena containing responses using over 121 languages and (c) the binarized version of the Ultrafeedback dataset (`ultrafeedback-binarized-preferences`), which is a filtered version of Ultrafeedback dataset (Cui et al., 2023) containing 63K pairwise examples from 17 LLMs. For pairwise evaluation, we consider all model pairs in PPE and retain only pairs with at least 200 labeled examples in LMArena and UltraFeedback datasets. For LLM ranking, we first construct a graph with LLMs as nodes, connecting two nodes with available evaluation data. We then extract the largest clique using the algorithm of Boppana & Halldórsson (1992) and restrict the ranking to models within this clique. This procedure yields 46 LLMs in LMArena, 17 in UltraFeedback, and 19 in PPE.

4.2 Pairwise Evaluation

Baselines. We construct baselines to independently evaluate two decoupled components of our pairwise evaluation approach: estimating a prompt's latent preference probability ($\pi(x)$) and the

Table 1: We report the sample efficiency (win rate against uniform sampling; higher is better) and accuracy of pairwise evaluation for different methods across 3 datasets. ROUGE-L with greyed out results has prior access to model answers.

Method	LMArena		PPE		UltraFeedback	
	Win Rate (↑)	Acc. (↑)	Win Rate (↑)	Acc. (↑)	Win Rate (↑)	Acc. (↑)
`Active Testing` **[AT]**	12.20±4.15	93.17±1.76	38.04±2.60	86.41±1.13	17.37±3.19	97.46±0.65
`BLR exploit` **[BLE]**	53.66±7.40	**94.63**±1.57	48.10±2.83	85.87±1.15	44.07±4.37	97.97±0.58
`BLR` ϵ`-greedy` **[BLG]**	45.12±7.28	93.17±1.76	53.53±2.80	**87.39**±1.09	44.49±4.39	**98.31**±0.53
`BLR Thompson` **[BLT]**	26.83±6.24	93.66±1.70	40.49±2.64	87.28±1.10	36.02±4.19	98.14±0.56
`LLM` **[LM]**	**57.32**±7.02	91.71±1.93	**56.79**±2.90	85.65±1.16	**59.32**±4.32	97.46±0.65
`ROUGE-L` **[RL]**	42.68 ±7.02	94.63 ±1.57	44.29 ±2.83	85.43 ±1.16	41.10 ±4.39	96.95 ±0.71

active sampling distribution, $\tilde{q}_t(x)$ (Eq. 9). To evaluate *preference-probability estimation*, we fix the sampling distribution to the derived optimal proposal and compare five $\hat{\pi}(x)$ predictors: **[LM]** an `LLM estimator`, where an LLM assigns $\pi(x)$ scores to prompts using the raw prompt text and previously observed comparison outcomes; three variants of *embedding-based estimation* using online Bayesian logistic regression (`BLR`) (Bishop & Nasrabadi, 2006) model, where we estimate $\hat{\pi}(x) = \sigma(\mu^\top \text{emb}(x))$ and sequentially update parameters after each observation. These variants differ in how the next example is selected using **[BLE]** `BLR exploit`, by plugging in $\hat{\pi}(x)$ estimates into the derived sampling distribution (Eq. 9), **[BLG]** `BLR` ϵ`-greedy` that uses the proposal distribution with probability $(1-\epsilon)$ and uniform sampling with probability, ϵ, and **[BLT]** `BLR Thompson`, which uses Thompson Sampling (Agrawal & Goyal, 2013) and considers the average proposal distribution across posterior parameter draws $\theta \sim \mathcal{N}(\mu, \Sigma)$. Next, we compare against two different active sampling distributions: **[AT]** `Active Testing` (Kossen et al., 2021), which targets prompts with maximal preference uncertainty ($q(x) \propto \sqrt{\hat{\pi}(x)(1-\hat{\pi}(x))}$), we use `BLR` ϵ`-greedy` to estimate $\hat{\pi}(x)$ in this setup, and **[RL]** `ROUGE-L divergence` (Lin, 2004) distribution, which completely bypasses preference probability estimation by prioritizing prompts that yield the most dissimilar candidate answers (low ROUGE-L). Unlike other baselines, ROUGE distribution requires access to model answers. We use Gemini 3.6 Flash model as the LLM estimator and obtain prompt embeddings from Gemini Embedding 2 (Shanbhogue et al., 2026). We report the average results across 5 runs for all baselines. We report additional results using open-sourced LLMs in Appendix E.2.

Metrics. We assess pairwise evaluation methods along two axes: *correctness*, measured as the percentage of evaluations that identified the correct winner; and *sample efficiency*, measured as the *win rate against uniform sampling*, i.e., the fraction of trials in which the method required fewer samples than the uniform baseline.

Results. We report the pairwise evaluation results in Table 1. The `LLM` estimator **[LM]** consistently achieves the highest win rate (better sample efficiency) across all three datasets, outperforming uniform sampling by up to ∼9% while maintaining competitive accuracy. Among embedding-based methods, `BLR exploit` **[BLE]** and `BLR` ϵ`-greedy` **[BLG]** show moderate sampling efficiency in a few settings. `Active Testing` **[AT]** and `ROUGE-L` **[RL]** both consistently underperform uniform sampling, indicating that neither uncertainty-based selection nor response divergence is an effective proxy for preference informativeness. Accuracy on PPE is relatively lower across all methods as all LLM pairs have fewer than 100 annotations. We observe that the `LLM estimator` significantly outperforms the `ROUGE-L divergence` baseline despite having no access to model answers. We use the LLM-based predictor to report the results of sorting-based variants in the following section.

4.3 Efficient Ranking

Baselines. We compare our approach against three classes of baselines. The first class provides stopping criteria for the ranking data acquisition process and consists of variants of our algorithm. We ablate two components—the sorting algorithm and error tolerance allocation, to construct the following baselines: **[QS]** quicksort with a constant error tolerance (α) across all pairs; **[QSD]** quicksort with a doubling error tolerance strategy, where the earliest recursion steps receive the tightest allocation (low α_0) that is gradually relaxed; **[QSA]** quicksort with adaptive error allocation (Eq. 17); **[MS]** merge sort with a constant error tolerance; and **[MSA]** merge sort with adaptive error allocation (Eq. 17). The quicksort baselines with constant error allocation (α) are a close variant of Maystre & Grossglauser (2017), where we plug-in e-process-based pairwise evaluation

Table 2: We report the quality of ranking (agreement with pairwise data) and fraction of samples used with standard error over 3 runs. A higher quality score and lower fraction of samples is better.

Method	LMArena		PPE		UltraFeedback	
	Quality (↑)	Samples (↓)	Quality (↑)	Samples (↓)	Quality (↑)	Samples (↓)
`Active Ranking` [AR]	59.8±0.5	72.9%±1.8	57.4±0.8	51.8%±4.4	65.4±3.3	64.8%±11.0
`Active Testing` (w/ QS) [ATQ]	60.6±0.5	19.0%±1.0	59.2±0.1	30.9%±1.2	75.0±0.1	19.9%±0.7
`Active Testing` (w/ MS) [ATM]	61.0±0.2	16.8%±0.5	59.0±0.5	29.8%±1.3	**75.1**±0.0	19.9%±1.2
`Quick Sort` (const. α) [QS]	**61.9**±0.1	15.0%±0.7	58.4±0.3	29.7%±0.1	74.9±0.1	11.5%±0.8
`Quick Sort` (doubling α) [QSD]	61.3±0.1	15.9%±0.4	58.4±0.2	31.3%±1.1	75.0±0.1	15.4%±0.2
`Merge Sort` (const. α) [MS]	60.8±0.2	14.4%±0.5	59.2±0.2	28.1%±0.8	75.0±0.1	12.2%±0.8
`Quick Sort` (adaptive α) [QSA] (Ours)	61.0±0.3	**9.7%**±0.8	**59.3**±0.1	21.8%±1.3	74.7±0.2	**6.8%**±0.5
`Merge Sort` (adaptive α) [MSA] (Ours)	60.1±0.5	**9.7%**±0.8	58.3±0.4	**21.7%**±1.3	74.5±0.3	7.6%±0.9
`Expected (MC)` [EX]	50.0 ±0.0	-	50.0 ±0.0	-	50.1 ±0.1	-
`Bradley-Terry` [BT]	62.6 ±0.0	100% ±0.0	59.8 ±0.0	100% ±0.0	75.0 ±0.0	100%±0.0
`HodgeRank` [HG]	62.5 ±0.0	100% ±0.0	59.8 ±0.0	100% ±0.0	75.0 ±0.0	100% ±0.0
`Oracle` [OE]	63.7 ±0.0	100% ±0.0	60.8 ±0.0	100% ±0.0	75.2 ±0.0	100% ±0.0

for noisy comparisons. The second class of methods consists of active estimation baselines: (a) Active Ranking [AR] (Heckel et al., 2019), which performs sequential ranking, and (b) Active Testing (Kossen et al., 2021), a pairwise evaluation method that we plug into quick sort [ATQ] and merge sort [ATM]. The third class comprises offline methods that use the entire dataset to construct a ranking, providing an estimate of the upper bound on ranking quality. We consider [BT] Bradley-Terry ranking (Bradley & Terry, 1952), which assumes a transitive total ordering by assigning each LLM a scalar strength parameter; [HG] HodgeRank (Jiang et al., 2011), which relaxes the transitivity assumption and accommodates cyclic preferences in pairwise comparisons; [OE] oracle estimation of the quality upper bound, $Q(\sigma_o) = \sum_{1 \le i < j \le n} \max\{\mu_{ij}, 1 - \mu_{ij}\}$ (note that no linear ordering is guaranteed to achieve this bound); and [EX] expected quality achieved by using 10k random permutations (irrespective of the underlying dataset size).

Metrics. Evaluating LLM rankings is challenging because we lack access to a gold ranking and the stochastic nature of pairwise evaluation introduces additional uncertainty. Moreover, when we construct a directed preference graph with LLMs as nodes, adding edges based on pairwise evaluation outcomes, we find many strongly connected components, indicating that a single optimal linear ordering may not exist (Fig. 3, 4). Therefore, we evaluate the quality of a predicted ranking (σ) by computing its agreement with the pairwise data, $Q(\sigma)$ (we report the fraction in Eq. 16 as percentages). For sample efficiency, we report the fraction of total pairwise evaluations used to construct the ranking. We execute all algorithms 3 times and use $\alpha = 0.05$ for pairwise comparison calls.

Results. Table 2 summarizes the ranking results. All sorting-based methods achieve quality close to the offline full-data baselines ([BT], [HG]), trailing them by less than 1% on average, while using only a fraction of the total evaluations. In contrast, [AR] requires 52-73% of the evaluations yet yields the lowest quality on every dataset. Plugging Active Testing into sorting ([ATQ], [ATM]) achieves competitive quality but consumes up to ∼3x more samples than the best performing method, [MSA]. Our adaptive sorting variants ([QSA], [MSA]) are the two most sample-efficient methods across all settings, reducing the sample budget by up to ∼40% compared to their constant error allocated counterparts while achieving similar quality. The ranking quality obtained by the adaptive variants is also comparable to the offline methods that use all evaluations. These results show that our framework produces high quality rankings using only a fraction of the total evaluation budget.

5 Conclusion

We propose a sample-efficient LLM ranking framework that optimizes both LLM-pair selection and prompt selection for evaluation. Our pairwise evaluator combines e-process-based hypothesis testing with active sampling. We theoretically analyze how variation in prompt preferences can accelerate evidence accumulation and reduce evaluation cost. An LLM-driven sampling strategy makes this approach practical without access to latent preference probabilities. We integrate the pairwise procedure into comparison-based sorting algorithm, and derive an error allocation rule that balances evaluation cost against direct comparison errors. Experiments on synthetic and real-world preference data demonstrate the potential of active prompt selection to improve pairwise evaluation efficiency. Future work can analyze how errors in predicted prompt preferences affect evaluation cost and how reliably the final ranking reflects human judgments when pairwise preferences conflict.

AI USE STATEMENT

In this work, we used generative AI tools for sentence-level rewrites, executing human-written code and logging results, and obtaining partial proofs of Theorem 3.2 and second half of Theorem 3.1. We have not used generative AI tools for writing large sections of the paper, curating references and writing the core logic behind e-process based sampling. We have reviewed all AI-assisted work including proofs, paper text, and analysis code. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

We acknowledge that this work adheres to ICLR's code of ethics. This works proposes a sample efficient evaluation algorithm to rank LLMs. We do not anticipate any misuse or harm feasible from the research of this work. In this work, we rely on publicly available and open-source datasets.

REPRODUCIBILITY STATEMENT

We have described the details of our experimental setup, datasets, baselines, prompts, and hyperparameters in Section 4, Appendix C, and Appendix E.1. We run every experiment multiple times (3-5 times) and report the mean along with the standard error for all results reported in Tables 1 and 2.

REFERENCES

Shipra Agrawal and Navin Goyal. Thompson sampling for contextual bandits with linear payoffs. In Sanjoy Dasgupta and David McAllester (eds.), *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pp. 127–135, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR. URL https://proceedings.mlr.press/v28/agrawal13.html.

Nir Ailon. Active learning ranking from pairwise preferences with almost optimal query complexity. *Advances in Neural Information Processing Systems*, 24, 2011.

Anastasios N Angelopoulos, Jacob Eisenstein, Jonathan Berant, Alekh Agarwal, and Adam Fisch. Cost-optimal active ai model evaluation. *arXiv preprint arXiv:2506.07949*, 2025.

Kazuoki Azuma. Weighted sums of certain dependent random variables. *Tohoku Mathematical Journal, Second Series*, 19(3):357–367, 1967.

Gabrielle Berrada, Jannik Kossen, Freddie Bickford Smith, Muhammed Razzak, Yarin Gal, and Thomas Rainforth. Scaling up active testing to large language models. *Advances in Neural Information Processing Systems*, 38:173670–173697, 2026.

Patrick Billingsley. *Convergence of probability measures*, volume 2. wiley New York, 1999.

Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*, volume 4. Springer, 2006.

Ravi Boppana and Magnús M Halldórsson. Approximating maximum independent sets by excluding subgraphs. *BIT Numerical Mathematics*, 32(2):180–196, 1992.

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.

Leo Breiman. Optimal gambling systems for favorable games, 1961.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. Chatbot arena: An open platform for evaluating LLMs by human preference. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=3MW8GKNyzI.

Charlie Cowen-Breen, Alekh Agarwal, Stephen Bates, William W Cohen, Jacob Eisenstein, Amir Globerson, and Adam Fisch. Multiple-prediction-powered inference. In *The Fourteenth International Conference on Learning Representations*, 2026.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Wei Zhu, Yuan Ni, Guotong Xie, Zhiyuan Liu, and Maosong Sun. Ultrafeedback: Boosting language models with high-quality feedback, 2023.

Moein Falahatgar, Alon Orlitsky, Venkatadheeraj Pichapati, and Ananda Theertha Suresh. Maximum selection and ranking under noisy comparisons. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1088–1096. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/falahatgar17a.html.

Sebastian Farquhar, Yarin Gal, and Tom Rainforth. On statistical bias in active learning: How and when to fix it. In *International Conference on Learning Representations*, 2021.

Evan Frick, Tianle Li, Connor Chen, Wei-Lin Chiang, Anastasios N. Angelopoulos, Jiantao Jiao, Banghua Zhu, Joseph E. Gonzalez, and Ion Stoica. How to evaluate reward models for rlhf, 2024. URL https://arxiv.org/abs/2410.14872.

Peter Grünwald, Rianne de Heide, and Wouter Koolen. Safe testing. *Statistical Methodology*, 86: 1091–1128, 2024.

Runzhe Gu, Wenguang Sun, Bowen Gang, and Xintao Xia. Anytime-valid inference for online ranking of large language models. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=7Y8xRnGQ47.

Reinhard Heckel, Nihar B. Shah, Kannan Ramchandran, and Martin J. Wainwright. Active ranking from pairwise comparisons and when parametric assumptions do not help. *The Annals of Statistics*, 47(6):pp. 3099–3126, 2019. ISSN 00905364, 21688966. URL https://www.jstor.org/stable/26867160.

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

Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. Bayesian active learning for classification and preference learning. *stat*, 1050:24, 2011.

Chia-Yu Hsu and Shubhanshu Shekhar. Efficient sequential evaluation of large language models. *arXiv preprint arXiv:2607.17409*, 2026.

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

Xiaoye Jiang, Lek-Heng Lim, Yuan Yao, and Yinyu Ye. Statistical ranking and combinatorial hodge theory. *Mathematical Programming*, 127(1):203–244, 2011.

Jannik Kossen, Sebastian Farquhar, Yarin Gal, and Tom Rainforth. Active testing: Sample-efficient model evaluation. In *International Conference on Machine Learning*, pp. 5753–5763. PMLR, 2021.

Tze Leung Lai. Nearly optimal sequential tests of composite hypotheses. *The Annals of Statistics*, pp. 856–886, 1988.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-1013/.

Lucas Maystre and Matthias Grossglauser. Just sort it! a simple and effective approach to active preference learning. In *International Conference on Machine Learning*, pp. 2344–2353. PMLR, 2017.

Aaditya Ramdas and Ruodu Wang. Hypothesis testing with e-values. *Foundations and Trends® in Statistics*, 1(1-2):1–390, 2025.

Herbert Robbins. Statistical methods related to the law of the iterated logarithm. *The Annals of Mathematical Statistics*, 41(5):1397–1409, 1970.

Christoph Sawade, Niels Landwehr, Steffen Bickel, and Tobias Scheffer. Active risk estimation. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pp. 951–958, 2010.

Christoph Sawade, Niels Landwehr, and Tobias Scheffer. Active comparison of prediction models. *Advances in neural information processing systems*, 25, 2012.

Glenn Shafer. Testing by betting: A strategy for statistical and scientific communication. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 184(2):407–431, 2021. doi: https://doi.org/10.1111/rssa.12647. URL https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/rssa.12647.

Madhuri Shanbhogue, Zhe Li, Shanfeng Zhang, Gustavo Hernández Ábrego, Shih-Cheng Huang, Aashi Jain, Daniel Salz, Sonam Goenka, Chaitra Hegde, Ji Ma, et al. Gemini embedding 2: A native multimodal embedding model from gemini. *arXiv preprint arXiv:2605.27295*, 2026.

Shubhanshu Shekhar and Aaditya Ramdas. Nonparametric two-sample testing by betting. *IEEE Transactions on Information Theory*, 70(2):1178–1203, 2024.

Lieven Vandenberghe and Stephen Boyd. *Convex optimization*, volume 1. Cambridge university press Cambridge, 2004.

Jean Ville. *Etude critique de la notion de collectif*, volume 3. Gauthier-Villars Paris, 1939.

Vladimir Vovk and Ruodu Wang. E-values: Calibration, combination and applications. *The Annals of Statistics*, 49(3):1736–1754, 2021.

Abraham Wald. Sequential tests of statistical hypotheses. In *Breakthroughs in statistics: Foundations and basic theory*, pp. 256–298. Springer, 1992.

Abraham Wald and Jacob Wolfowitz. Optimum character of the sequential probability ratio test. *The Annals of Mathematical Statistics*, pp. 326–339, 1948.

Ian Waudby-Smith and Aaditya Ramdas. Confidence sequences for sampling without replacement. *Advances in neural information processing systems*, 33:20204–20214, 2020.

Ian Waudby-Smith and Aaditya Ramdas. Estimating means of bounded random variables by betting. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 86(1), 2024.

Skyler Wu, Yash Nair, and Emmanuel J Candés. Efficient evaluation of llm performance with statistical guarantees. *arXiv preprint arXiv:2601.20251*, 2026.

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, 2025a.

Zhaoyi Zhou, Yuda Song, and Andrea Zanette. Accelerating unbiased llm evaluation via synthetic feedback. In *International Conference on Machine Learning*, pp. 79150–79167. PMLR, 2025b.

Vilém Zouhar, Julia Kreutzer, Alon Lavie, Tom Kocmi, Matt Post, Ondřej Bojar, and Mrinmaya Sachan. Dynamically allocating evaluation effort for model ranking. *arXiv preprint arXiv:2608.03437*, 2026.

A Analysis of Pairwise Evaluation

The proofs follow the same progression as the method. We first verify error control and derive the directional proposals. We then prove the general sample-cost bound and show how the oracle mixture satisfies its growth condition. The concentration and central limit results build on this same argument. Finally, we quantify the effects of prediction error, learning the bet, and shortlisting, before discussing other fixed bets and a deterministic-label boundary case.

The central chain of implications is: strict drift of the evidence ratio, geometric decay of the wrong-direction weight, summable KL loss, and then the leading stopping-cost bound. Appendix A.4 proves this chain. Appendix A.3 supplies the general growth-to-cost step rather than an additional assumption about the oracle. The learned-policy arguments replace the summable oracle loss by controlled prediction, coverage, and betting losses.

Throughout, the prompt population is finite, $p(x) > 0$, and proposals and bets are predictable. Full support is assumed for the growth and infinite-horizon cost analysis; Appendix A.13 explicitly relaxes support for validity. All logarithms are natural. In sample-cost proofs we label the models so that $\mu > 1/2$, and use

$$B = \log(1/\alpha), \qquad S_t = \log E_t^+.$$

The negative-direction case follows by exchanging labels. When an argument uses an infinite observation sequence, we continue the sampling policy beyond its stopping time. This continuation does not change the procedure before termination and lets us compare different thresholds on the same sample path.

A.1 Validity of the pairwise evaluator

Proof of Lemma 3.1. The only inputs here are the fixed conditional label law, predictability, and full support. Prediction accuracy is not used. Full support lets the importance weights remove the proposal on every target prompt; predictability lets us hold those weights and the bet fixed inside the conditional expectation. Equation 3 gives

$$\begin{aligned}\mathbb{E}\left[\frac{p(X_t)}{q_t(X_t)} e_t^+(Y_t) \,\middle|\, \mathcal{F}_{t-1}\right] &= \sum_x p(x)\{2\nu_t\pi(x) + 2(1-\nu_t)(1-\pi(x))\} \\ &= 1 + 4(\nu_t - 1/2)(\mu - 1/2).\end{aligned}$$

Under H_0^+, the last expression is at most one. Multiplying by the nonnegative, $\mathcal{F}_{t-1}$-measurable E_{t-1}^+ proves the supermartingale property. Positivity follows from $p/q_t > 0$ and $1/2 \le \nu_t < 1$. For the negative direction the same calculation gives

$$\mathbb{E}\left[\frac{p(X_t)}{q_t(X_t)} e_t^-(Y_t) \,\middle|\, \mathcal{F}_{t-1}\right] = 1 - 4(\nu_t - 1/2)(\mu - 1/2) \le 1$$

under H_0^-.

Now suppose $\mu > 1/2$. An incorrect declaration requires the negative process to cross $1/\alpha$, because that is the process supporting the inferior model. Ville's inequality (Ville, 1939; Ramdas & Wang, 2025) therefore yields

$$\mathbb{P}\{\text{an incorrect declaration}\} \le \mathbb{P}\{\sup_{t\ge 0} E_t^- \ge 1/\alpha\} \le \alpha.$$

The event inclusion also holds for simultaneous crossings, regardless of the fixed tie-breaking rule. Exchange the directions when $\mu < 1/2$.

At $\mu = 1/2$, both processes are supermartingales. At threshold $2/\alpha$, each crossing probability is at most $\alpha/2$, so their union has probability at most α. This is an error statement, not a finite-sample termination guarantee at a tie. □

A.2 Local e-power and the directional oracle

Definition A.1 (Local e-power). *For direction $s \in \{+,-\}$, the local e-power of prompt x is*

$$h_t^s(x) = \mathbb{E}[\log e_t^s(Y_t) \mid \mathcal{F}_{t-1}, X_t = x]. \tag{19}$$

Lemma A.1. *The local e-powers of the factors in Eq. 4 are*

$$h_t^+(x) = D_{\mathrm{KL}}(\pi(x)\|1/2) - D_{\mathrm{KL}}(\pi(x)\|\nu_t),$$
$$h_t^-(x) = D_{\mathrm{KL}}(\pi(x)\|1/2) - D_{\mathrm{KL}}(\pi(x)\|1-\nu_t).$$

Lemma A.2 (Directional oracle)**.** *For either direction and every proposal q,*

$$G_t^s(q) = G_t^{s,\star} - D_{\mathrm{KL}}(q\|q_t^{s,\star}).$$

Consequently, $q_t^{s,\star}$ uniquely maximizes G_t^s.

Proof of Lemmas A.1 and A.2. We compute the unweighted contribution first, then include the importance weight. This order explains why the proposal maximizes a reward minus a KL cost rather than local e-power alone.

Step 1: compute a prompt's local contribution. For Bernoulli probabilities $a \in [0,1]$ and $b \in (0,1)$, expanding the two divergences gives

$$D_{\mathrm{KL}}(a\|1/2) - D_{\mathrm{KL}}(a\|b) = a\log(2b) + (1-a)\log(2(1-b)).$$

The convention $0\log 0 = 0$ covers the endpoints. Substituting $a = \pi(x)$ and $b = \nu_t$ or $1-\nu_t$ proves

$$h_t^+(x) = D_{\mathrm{KL}}(\pi(x)\|1/2) - D_{\mathrm{KL}}(\pi(x)\|\nu_t),$$
$$h_t^-(x) = D_{\mathrm{KL}}(\pi(x)\|1/2) - D_{\mathrm{KL}}(\pi(x)\|1-\nu_t). \tag{20}$$

Step 2: include the sampling cost. Under proposal q, the complete log factor has conditional expectation

$$\sum_x q(x)\left\{h_t^s(x) + \log\frac{p(x)}{q(x)}\right\} = \sum_x q(x)h_t^s(x) - D_{\mathrm{KL}}(q\|p).$$

The order of the KL arguments matters: the expectation is under q, the distribution actually used to draw the prompt.

Step 3: identify the optimizer and the exact loss. The definition of $q_t^{s,\star}$ implies

$$\log\frac{q_t^{s,\star}(x)}{p(x)} = h_t^s(x) - G_t^{s,\star}.$$

Substitution into the divergence gives

$$\begin{aligned} D_{\mathrm{KL}}(q\|q_t^{s,\star}) &= D_{\mathrm{KL}}(q\|p) - \sum_x q(x)h_t^s(x) + G_t^{s,\star} \\ &= G_t^{s,\star} - G_t^s(q). \end{aligned}$$

The divergence is nonnegative and equals zero only when $q = q_t^{s,\star}$. This proves both optimality and uniqueness.

Finally,

$$h_t^+(x) = \log(2(1-\nu_t)) + \pi(x)\log\frac{\nu_t}{1-\nu_t},$$

and the negative expression replaces the constant by $\log(2\nu_t)$ and reverses the sign of the second term. Constants independent of x cancel on normalization, giving Eq. 8. □

For reference, the matched target-sampling bet and the estimated mixture are

$$\nu^\star = \max\{\mu, 1-\mu\}, \qquad \mu \in (0,1), \tag{21}$$

$$\widehat{q}_t = w_{t-1}\widehat{q}_t^{+} + (1-w_{t-1})\widehat{q}_t^{-}. \tag{22}$$

Neither matching the target bet nor knowing the preferences is needed for the validity lemma.

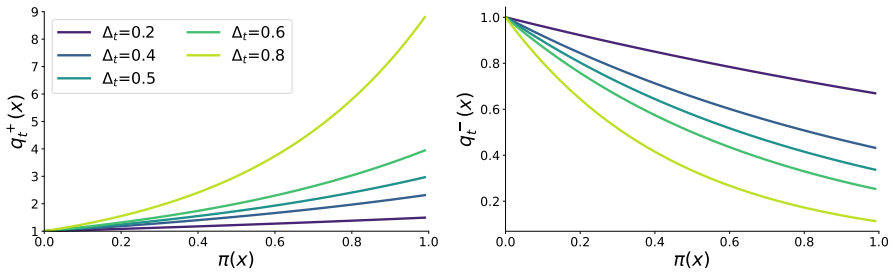


Figure 2: **Directional prompt selection.** Unnormalized tilt factors increase with the preference probability for the positive direction and decrease for the negative direction. The curves are indexed by bet strength $\Delta_t = 2\nu_t - 1$. The normalized proposals also include the target weight $p(x)$, as in Eq. 8.

A.3 Sample cost from cumulative evidence growth

The conditional growth defined in Eq. 6 obeys

$$\mathbb{E}[\log(E_t^+/E_{t-1}^+) \mid \mathcal{F}_{t-1}] = G_t^+(q_t) = G_t^{+,\star} - D_{\mathrm{KL}}(q_t \| q_t^{+,\star}). \tag{23}$$

The next theorem converts cumulative growth into stopping cost. Its assumptions must be established for the policy under study; they do not follow from error validity alone.

Theorem A.1 (Sample complexity under adaptive sampling)**.** *Suppose $\mu > 1/2$, $q_t \geq \kappa p$ for fixed $\kappa > 0$, and $1/2 \leq \nu_t \leq \bar{\nu} < 1$. Let b be a finite bound on the absolute log increment,*

$$\left|\log(E_t^+/E_{t-1}^+)\right| \leq b.$$

Such a bound exists under these conditions. Extend the policy beyond termination, and suppose that, for some $G > 0$ and nonnegative R with $\mathbb{E}[R] < \infty$,

$$\sum_{u=1}^{t} G_u^+(q_u) \geq tG - R \qquad \textit{for every } t \geq 1 \quad \textit{almost surely}. \tag{24}$$

Then

$$\mathbb{E}[\tau_\alpha] \leq \frac{B + b + \mathbb{E}[R]}{G}. \tag{25}$$

For every $r \geq 0$ and integer t with $tG > B + r$,

$$\mathbb{P}(\tau_\alpha > t) \leq \mathbb{P}(R > r) + \exp\left\{-\frac{(tG - B - r)^2}{8b^2 t}\right\}. \tag{26}$$

The policy and the bounds b, G, R are taken uniformly over α.

Proof of Theorem A.1. We separate the proof into the predictable part of log evidence and its random fluctuations. The growth assumption controls the first part; bounded increments control the second. The statement is conditional on the supplied growth bound; no step of this proof establishes that bound for an arbitrary predictor. The subsequent oracle and approximation proofs provide it for the policies they analyze.

Step 1: bound increments and isolate a martingale. Both directional factors lie in $[2(1-\bar{\nu}), 2\bar{\nu}]$. Since $q_t(x) \leq 1$ and $q_t(x) \geq \kappa p(x)$,

$$\log\{2(1-\bar{\nu}) \min_x p(x)\} \leq S_t - S_{t-1} \leq \log(2\bar{\nu}/\kappa). \tag{27}$$

Thus one valid choice in the theorem is

$$b = \max\left\{\log \frac{1}{2(1-\bar{\nu}) \min_x p(x)}, \log \frac{2\bar{\nu}}{\kappa}\right\}.$$

The first term is nonnegative because $\bar{\nu} > 1/2$ and $\min_x p(x) \le 1$. By Eq. 23,

$$S_t = \sum_{u=1}^{t} G_u^+(q_u) + M_t, \qquad M_t := \sum_{u=1}^{t} \{S_u - S_{u-1} - G_u^+(q_u)\}. \tag{28}$$

The process M_t is a martingale starting at zero. Each increment has absolute value at most $2b$, because both the observed increment and its conditional mean lie in $[-b, b]$.

Step 2: bound expected cost without assuming it is finite. We cannot apply a stopping argument at τ_α before proving its integrability. Instead, fix an integer m and use the bounded stopping time $T_m = \tau_\alpha \wedge m$. The event $\{T_m \ge t\}$ is known at time $t-1$, so

$$\begin{aligned}\mathbb{E}[S_{T_m}] &= \sum_{t=1}^{m} \mathbb{E}[\mathbf{1}_{\{T_m \ge t\}}(S_t - S_{t-1})] \\ &= \mathbb{E}\left[\sum_{t=1}^{T_m} G_t^+(q_t)\right] \ge G\mathbb{E}[T_m] - \mathbb{E}[R].\end{aligned}$$

The last inequality applies the pathwise growth assumption at the random integer T_m; the same assumption holds at every integer on a common probability-one event.

Before termination, both log processes are below B. At termination the correct process can increase by at most b, whether or not it is the process that causes stopping. Therefore $S_{T_m} \le B + b$ in every case. Combining the two bounds gives

$$G\mathbb{E}[T_m] \le B + b + \mathbb{E}[R].$$

As $m \to \infty$, monotone convergence proves Eq. 25. In particular, τ_α is finite almost surely.

Step 3: bound the probability of taking too many samples. On $\{\tau_\alpha > t, R \le r\}$, the correct log evidence has not reached B, while its cumulative conditional growth is at least $tG - r$. Equation 28 therefore implies

$$M_t < -(tG - B - r).$$

For $tG > B + r$, the Azuma–Hoeffding inequality for the bounded martingale differences (Azuma, 1967; Hoeffding, 1963) yields

$$\mathbb{P}\{M_t \le -(tG - B - r)\} \le \exp\left\{-\frac{(tG - B - r)^2}{8b^2 t}\right\}.$$

Adding the probability of $R > r$ proves Eq. 26. Independence between the deficit and the martingale is not required. □

An explicit sample budget. Given $\beta \in (0, 1)$, choose a deterministic $r \ge 0$ satisfying $\mathbb{P}(R > r) \le \beta/2$. Then a valid upper quantile is

$$U(\alpha, \beta) = \left\lceil \left\{ \frac{b\sqrt{2\log(2/\beta)} + \sqrt{2b^2\log(2/\beta) + G(B + r)}}{G} \right\}^2 \right\rceil. \tag{29}$$

Indeed, the expression before rounding solves $Gt - B - r = 2b\sqrt{2t\log(2/\beta)}$. For larger t, the martingale term in Eq. 26 is at most $\beta/2$. The deficit contributes the other $\beta/2$.

A convenient, slightly larger expression is

$$\left\lceil \frac{B + r}{G} + \frac{2b\sqrt{2(B + r)\log(2/\beta)}}{G^{3/2}} + \frac{8b^2\log(2/\beta)}{G^2} \right\rceil. \tag{30}$$

It follows from Eq. 29 using $\sqrt{x + y} \le \sqrt{x} + \sqrt{y}$ after expanding the square. This form makes the leading threshold term, the square-root fluctuation term, and the confidence term explicit.

For sharper constants, the expected-cost numerator can use the upper endpoint $\log(2\bar{\nu}/\kappa)$ in Eq. 27 instead of b. Likewise, the martingale-increment bound $2b$ can be replaced by the width of that interval. The single constant b in the theorem avoids carrying both endpoints through the statements.

An expectation-only variant. A bound on expected cost does not need a pathwise deficit or a tail bound for that deficit. It suffices to control the expected growth accumulated up to bounded stopping times.

Proposition A.1 (Sample complexity from expected cumulative growth)**.** *Under the bounded-increment assumptions above, suppose that for constants $G > 0$ and $C \geq 0$, independent of α,*

$$\mathbb{E}\left[\sum_{t=1}^{\tau_\alpha \wedge m} G_t^+(q_t)\right] \geq G\mathbb{E}[\tau_\alpha \wedge m] - C \qquad \textit{for every } m \geq 1. \tag{31}$$

Then

$$\mathbb{E}[\tau_\alpha] \leq \frac{B + b + C}{G}. \tag{32}$$

Proof. Step 2 above gives $G\mathbb{E}[T_m] - C \leq \mathbb{E}[S_{T_m}] \leq B + b$. Monotone convergence completes the argument. □

The expectation-only condition does not itself supply an exponential upper tail. For that conclusion, one must also control realized cumulative growth, as in Theorem A.1.

A.4 Expected sample cost of the oracle mixture

We prove the expected-cost and asymptotic claims in Theorem 3.1; the upper-tail claim follows in the next subsection. The key intermediate conclusion is

$$\sum_{t \geq 1} \mathbb{E}[D_{\mathrm{KL}}(q_t \| q^+)] < \infty. \tag{33}$$

Proof of Theorem 3.1. The proof has three ingredients. First, even the opposite proposal favors the correct label on average. Second, this makes the expected weight on that proposal decrease geometrically. Third, the resulting total loss in log growth is finite. This verifies the general sample-cost condition with rate $G^\star$. A final growth argument gives the matching asymptotic lower bound.

For this proof, write

$$\ell = \log \frac{\mu}{1-\mu}, \qquad \eta = \mathbb{E}_{q^-}[\pi(X)].$$

The fixed matched proposals are mixed as $q_t = w_{t-1}q^+ + (1 - w_{t-1})q^-$.

Step 1: the opposite proposal still favors the correct label. Since $q^-(x) \propto p(x)e^{-\ell\pi(x)}$, proving $\eta > 1/2$ amounts to proving that the numerator below is positive:

$$\eta - \frac{1}{2} = \frac{\mathbb{E}_p[(\pi(X) - 1/2)e^{-\ell\pi(X)}]}{\mathbb{E}_p[e^{-\ell\pi(X)}]}.$$

For $z \in [0, 1]$,

$$(z - 1/2)e^{-\ell z} \geq -\frac{1}{2} + \frac{1 + e^{-\ell}}{2} z, \tag{34}$$

with equality only at $z = 0$ or $z = 1$. To verify it, for $z \leq 1/2$ use $(z - 1/2)e^{-\ell z} \geq z - 1/2$. Its difference from the right side is $(1 - e^{-\ell})z/2$. For $z \geq 1/2$, use $(z - 1/2)e^{-\ell z} \geq e^{-\ell}(z - 1/2)$; the corresponding difference is $(1 - e^{-\ell})(1 - z)/2$.

Taking expectations under p and using $e^{-\ell} = (1 - \mu)/\mu$ gives

$$\mathbb{E}_p[(\pi(X) - 1/2)e^{-\ell\pi(X)}] \geq -\frac{1}{2} + \frac{1 + e^{-\ell}}{2}\mu = 0.$$

At least one prompt has an interior preference probability and positive target mass, so the inequality is strict. This is where the interior-preference assumption is used.

The decreasing tilt also gives $\eta \leq \mu$. One direct verification is

$$2\,\mathrm{Cov}_p(\pi(X), e^{-\ell\pi(X)}) = \sum_{x,z} p(x)p(z)(\pi(x) - \pi(z))(e^{-\ell\pi(x)} - e^{-\ell\pi(z)}) \leq 0.$$

Using the increasing tilt instead gives $\mathbb{E}_{q^+}[\pi(X)] \geq \mu$. Consequently,

$$\mathbb{E}[Y_t \mid \mathcal{F}_{t-1}] = w_{t-1}\mathbb{E}_{q^+}[\pi(X)] + (1 - w_{t-1})\eta \geq \eta > 1/2. \tag{35}$$

The mixture therefore receives evidence favoring the correct direction even when it puts most of its weight on q^-.

Step 2: bound the weight on the wrong direction. Define the log evidence ratio

$$H_t = \log(E_t^+/E_t^-), \qquad H_0 = 0.$$

The same importance factor appears in both updates and cancels from the ratio. Thus

$$H_t - H_{t-1} = \ell(2Y_t - 1), \qquad 1 - w_t = \frac{1}{1 + e^{H_t}}. \tag{36}$$

Positive drift alone suggests that H_t increases; an exponential-moment calculation gives the quantitative bound we need. Choose

$$\theta = \frac{\log(\eta/(1-\eta))}{2\ell} \in (0, 1/2], \qquad \rho = 2\sqrt{\eta(1-\eta)} < 1.$$

The range of θ follows from $1/2 < \eta \le \mu < 1$. Equation 35 implies

$$\mathbb{E}[e^{-\theta(H_t - H_{t-1})} \mid \mathcal{F}_{t-1}] \le \eta e^{-\theta\ell} + (1-\eta)e^{\theta\ell} = \rho.$$

Iterating conditional expectations yields $\mathbb{E}[e^{-\theta H_t}] \le \rho^t$. For $0 \le \theta \le 1$ and every real h, $(1 + e^h)^{-1} \le e^{-\theta h}$. Hence

$$\mathbb{E}[1 - w_t] \le \rho^t, \qquad \sum_{t \ge 0} \mathbb{E}[1 - w_t] \le \frac{1}{1-\rho}. \tag{37}$$

Step 3: convert the weight bound into a growth bound. Let $D = D_{\mathrm{KL}}(q^- \| q^+)$. This is finite because both proposals have full support. By the directional-oracle identity and convexity of KL divergence in its first argument,

$$G_t^+(q_t) = G^\star - D_{\mathrm{KL}}(q_t \| q^+), \qquad D_{\mathrm{KL}}(q_t \| q^+) \le D(1 - w_{t-1}). \tag{38}$$

Define the total loss

$$R = \sum_{t \ge 1} D_{\mathrm{KL}}(q_t \| q^+), \qquad \mathbb{E}[R] \le \frac{D}{1-\rho} < \infty. \tag{39}$$

This proves Eq. 33, and verifies Eq. 24 with $G = G^\star$.

For completeness, all increment bounds required by the general theorem hold here. The normalized tilts satisfy $q^\pm \ge e^{-\ell} p$, so their mixture does too. The largest correct log increment is at most $\log(2\mu) + \ell$; a finite lower bound follows from $p(x) > 0$ and the finite population. Also, directional optimality gives $G^\star \ge D_{\mathrm{KL}}(\mu \| 1/2) > 0$. Applying the stopped-expectation argument with this upper increment gives the explicit bound

$$\mathbb{E}[\tau_\alpha] \le \frac{B + \log(2\mu) + \ell + D/(1-\rho)}{G^\star}. \tag{40}$$

The finite additive constant depends only on p and π. This calculation applies to the first crossing of either process, not just to the correct process: before that crossing $S_t < B$, and its final increment is still bounded if the other process triggers stopping.

Step 4: identify the limiting sample cost. The martingale in Eq. 28 gives

$$S_t = tG^\star - \sum_{u=1}^{t} D_{\mathrm{KL}}(q_u \| q^+) + M_t.$$

Since $\mathbb{E}[R] < \infty$, the nonnegative total loss R is finite almost surely. Moreover, bounded martingale increments and Azuma–Hoeffding give a summable bound on $\mathbb{P}(|M_t| > \varepsilon t)$ for each $\varepsilon > 0$. The Borel–Cantelli lemma then gives $M_t/t \to 0$ almost surely. Dividing the display by t proves $S_t/t \to G^\star$ almost surely.

We must also rule out an earlier crossing by the opposite process at large thresholds. Ville's inequality gives

$$\mathbb{P}\{\sup_{t \ge 0} E_t^- \ge u\} \le u^{-1}, \qquad u \ge 1.$$

Letting $u \to \infty$ shows that $\sup_t E_t^- < \infty$ almost surely. Therefore, on almost every infinite path, sufficiently large thresholds can be crossed only by E_t^+. For any $0 < \varepsilon < G^\star$, the growth limit eventually places S_t between $(G^\star - \varepsilon)t$ and $(G^\star + \varepsilon)t$. The finitely many earlier values cannot cross arbitrarily large thresholds. Comparing the hitting time with $B/(G^\star + \varepsilon)$ and $B/(G^\star - \varepsilon)$, then letting $\varepsilon \downarrow 0$, yields

$$\frac{\tau_\alpha}{B} \longrightarrow \frac{1}{G^\star} \quad \text{almost surely as } \alpha \downarrow 0.$$

Fatou's lemma gives the corresponding lower bound on the limiting expected ratio. Equation 40 supplies the matching upper bound. This proves Eq. 14 without assuming uniform integrability in advance. □

A.5 Concentration for the oracle mixture

Corollary A.1 (An upper bound on realized sample cost)**.** *Under the oracle assumptions of Theorem 3.1, there are finite constants C_1, C_2, C_3, depending only on (p, π), such that, for every $\alpha, \beta \in (0, 1)$,*

$$\mathbb{P}\left\{\tau_\alpha > \frac{B}{G^\star} + C_1\sqrt{B\log(2/\beta)} + C_2\log(2/\beta) + C_3\right\} \le \beta. \tag{41}$$

The main-text bound Eq. 15 follows by taking one constant at least as large as C_1, C_2, C_3 and the expectation constant.

The expected-cost proof controls the mean of the total growth loss. For a high-probability sample budget, we also need its tail. Geometric decay of the expected wrong-direction weight is enough because each round's loss is bounded. No independence is needed for this step.

Lemma A.3 (A sum with geometrically decaying means)**.** *Suppose $0 \le A_t \le C$ and $\mathbb{E}[A_t] \le CC_0\rho^t$ for $t \ge 0$, where $C \ge 0$, $C_0 \ge 1$, and $0 < \rho < 1$. Then*

$$\mathbb{E}\left[\sum_{t\ge 0} A_t\right] \le \frac{CC_0}{1-\rho} < \infty,$$

and, for every integer $m \ge 0$,

$$\mathbb{P}\left\{\sum_{t\ge 0} A_t > C(m+1)\right\} \le \frac{C_0\rho^m}{1-\rho}. \tag{42}$$

Proof. Summing the nonnegative expectations proves the first assertion and almost-sure finiteness. If $C = 0$, the sum is zero. Otherwise split it after its first m terms. Those terms contribute at most Cm, while the remaining terms have expectation at most $CC_0\rho^m/(1-\rho)$. For the full sum to exceed $C(m+1)$, this tail must exceed C. Markov's inequality gives the stated bound. □

Proof of Corollary A.1. Recall from Appendix A.4 that

$$D = D_{\mathrm{KL}}(q^-\|q^+), \qquad \rho = 2\sqrt{\eta(1-\eta)} < 1,$$

and that the total growth loss satisfies

$$R \le D\sum_{t\ge 0}(1-w_t), \qquad \mathbb{E}[1-w_t] \le \rho^t.$$

Apply Lemma A.3 to $A_t = D(1-w_t)$. For $\beta \in (0,1)$, choose

$$r = D\left[1 + \left\lceil \frac{\log\{2/((1-\rho)\beta)\}}{\log(1/\rho)} \right\rceil\right]. \tag{43}$$

Then $\mathbb{P}(R > r) \le \beta/2$. If $D = 0$, simply take $r = 0$.

The growth condition holds with $G = G^\star$, and the increment bound depends only on the fixed prompt population. Substituting this r into Eq. 30 gives an explicit sample budget. In particular,

r is at most a pair-dependent constant times $1 + \log(2/\beta)$. Expanding the square-root term with $\sqrt{x+y} \le \sqrt{x} + \sqrt{y}$ therefore gives

$$\frac{B}{G^\star} + C_1\sqrt{B\log(2/\beta)} + C_2\log(2/\beta) + C_3.$$

The constants absorb the bounded rounding term and are independent of α and β. This is Eq. 41. □

A.6 A CENTRAL LIMIT THEOREM FOR PAIRWISE SAMPLE COST

The preceding bound controls unusually large sample costs at a fixed threshold. The next theorem instead describes typical fluctuations as the threshold increases. The same finite total growth loss is now negligible relative to the square-root scale.

Theorem A.2 (Pairwise stopping-time fluctuations)**.** *Under the matched-oracle assumptions of Theorem 3.1, with $\mu > 1/2$, let*

$$v^\star = \left(\log\frac{\mu}{1-\mu}\right)^2 \mathbb{E}_{q^+}[\pi(X)(1-\pi(X))]. \tag{44}$$

Then $v^\star > 0$ and, as $\alpha \downarrow 0$,

$$\frac{\tau_\alpha - B/G^\star}{\sqrt{B}} \Rightarrow \mathcal{N}\left(0, \frac{v^\star}{(G^\star)^3}\right). \tag{45}$$

For $\mu < 1/2$, exchange labels and use the correct negative-direction proposal.

Proof. We first identify the limiting variance of log evidence. We then apply a martingale functional central limit theorem and translate the log-evidence limit into a hitting-time limit.

Step 1: identify the remaining randomness under the oracle. Under q^+, its normalizing tilt cancels the prompt-specific mean of the log factor:

$$\log\left\{\frac{p(X)}{q^+(X)}e^+(Y)\right\} = G^\star + \ell(Y - \pi(X)), \qquad \ell = \log\frac{\mu}{1-\mu}. \tag{46}$$

Indeed, $\log(p/q^+) = G^\star - h^+(X)$ and $\log e^+(Y) - h^+(X) = \ell(Y - \pi(X))$. Conditional on the prompt, the second term has mean zero and variance $\ell^2\pi(X)(1-\pi(X))$. Averaging over q^+ gives $v^\star$. It is positive because $\ell > 0$ and some prompt with positive q^+ mass has an interior preference probability.

For the actual mixture, the decomposition from Eq. 28 is

$$S_t = tG^\star - \sum_{u=1}^{t} D_{\mathrm{KL}}(q_u \| q^+) + M_t.$$

Appendix A.4 shows that the entire loss sum is bounded by an almost-surely finite R. It also shows $\sum_t \mathbb{E}[1 - w_t] < \infty$, which implies $w_t \to 1$ and hence $q_t \to q^+$ almost surely.

Step 2: obtain a functional limit for the martingale. The conditional second moment of a log increment is a finite sum of continuous functions of $q_t(x)$. The floor $q_t \ge e^{-\ell}p$ keeps all denominators away from zero. Therefore convergence to q^+ implies

$$\mathbb{E}[(M_t - M_{t-1})^2 \mid \mathcal{F}_{t-1}] \longrightarrow v^\star \quad \text{almost surely.}$$

Averaging these conditional variances gives, for every $u \ge 0$,

$$\frac{1}{T}\sum_{t=1}^{\lfloor Tu \rfloor} \mathbb{E}[(M_t - M_{t-1})^2 \mid \mathcal{F}_{t-1}] \longrightarrow uv^\star \quad \text{almost surely.}$$

This verifies convergence of the predictable quadratic variation. Since the martingale increments are uniformly bounded, the scaled increments satisfy the Lindeberg condition, as in the scaled martingale's maximum squared jump is $O(T^{-1})$, and the maximum jump of its predictable quadratic

variation is also $O(T^{-1})$. The martingale functional central limit theorem Billingsley (1999) therefore applies. Both bounds are deterministic and therefore also hold in expectation. The martingales are square-integrable. Consequently,

$$\left\{ \frac{M_{\lfloor Tu \rfloor}}{\sqrt{T}} : u \geq 0 \right\} \Rightarrow \{\sqrt{v^\star} W(u) : u \geq 0\}, \tag{47}$$

where W is standard Brownian motion.

Step 3: convert log-evidence fluctuations into time fluctuations. First consider only the correct-process hitting time

$$\tau_B^+ = \inf\{t \geq 1 : S_t \geq B\}.$$

The almost-sure growth limit proved in Appendix A.4 gives $\tau_B^+/B \to 1/G^\star$. Since this limit is deterministic, it holds jointly with the functional limit above. Evaluation at a converging time is continuous at continuous paths (Billingsley, 1999); hence

$$\frac{M_{\tau_B^+}}{\sqrt{B}} \Rightarrow \sqrt{v^\star} W(1/G^\star) \sim \mathcal{N}(0, v^\star/G^\star).$$

The overshoot is bounded: $0 \leq S_{\tau_B^+} - B \leq b$ for a finite increment bound b. Rearranging the decomposition of S_t at the hitting time gives

$$\begin{aligned} \frac{\tau_B^+ - B/G^\star}{\sqrt{B}} = & -\frac{M_{\tau_B^+}}{G^\star \sqrt{B}} \\ & + \frac{\sum_{u=1}^{\tau_B^+} D_{\mathrm{KL}}(q_u \| q^+) + S_{\tau_B^+} - B}{G^\star \sqrt{B}}. \end{aligned}$$

The second line tends to zero almost surely: its numerator is at most $R + b$. The first line converges to a centered normal with variance $v^\star/(G^\star)^3$, proving the desired limit for τ_B^+.

Step 4: return to the implemented two-process rule. The two stopping times can differ only if the negative process also reaches e^B. Since it is a null supermartingale, Ville's inequality (Ville, 1939) gives

$$\mathbb{P}\{\tau_\alpha \neq \tau_B^+\} \leq \mathbb{P}\{\sup_t E_t^- \geq e^B\} \leq e^{-B} = \alpha.$$

This probability vanishes, so the implemented stopping time has the same weak limit. □

The variance has two sources of scaling. About $B/G^\star$ judgments contribute martingale variance, and converting log-evidence error to time error divides by $G^\star$. This explains the denominator $(G^\star)^3$. A corresponding theorem for learned proposals would additionally need convergence of their conditional variances and growth errors negligible relative to $\sqrt{B}$; error validity alone does not imply such a limit.

A.7 The gain from prompt heterogeneity

Under target sampling, the matched bet gives log-growth rate $D_{\mathrm{KL}}(\mu \| 1/2)$. The oracle rate decomposes as

$$G^\star = D_{\mathrm{KL}}(\mu \| 1/2) + \log \mathbb{E}_{X \sim p}\left[\left(\frac{\mu}{1-\mu} \right)^{\pi(X) - \mu} \right]. \tag{48}$$

The second term is nonnegative by Jensen's inequality and is strictly positive when preference probabilities vary across prompts. Let $\tau_{p,\alpha}$ denote the stopping time under $q_t = p$, using the same matched bet and two-process stopping rule.

Corollary A.2 (Gain from prompt heterogeneity)**.** *Under the oracle assumptions,*

$$\lim_{\alpha \downarrow 0} \frac{\mathbb{E}[\tau_\alpha]}{\mathbb{E}[\tau_{p,\alpha}]} = \frac{D_{\mathrm{KL}}(\mu \| 1/2)}{G^\star} \leq 1, \tag{49}$$

with strict inequality exactly when $\pi(X)$ is nonconstant under p.

The two-prompt population with preferences 0.9 and 0.6 therefore admits a strict asymptotic gain. If both probabilities were 0.75, the target win rate would be unchanged, but both directional proposals would equal p and the gain would disappear.

Proof of Corollary A.2. We first identify the growth gain and then compare the two stopping costs.

Step 1: separate the target rate from the gain. At the matched bet, the positive local e-power is

$$h^+(x) = \log(2(1-\mu)) + \ell\pi(x) = D_{\mathrm{KL}}(\mu\|1/2) + \ell(\pi(x) - \mu).$$

Taking its exponential moment under p gives

$$G^\star = D_{\mathrm{KL}}(\mu\|1/2) + \log \mathbb{E}_p[e^{\ell(\pi(X)-\mu)}],$$

which is Eq. 48. The centered preference has mean zero. Jensen's inequality makes the second term nonnegative, and strict convexity makes it positive exactly when $\pi(X)$ is nonconstant under p.

Step 2: evaluate the target-sampling benchmark. Under $q_t = p$, the correct log increment has constant conditional mean $D_{\mathrm{KL}}(\mu\|1/2) > 0$ and upper bound $\log(2\mu)$. The bounded-stopping argument therefore gives

$$\mathbb{E}[\tau_{p,\alpha}] \le \frac{B + \log(2\mu)}{D_{\mathrm{KL}}(\mu\|1/2)}.$$

As in Step 4 of Appendix A.4, bounded martingale fluctuations give the almost-sure growth rate, and the wrong-direction supermartingale is bounded over the infinite path almost surely. Thus

$$\frac{\tau_{p,\alpha}}{B} \longrightarrow \frac{1}{D_{\mathrm{KL}}(\mu\|1/2)} \quad \text{almost surely.}$$

Fatou's lemma and the displayed expectation bound give the same limit for the expected ratio. Combining this with Theorem 3.1 proves Eq. 49 and its strictness condition. □

A.8 Prediction error and proposal loss

We now verify the general growth condition for LLM-guided sampling. Let $\ell_t = \log(\nu_t/(1-\nu_t))$. If predictions on the full population satisfy $\sup_x |\widehat{\pi}_{t-1}(x) - \pi(x)| \le \epsilon_t$, then

$$D_{\mathrm{KL}}(\widehat{q}_t^{\,s}\|q_t^{s,\star}) \le \frac{\ell_t^2\epsilon_t^2}{2}, \qquad s \in \{+,-\}. \tag{50}$$

Thus prediction error incurs a quadratic loss in the growth of the estimated directional proposal. The proof below derives this bound from Hoeffding's exponential-moment inequality (Hoeffding, 1963). For the final mixture, convexity gives

$$\begin{aligned} D_{\mathrm{KL}}(q_t\|q_t^{+,\star}) \le{}& \kappa D_{\mathrm{KL}}(p\|q_t^{+,\star}) \\ &+ (1-\kappa)\left\{ w_{t-1}\frac{\ell_t^2\epsilon_t^2}{2} + (1-w_{t-1})D_{\mathrm{KL}}(\widehat{q_t^-}\|q_t^{+,\star}) \right\}. \end{aligned} \tag{51}$$

The terms have distinct roles: exploration, prediction error in the correct component, and weight assigned to the opposite component. To turn this inequality into a sample bound, we must also show that the last weight decreases.

Shortlisting prompts. Prediction accuracy on a shortlist is not sufficient if the shortlist omits useful prompts. For a nonempty shortlist C_t, let $\widehat{q}_{t,C_t}^{\,s}$ be the directional proposal normalized on that set. If prediction error on C_t is at most ϵ_t, then

$$D_{\mathrm{KL}}(\widehat{q}_{t,C_t}^{\,s}\|q_t^{s,\star}) \le \frac{\ell_t^2\epsilon_t^2}{2} + \log\frac{1}{q_t^{s,\star}(C_t)}. \tag{52}$$

The additional term is small when the shortlist retains most of the oracle proposal's mass. The proof below gives this penalty; Appendix A.10 converts it into a sample bound under coverage and directional-learning conditions. For adaptive bets and shortlists, Theorem A.1 applies whenever the final proposal satisfies the cumulative-growth condition.

The estimated proposal is an exponential perturbation of the true proposal. Bounding this perturbation serves two different purposes: a KL bound controls evidence growth, while a label-mean bound ensures that the mixture still learns the correct direction.

Lemma A.4 (Perturbing a directional proposal). *Fix $\nu \in [1/2, 1)$, and put $\ell = \log(\nu/(1-\nu))$. Let q^s and $\widehat{q}^s$ be the true and estimated directional proposals at that bet. If $\sup_x |\widehat{\pi}(x) - \pi(x)| \le \epsilon$, then*

$$D_{\mathrm{KL}}(\widehat{q}^s \| q^s) \le \frac{\ell^2 \epsilon^2}{2}, \qquad |\mathbb{E}_{\widehat{q}^s}[\pi(X)] - \mathbb{E}_{q^s}[\pi(X)]| \le \frac{\ell \epsilon}{2}. \tag{53}$$

If the estimated proposal is normalized on a nonempty shortlist C, the same bounds hold relative to $q^s(\cdot \mid C)$ when the error bound holds on C. In particular,

$$D_{\mathrm{KL}}(\widehat{q}^s_C \| q^s) \le \frac{\ell^2 \epsilon^2}{2} + \log \frac{1}{q^s(C)}. \tag{54}$$

The statements also hold conditionally for predictable bets, predictions, and shortlists.

Proof. We first control the exponential perturbation of the proposal, then its effect on growth and on the mean label. The shortlist result follows by conditioning the reference proposal.

Step 1: express prediction error as an exponential tilt. For the positive direction let $d(x) = \ell(\widehat{\pi}(x) - \pi(x))$; use its negative for the other direction. Then

$$\widehat{q}^s(x) = \frac{q^s(x) e^{d(x)}}{\mathbb{E}_{q^s}[e^{d(X)}]}.$$

The range of d has length at most $2\ell\epsilon$. The normalized relation also implies $\mathbb{E}_{\widehat{q}^s}[e^{-d(X)}] = 1/\mathbb{E}_{q^s}[e^{d(X)}]$.

Step 2: bound the KL loss. Using the last identity,

$$\begin{aligned} D_{\mathrm{KL}}(\widehat{q}^s \| q^s) &= \mathbb{E}_{\widehat{q}^s}[d(X)] + \log \mathbb{E}_{\widehat{q}^s}[e^{-d(X)}] \\ &\le \frac{(\sup_x d(x) - \inf_x d(x))^2}{8} \le \frac{\ell^2 \epsilon^2}{2}. \end{aligned}$$

The first inequality is Hoeffding's bounded-variable exponential-moment inequality (Hoeffding, 1963). It applies under $\widehat{q}^s$ to the centered variable $-d(X)$. Only the range matters; a constant additive prediction error cancels in the normalization.

Step 3: bound the change in the label mean. Interpolate between the proposals using $q_u(x) \propto q^s(x) e^{u d(x)}$, $0 \le u \le 1$. Differentiating the finite normalized sum gives

$$\frac{d}{du} \mathbb{E}_{q_u}[\pi(X)] = \mathrm{Cov}_{q_u}(\pi(X), d(X)).$$

A variable supported on an interval of length a has variance at most $a^2/4$: its squared distance from the interval midpoint is at most that value, and variance minimizes expected squared distance over all centers. Thus $\mathrm{Var}_{q_u}(\pi(X)) \le 1/4$ and $\mathrm{Var}_{q_u}(d(X)) \le \ell^2 \epsilon^2$. Cauchy–Schwarz bounds the absolute covariance by $\ell\epsilon/2$. Integrating from $u = 0$ to $u = 1$ proves the mean bound.

Step 4: separate prediction error from shortlist coverage. Repeat the preceding argument with reference distribution $q^s(\cdot \mid C)$. Since the estimated shortlist proposal is supported on C,

$$D_{\mathrm{KL}}(\widehat{q}^s_C \| q^s) = D_{\mathrm{KL}}(\widehat{q}^s_C \| q^s(\cdot \mid C)) + \log \frac{1}{q^s(C)}.$$

The first term is the prediction penalty on the retained prompts; the second is the exact cost of restricting the oracle's support. This proves Eq. 54. All calculations condition on choices already made before the draw, so they remain valid for predictable adaptive choices. □

A.9 Sample cost with predicted preferences and a learned bet

For an explicit sufficient condition, suppose $\mu \in (1/2, 1)$ and the full-population prediction error is at most ϵ almost surely at every round. At the matched bet, write

$$\ell = \log \frac{\mu}{1-\mu}, \qquad \eta = \mathbb{E}_{q^-}[\pi(X)].$$

Assume

$$\eta - \frac{\ell\epsilon}{2} > \frac{1}{2}, \qquad G_{\mathrm{LLM}} = \kappa D_{\mathrm{KL}}(\mu\|1/2) + (1-\kappa)\left(G^\star - \frac{\ell^2\epsilon^2}{2}\right) > 0. \tag{55}$$

The first condition keeps both estimated components on the correct side of mean label $1/2$, so the evidence weights can learn the direction. The second leaves a positive growth rate after the prediction penalty. These are sufficient conditions; their failure alone does not imply that the procedure is inefficient. In particular, a bound on average prediction error is not the uniform-in-prompt, every-round bound assumed here.

Corollary A.3 (LLM-guided sampling)**.** *Under these conditions, use the estimated mixture in Eq. 10. For a fixed matched bet and $\kappa \in [0,1]$, there is a finite C, independent of α, such that*

$$\mathbb{E}[\tau_\alpha] \le \frac{B}{G_{\mathrm{LLM}}} + C. \tag{56}$$

An upper-tail bound of the form in Eq. 41 also holds with leading rate G_{LLM}.

For the plug-in bet in Eq. 12, suppose $\kappa > 0$ and $\bar\nu \ge \mu$. For every fixed $0 < G < G_{\mathrm{LLM}}$, the same expected-cost and upper-tail bounds hold with leading rate G and finite constants that may depend on G. In particular,

$$\limsup_{\alpha\downarrow 0} \frac{\mathbb{E}[\tau_\alpha]}{B} \le \frac{1}{G_{\mathrm{LLM}}}. \tag{57}$$

The proof below first controls the wrong-direction weight and then the cost of learning the bet. The rate is a lower bound, not an exact rate or a guarantee of any particular LLM's prediction accuracy. These assumptions affect efficiency only; Lemma 3.1 continues to hold without them.

Proof of Corollary A.3. The proof verifies the same cumulative-growth condition as in the oracle analysis. For a fixed bet, prediction error reduces the useful growth rate, while learning the direction contributes a finite total deficit. For a learned bet, we additionally show that rounds with a poor bet have geometrically decreasing probabilities. The predictor itself may change arbitrarily with the history, as long as the assumed error bound holds.

Step 1: predicted proposals still reveal the correct direction. First fix $\nu_t = \mu$. The matched true proposals satisfy $\mathbb{E}_{q^+}[\pi(X)] \ge \mathbb{E}_{q^-}[\pi(X)] = \eta$. By the label-mean bound in Lemma A.4, each estimated component has label mean at least $\eta - \ell\epsilon/2$. Target sampling has mean $\mu \ge \eta$. Thus

$$\mathbb{E}[Y_t \mid \mathcal{F}_{t-1}] \ge \eta - \ell\epsilon/2 > 1/2 \tag{58}$$

under the final proposal, including exploration.

As in Eq. 36, $H_t = \log(E_t^+/E_t^-)$ changes by $\ell(2Y_t - 1)$. Choose a sufficiently small $\theta \in (0,1]$ such that

$$\rho = (\eta - \ell\epsilon/2)e^{-\theta\ell} + (1 - \eta + \ell\epsilon/2)e^{\theta\ell} < 1.$$

This is possible because the expression equals one at $\theta = 0$ and has negative derivative there. The conditional exponential-moment argument from Appendix A.4 gives

$$\mathbb{E}[1 - w_t] \le \mathbb{E}[e^{-\theta H_t}] \le \rho^t. \tag{59}$$

This step explains the first condition in Eq. 55: it prevents prediction error from reversing the direction learned by the evidence ratio.

Step 2: identify the fixed-bet growth rate. For any distribution q on the population,

$$D_{\mathrm{KL}}(q\|q^+) \le \log \frac{1}{\min_x q^+(x)}.$$

Indeed, $\sum_x q(x)\log q(x) \le 0$, and the remaining cross-entropy is bounded by the displayed maximum. Denote this finite bound by C. Using the oracle identity, convexity, and the prediction-error bound,

$$\begin{aligned} G_t^+(q_t) &\ge G^\star - \kappa D_{\mathrm{KL}}(p\|q^+) \\ &\quad - (1-\kappa)\left\{ w_{t-1}\frac{\ell^2\epsilon^2}{2} + (1 - w_{t-1})C \right\} \\ &\ge G_{\mathrm{LLM}} - C(1 - w_{t-1}). \end{aligned}$$

Here $G^\star - D_{\mathrm{KL}}(p\|q^+) = D_{\mathrm{KL}}(\mu\|1/2)$, which accounts for the target-sampling part of G_{LLM}. Thus the cumulative-growth condition holds with rate G_{LLM} and deficit

$$R = C\sum_{t\geq 0}(1-w_t).$$

By Eq. 59 and Lemma A.3, this deficit has finite expectation and an exponential tail. Since predictions lie in $[0,1]$, full-population estimated tilts satisfy $\widehat{q}_t^s \geq e^{-\ell}p$. The final proposal has floor $\{\kappa + (1-\kappa)e^{-\ell}\}p$, which is positive even when $\kappa = 0$. Therefore the log increments are bounded and Theorem A.1 applies. Its expectation bound gives Eq. 56; its upper-quantile formula and the exponential deficit tail give the stated concentration bound.

Step 3: a learned bet is eventually close to the matched bet. Now use Eq. 12, with $\kappa > 0$ and $\bar{\nu} \geq \mu$. Fix a desired rate $0 < G < G_{\mathrm{LLM}}$. We will work in a small neighborhood of the matched bet, where both useful growth and directional drift remain positive.

Use a subscript ν for a true proposal or growth rate evaluated at a fixed bet ν, and write $\ell_\nu = \log(\nu/(1-\nu))$. The two quantities

$$\begin{aligned}&\kappa G_\nu^+(p) + (1-\kappa)\{G_\nu^\star - \ell_\nu^2\epsilon^2/2\},\\ &\min\{\mu, \mathbb{E}_{q_\nu^-}[\pi(X)] - \ell_\nu\epsilon/2\}\end{aligned}$$

are continuous in $\nu \in [1/2, \bar{\nu}]$. At $\nu = \mu$, the first equals $G_{\mathrm{LLM}} > G$ and the second exceeds $1/2$. Choose $\epsilon_0 > 0$ small enough that these inequalities persist whenever $|\nu - \mu| \leq \epsilon_0$, and take $\epsilon_0 < (\mu - 1/2)/2$. In this neighborhood, ℓ_ν is also bounded away from zero.

All true positive proposals with $\nu \leq \bar{\nu}$ have a common positive floor. Hence their KL losses against arbitrary distributions admit a uniform finite bound C, as in Step 2. Continuity and the mean-perturbation bound now give a constant $c > 0$ such that, on every round with $|\nu_t - \mu| \leq \epsilon_0$,

$$\begin{aligned}G_t^+(q_t) &\geq G - C(1-w_{t-1}),\\ \mathbb{E}[H_t - H_{t-1} \mid \mathcal{F}_{t-1}] &\geq c.\end{aligned} \tag{60}$$

For the second inequality we used $H_t - H_{t-1} = \ell_t(2Y_t - 1)$: both the log odds and the conditional label advantage are uniformly positive in this neighborhood.

It remains to bound the probability of being outside the neighborhood. Each importance-weighted observation $p(X_t)Y_t/q_t(X_t)$ has conditional mean μ and lies in $[0, 1/\kappa]$. Applying Hoeffding's exponential-moment bound conditionally and iterating over the observations yields the usual bounded-martingale concentration inequality (Hoeffding, 1963; Azuma, 1967). Clipping the average to $[0,1]$, applying $z \mapsto \max\{z, 1-z\}$, and capping at $\bar{\nu} \geq \mu$ are all 1-Lipschitz operations; their composition maps μ to itself. Therefore

$$\mathbb{P}\{|\nu_t - \mu| > \epsilon_0\} \leq 2\exp\{-2\kappa^2\epsilon_0^2(t-1)\}, \qquad t \geq 2. \tag{61}$$

Thus poor-bet rounds become geometrically unlikely. This is where fixed positive exploration is used to learn the bet: it uniformly bounds the importance-weighted observations.

Step 4: changing bets still lead to a geometric weight bound. Let $\bar{\ell} = \log(\bar{\nu}/(1-\bar{\nu}))$. Every ratio increment has absolute value at most $\bar{\ell}$. For a large integer t, set

$$m = \left\lfloor \frac{ct}{4(c+\bar{\ell})} \right\rfloor.$$

This reserves a short initial segment for potentially poor behavior. Regardless of what happens in that segment, $H_m \geq -\bar{\ell}m$. If all bets in rounds $m+1, \ldots, t$ are in the neighborhood, Eq. 60 gives

$$\begin{aligned}H_t &\geq ct - (c+\bar{\ell})m\\ &\quad + \sum_{u=m+1}^{t}\{H_u - H_{u-1} - \mathbb{E}[H_u - H_{u-1} \mid \mathcal{F}_{u-1}]\}.\end{aligned}$$

The first line is at least $3ct/4$. The centered increments in the second line have absolute value at most $2\bar{\ell}$. Consequently, the event $H_t < ct/2$ requires either a poor bet in the final segment or a

martingale fluctuation below $-ct/4$. A union bound and Azuma–Hoeffding yield

$$\mathbb{P}\{H_t < ct/2\} \le \sum_{u=m+1}^{t} \mathbb{P}\{|\nu_u - \mu| > \epsilon_0\} + \exp\{-c^2 t/(128\bar{\ell}^2)\}.$$

The first term decreases geometrically by Eq. 61, because m grows linearly with t. So does the second. This argument uses event inclusion, rather than conditioning the martingale on a favorable future event.

On $\{H_t \ge ct/2\}$, we also have $1 - w_t \le e^{-ct/2}$. Combining the two cases and absorbing finitely many early rounds gives finite $C_0 \ge 1$ and $0 < \rho < 1$ such that

$$\mathbb{E}[1 - w_t] \le C_0 \rho^t, \qquad t \ge 0. \tag{62}$$

Step 5: include both learning costs in one deficit. All log increments are bounded by some b, so $G_t^+(q_t) \ge -b$ even outside the neighborhood. Increase the constant C in Eq. 60, if necessary, so that $C \ge G + b$. Then every round satisfies

$$G_t^+(q_t) \ge G - C\{(1 - w_{t-1}) + \mathbf{1}_{\{|\nu_t - \mu| > \epsilon_0\}}\}.$$

Thus a valid total deficit is

$$R = C \sum_{t \ge 1} \{(1 - w_{t-1}) + \mathbf{1}_{\{|\nu_t - \mu| > \epsilon_0\}}\}.$$

Its nonnegative summands are bounded, and their expectations decrease geometrically by Eqs. 61 and Eq. 62. Lemma A.3 gives finite expectation and an exponential tail. Theorem A.1 now proves both sample-cost bounds at the chosen rate G.

Finally, for every fixed $G < G_{\mathrm{LLM}}$, the additive constants disappear after dividing by B and taking $\limsup_{\alpha \downarrow 0}$. Letting $G \uparrow G_{\mathrm{LLM}}$ afterward proves Eq. 57. This order of limits is important: the finite-threshold constants may grow as G approaches G_{LLM}. □

A.10 Sample cost with shortlisted proposals

A shortlist must satisfy two distinct requirements. It must retain enough mass of the correct oracle proposal to preserve useful growth, and its labels must still reveal the winning direction. The following proposition makes both requirements explicit.

Proposition A.2 (Shortlist coverage and directional learning)**.** *Suppose $\mu \in (1/2, 1)$, use the fixed matched bet, and let $\kappa \in (0, 1]$. Each round selects a predictable nonempty shortlist C_t, forms both estimated directional proposals on C_t, and uses Eq. 10. Suppose, almost surely at every round,*

$$\sup_{x \in C_t} |\widehat{\pi}_{t-1}(x) - \pi(x)| \le \epsilon, \qquad q^+(C_t) \ge r_0 > 0,$$

and, for both $s \in \{+, -\}$,

$$\mathbb{E}_{q^s(\cdot|C_t)}[\pi(X)] \ge \eta_C, \qquad \eta_C - \ell\epsilon/2 > 1/2, \qquad \ell = \log \frac{\mu}{1 - \mu}.$$

If

$$G_C = \kappa D_{\mathrm{KL}}(\mu \| 1/2) + (1 - \kappa)\{G^\star - \ell^2\epsilon^2/2 - \log(1/r_0)\} > 0,$$

then $\mathbb{E}[\tau_\alpha] \le B/G_C + C$ for a finite C independent of α. An upper-tail bound of the form in Eq. 41 holds with leading rate G_C.

Proof. The mean bound in Lemma A.4, applied on C_t, shows that each estimated component has label mean at least $\eta_C - \ell\epsilon/2 > 1/2$. After adding exploration, the mean is still at least the smaller of this quantity and μ, hence uniformly above $1/2$. The ratio argument in Step 1 of Appendix A.9 therefore gives geometrically decreasing $\mathbb{E}[1 - w_t]$.

Next, the prediction and coverage assumptions give

$$D_{\mathrm{KL}}(\widehat{q}^+_{t,C_t} \| q^+) \le \ell^2\epsilon^2/2 + \log(1/r_0).$$

The negative component has divergence at most $\log(1/\min_x q^+(x))$, independently of which shortlist is chosen. The convexity calculation from Step 2 of Appendix A.9 now gives

$$G_t^+(q_t) \geq G_C - C(1 - w_{t-1})$$

for a finite C. Thus the total deficit is bounded by $C\sum_{t\geq 0}(1 - w_t)$, with finite mean and an exponential tail. Exploration gives the full-support floor κp and bounded log increments. Theorem A.1 proves both conclusions. □

These are assumptions on the selected prompts, not consequences of asking an LLM for a top-k list. Prediction accuracy on retained prompts does not imply coverage, and coverage alone does not imply the directional-mean condition. For changing bets, the general theorem remains available, but this fixed-bet proposition does not by itself verify the required adaptive growth condition.

A.11 Other fixed bets

The matched-oracle proof uses two properties of the bet: positive growth of the correct oracle and a label mean above $1/2$ under the opposite proposal. These properties, rather than exact matching itself, are enough for the argument.

Fix $\nu \in (1/2, 1)$, let q_ν^+ and q_ν^- be its true directional proposals, and let $G_\nu^\star$ be the positive-direction oracle growth rate. Use their evidence-weighted mixture with this fixed bet.

Proposition A.3 (A fixed-bet mixture)**.** *Suppose $\mu > 1/2$, $\mathbb{E}_{q_\nu^-}[\pi(X)] > 1/2$, and $G_\nu^\star > 0$. Then a finite constant $C = C(p, \pi, \nu)$ satisfies*

$$\mathbb{E}[\tau_\alpha] \leq \frac{B}{G_\nu^\star} + C, \qquad \mathbb{E}[\tau_\alpha] \sim \frac{B}{G_\nu^\star} \quad \text{as } \alpha \downarrow 0.$$

Both growth conditions hold whenever $1/2 < \nu < \mu$, including for deterministic prompt preferences.

Proof. **Step 1: apply the mixture argument at the chosen bet.** Write $\ell = \log(\nu/(1-\nu))$ and $\eta = \mathbb{E}_{q_\nu^-}[\pi(X)]$. Monotone tilting gives $\mathbb{E}_{q_\nu^+}[\pi(X)] \geq \eta$, so every mixture has label mean at least $\eta > 1/2$. Choose $\theta \in (0, 1]$ small enough that

$$\rho = \eta e^{-\theta\ell} + (1 - \eta)e^{\theta\ell} < 1.$$

The derivative at zero is negative, so such a choice exists, including when $\eta = 1$. As before, $\mathbb{E}[1 - w_t] \leq \rho^t$, and the total expected KL loss is at most $D_{\mathrm{KL}}(q_\nu^- \| q_\nu^+)/(1 - \rho)$.

The component floor is $e^{-\ell}p$, and the correct log increment is at most $\log(2\nu) + \ell$. The stopped-expectation argument therefore gives

$$\mathbb{E}[\tau_\alpha] \leq \frac{B + \log(2\nu) + \ell + D_{\mathrm{KL}}(q_\nu^- \| q_\nu^+)/(1 - \rho)}{G_\nu^\star}. \tag{63}$$

Finite total loss, bounded martingale fluctuations, and the wrong-direction supermartingale give the almost-sure hitting-time limit exactly as in Step 4 of Appendix A.4. Fatou's lemma and this upper bound then give convergence of expected costs.

Step 2: verify the conditions for a conservative bet. Suppose $1/2 < \nu < \mu$. The pointwise inequality in Eq. 34 applies with the present ℓ, and gives

$$\mathbb{E}_p[(\pi(X) - 1/2)e^{-\ell\pi(X)}] \geq -\frac{1}{2} + \frac{1 + e^{-\ell}}{2}\mu = \frac{\mu - \nu}{2\nu} > 0.$$

Dividing by the positive normalizing constant proves $\mathbb{E}_{q_\nu^-}[\pi(X)] > 1/2$, without any interior-preference assumption.

Finally, Jensen's inequality yields

$$G_\nu^\star \geq \mu \log(2\nu) + (1 - \mu)\log(2(1 - \nu)).$$

The right side is zero at $\nu = 1/2$ and has derivative $(\mu - \nu)/(\nu(1 - \nu)) > 0$ on $(1/2, \mu)$. It is therefore positive, proving the other growth condition. □

A known lower bound $\underline{\mu} > 1/2$ on μ permits choosing $\nu \in (1/2, \underline{\mu})$. A point estimate alone does not certify that the bet is conservative.

A.12 A DETERMINISTIC-LABEL BOUNDARY UNDER MATCHED BETTING

The interior-preference assumption in the matched-oracle theorem cannot simply be removed. Consider two prompts with

$$p = (3/4, 1/4), \qquad \pi = (1, 0), \qquad \mu = \nu = 3/4.$$

Then $q^+ = (9/10, 1/10)$ and $q^- = (1/2, 1/2)$, so the opposite proposal produces the correct label only half the time. We show that $\mathbb{E}[\tau_\alpha] = \infty$ for every $\alpha < 4/5$.

Step 1: a wrong initial label forces an excursion. Let $J_t = \log_3(E_t^+/E_t^-)$. It starts at zero and moves by $+1$ or -1 according to the label. At state j, its upward transition probability is

$$\frac{1}{2} + \frac{2}{5}\frac{3^j}{1+3^j}.$$

The first label is zero with probability $3/10$. On this event,

$$J_1 = -1, \qquad E_1^+ = 5/12, \qquad E_1^- = 5/4.$$

While $J_t \le -1$, the weight on q^+ is at most $1/4$, so the proposal probability of label one lies in $[1/2, 3/5]$. A label-one update multiplies E_t^- by at most $(3/8)/(1/2) = 3/4$. A label-zero update multiplies it by at most $(3/8)/(2/5) = 15/16$. Thus E_t^- never rises above $5/4$ during this excursion. The positive wealth is smaller while $J_t < 0$, and the two are equal upon returning to zero. For a threshold exceeding $5/4$, neither process can stop before that return.

Step 2: the return time has infinite mean. Conditional on the first label being zero, let $Z_s = -J_{1+s}$ until it hits zero, so $Z_0 = 1$. At state $k \ge 1$, the down and up probabilities are

$$d_k = \frac{1}{2} + \frac{2}{5}\frac{3^{-k}}{1+3^{-k}}, \qquad b_k = 1 - d_k.$$

Write $T_n = \inf\{s \ge 0 : Z_s = n\}$, and use $\mathbb{P}_1$ for the law starting at one. For fixed n, let $h_k = \mathbb{P}_k(T_n < T_0)$. The nearest-neighbor recurrence is

$$h_k = d_k h_{k-1} + b_k h_{k+1}, \qquad h_0 = 0, \quad h_n = 1.$$

Rearranging gives $h_{k+1} - h_k = (d_k/b_k)(h_k - h_{k-1})$. Iterating these increments and summing from zero to n gives

$$\mathbb{P}_1(T_n < T_0) = \left(\sum_{j=1}^{n}\prod_{k=1}^{j-1}\frac{d_k}{b_k}\right)^{-1}.$$

Here $d_k/b_k - 1 = O(3^{-k})$, so the products are bounded above by a finite constant C, uniformly in j. For example, their logarithms are bounded by the convergent sum $\sum_k (d_k/b_k - 1)$. Consequently,

$$\mathbb{P}_1(T_n < T_0) \ge \frac{1}{Cn}.$$

Reaching level n from one takes at least $n-1$ steps. Hence $\mathbb{P}_1(T_0 \ge n-1) \ge 1/(Cn)$, and summing these tail probabilities shows $\mathbb{E}_1[T_0] = \infty$. The initial wrong-label event has positive probability, so Step 1 implies $\mathbb{E}[\tau_\alpha] = \infty$ whenever $1/\alpha > 5/4$.

This is a failure of finite expected stopping cost, not of directional error validity. Proposition A.3 restores finite expected cost on the same instance for every fixed $\nu \in (1/2, 3/4)$.

A.13 SAMPLING A FINITE ARCHIVE WITHOUT REPLACEMENT

Removing a queried prompt changes the sampling support. It does not, by itself, invalidate the nonnegative multiplicative factors in Eq. 4. The crucial distinction is between keeping the original target weight $p(x)$ and renormalizing that target over the remaining prompts. The following calculation covers the former. It does not assert that an uninspected implementation uses the required weights.

Proposition A.4 (Validity with a restricted sampling support)**.** *At each round before exhaustion, let $S_t \subseteq \mathcal{X}$ be a predictable nonempty set and let q_t be positive on S_t and zero outside it. Suppose Eq. 3 holds for every sampled prompt. Use predictable bets in $[1/2, 1)$ and the original target weight $p(X_t)$ in Eq. 4. Freeze the processes when no prompts remain. Then both processes are nonnegative supermartingales under their respective nulls, and the declaration-error conclusion of Lemma 3.1 still holds.*

Proof. For the positive direction, conditional on the past,

$$\begin{aligned}\mathbb{E}\left[\frac{p(X_t)}{q_t(X_t)}e_t^+(Y_t) \mid \mathcal{F}_{t-1}\right] &= \sum_{x\in S_t} p(x)\{2\nu_t\pi(x) + 2(1-\nu_t)(1-\pi(x))\} \\ &\le \sum_{x\in\mathcal{X}} p(x)\{2\nu_t\pi(x) + 2(1-\nu_t)(1-\pi(x))\} \\ &= 1 + 4(\nu_t - 1/2)(\mu - 1/2) \le 1\end{aligned}$$

under $\mu \le 1/2$. The inequality holds because every omitted summand is nonnegative. The same calculation with the negative factor yields $1 - 4(\nu_t - 1/2)(\mu - 1/2) \le 1$ under $\mu \ge 1/2$. Freezing after exhaustion preserves both properties. Ville's inequality completes the proof (Ville, 1939; Ramdas & Wang, 2025). □

The target of archived-data replay. Condition on a finite archive with fixed labels $y(x) \in \{0,1\}$ and let $\pi(x) = y(x)$. Revealing labels by any adaptive without-replacement proposal satisfies the proposition for the fixed-record mean $\sum_x p(x)y(x)$; randomness is in the sampling procedure. This is an exact statement about the archive, not inference about an unknown population of fresh human judgments. Under a latent-preference interpretation the stipulated conditional label law is instead required. If all archive labels have been revealed, their weighted mean is known exactly, so an exact terminal decision about that archive adds no error. Forcing a winner before full exhaustion without a valid crossing is not covered.

Why the earlier cost rates do not transfer. For q_t supported on S_t, the log-growth identity remains

$$G_t^s(q_t) = G_t^{s,\star} - D_{\mathrm{KL}}(q_t \| q_t^{s,\star}),$$

with $0 \log 0 = 0$. Maximizing over distributions supported on S_t instead gives

$$\max_{q:\, q(S_t)=1} G_t^s(q) = G_t^{s,\star} + \log q_t^{s,\star}(S_t).$$

The extra term is nonpositive and grows more costly as useful prompts are removed. This is the same omitted-mass mechanism as the shortlist penalty. Also,

$$\mathbb{E}\left[\frac{p(X_t)}{q_t(X_t)}Y_t \mid \mathcal{F}_{t-1}\right] = \sum_{x\in S_t} p(x)\pi(x),$$

not μ. Consequently, the plug-in estimator's convergence proof and the infinite-horizon oracle cost and CLT arguments cannot be reused unchanged. Its predictable capped bets still preserve validity. A finite archive can be exhausted without a threshold crossing, and a fixed finite horizon does not support the same vanishing-error asymptotics.

Renormalizing the target is different. Replacing $p(x)$ by $p(x)/p(S_t)$ in the multiplier divides the conditional expectation by $p(S_t)$, which can exceed the null bound. For example, take two equally weighted deterministic labels $0, 1$, a bet $\nu_t = 3/4$, and suppose label 0 was sampled first. The original mean is $1/2$. On the remaining singleton, renormalized target weighting gives a positive factor $3/2 > 1$; the required supermartingale inequality fails. Keeping the original target weight instead gives $3/4 \le 1$. This example shows why the exact weighting convention must be checked, rather than treating all without-replacement implementations alike.

Finite-population confidence sequences can exploit sampling without replacement more directly; see Waudby-Smith & Ramdas (2020). The proposition above is only a support-restriction result for this paper's factors, not an efficiency claim for those alternative constructions.

B Analysis of Ranking

The pairwise results control one call to the evaluator. We now explain how those guarantees pass through an adaptive sorter. The finite-sample argument uses conditional bounds and does not require independent calls. The central limit theorem needs a stronger assumption: independent evaluation streams along a fixed reference execution.

B.1 CONDITIONAL ERROR SPENDING AND RANKING COST

The general, call-indexed statement. This version allows fresh repeated evaluations and history-dependent budgets. It does not require the caching convention of Algorithm 1. Each call begins from unit evidence, and the ranker observes only its returned winner when selecting the next pair. Let $\mathcal{H}_{k-1}$ be all information available before call k. Within that call, $\mathcal{F}_{k,t}$ contains this information and the first t new judgments, so $\mathcal{H}_{k-1} \subseteq \mathcal{F}_{k,0} \subseteq \mathcal{F}_{k,t}$. The pairwise notation $\mathcal{F}_t$ suppresses k. Assume that all requested pairs are non-tied and that the sorter makes at most M calls on every possible execution. Repeated evaluations count as separate calls. Let $\widehat{\sigma}_{\mathcal{A}}$ be its returned ordering, and $\sigma^{\star}_{\mathcal{A}}$ the output of the same sorter given exact pairwise winners, with the same initial order and random seed.

Suppose the pairwise results give finite bounds V_k and U_k, chosen from $\mathcal{H}_{k-1}$, such that

$$\mathbb{E}[\tau_k \mid \mathcal{H}_{k-1}] \le V_k, \qquad \mathbb{P}(\tau_k > U_k \mid \mathcal{H}_{k-1}) \le \beta_k. \tag{64}$$

Here τ_k is the cost of call k; V_k bounds its conditional expectation and U_k its upper quantile. No independence across calls is needed, but the pairwise assumptions must hold conditional on the history at each call. This is why a pilot may choose tolerances, but its labels cannot simply be reused as fresh observations in a new unit-evidence call.

Theorem B.1 (Ranking with a comparison-based algorithm)**.** *Let $K \le M$ be the number of executed calls, N_0 the pilot cost, and $N = N_0 + \sum_{k=1}^{K} \tau_k$. If the predictable budgets satisfy*

$$\sum_{k=1}^{K} \alpha_k \le \delta, \qquad \sum_{k=1}^{K} \beta_k \le \beta \quad \textit{on every execution}, \tag{65}$$

then

$$\mathbb{P}(\widehat{\sigma}_{\mathcal{A}} \ne \sigma^{\star}_{\mathcal{A}}) \le \delta, \tag{66}$$

$$\mathbb{E}[N] \le \mathbb{E}[N_0] + \mathbb{E}\left[\sum_{k=1}^{K} V_k\right], \tag{67}$$

$$\mathbb{P}\left(N > N_0 + \sum_{k=1}^{K} U_k\right) \le \beta. \tag{68}$$

Thus, with probability at least $1 - \delta - \beta$, the output equals $\sigma^{\star}_{\mathcal{A}}$ and the sample cost is within the stated bound.

Proof outline. We first control the probability that any comparison is incorrect. On the complementary event, the actual sorter follows exactly the same path as the sorter given true pairwise winners. We then add the conditional sample-cost bounds along the executed path. The key point is that whether call k is executed is already determined when that call begins.

Step 1: control comparison errors along the adaptive path. Recall that $K \le M$ is the number of executed calls and $\mathcal{H}_{k-1}$ contains the information available before call k. Set the cost and error indicators of unexecuted calls to zero. The event $\{k \le K\}$ is $\mathcal{H}_{k-1}$-measurable. Conditional pairwise validity therefore gives

$$\mathbb{P}\{k \le K, \text{ call } k \text{ is incorrect}\} \le \mathbb{E}\left[\mathbf{1}_{\{k \le K\}} \alpha_k\right].$$

Summing over the deterministic range $k = 1, \ldots, M$ and using the pathwise spending constraint yields

$$\mathbb{P}\{\text{some executed comparison is incorrect}\} \le \mathbb{E}\left[\sum_{k=1}^{K} \alpha_k\right] \le \delta.$$

Each executed call terminates almost surely because its conditional expected cost is finite. Thus the sorter also terminates after at most M calls.

Couple the actual and noiseless sorters using the same initial input and internal random seed. Their states and requested pairs agree before the first incorrect return: both algorithms have then received exactly the same sequence of winners. If no return is incorrect, their final outputs agree as well. This

proves Eq. 66. Transitivity is not used; with cycles, the reference output is simply the output of this particular sorter on the true pairwise winners.

Step 2: add the conditional sample-cost bounds. The same pre-call measurability justifies applying the tower property to each executed cost:

$$\begin{aligned}\mathbb{E}[N - N_0] &= \sum_{k=1}^{M} \mathbb{E}\big[\mathbf{1}_{\{k\le K\}}\mathbb{E}[\tau_k \mid \mathcal{H}_{k-1}]\big] \\ &\le \mathbb{E}\left[\sum_{k=1}^{K} V_k\right].\end{aligned}$$

For the upper-tail bound, the conditional inequality $\mathbb{P}(\tau_k > U_k \mid \mathcal{H}_{k-1}) \le \beta_k$ gives

$$\mathbb{P}\{\text{some executed call exceeds its } U_k\} \le \mathbb{E}\left[\sum_{k=1}^{K} \beta_k\right] \le \beta.$$

When no call exceeds its bound, $N \le N_0 + \sum_{k=1}^{K} U_k$. This proves Eq. 68. A final union bound gives simultaneous agreement with the noiseless output and the sample budget with probability at least $1-\delta-\beta$. These arguments use conditional guarantees, not independence or a comparison path fixed in advance.

Step 3: translate output agreement into a quality bound. For any pair, changing its relative order changes its contribution to Q by at most $\Delta_{ij} = |2\mu_{ij} - 1|$. Consequently,

$$|Q(\sigma) - Q(\sigma')| \le \sum_{i<j} \Delta_{ij}$$

for any two permutations. The quality difference is zero when the actual and reference outputs agree. Bounding it by this sum on the remaining event proves

$$\mathbb{E}[(Q(\sigma^\star_{\mathcal{A}}) - Q(\widehat{\sigma}_{\mathcal{A}}))_+] \le \delta \sum_{i<j} \Delta_{ij}.$$

This controls the loss caused by statistical comparison errors. It does not compare the noiseless sorter with an optimizer of Q.

Example: a budget fixed before merge sort begins. A merge of two lists containing r items uses at most $r-1$ comparisons, whatever winners the evaluator returns. There are at most $\lceil \log_2 n \rceil$ merge levels, and list lengths sum to n at each level. Hence $M = n\lceil \log_2 n \rceil$ is a valid deterministic bound for $n \ge 2$.

Choose $\alpha_k = \delta/M$ and $\beta_k = \beta/M$. Under the oracle assumptions, Corollary A.1 supplies the bound

$$\begin{aligned}U_e = \Bigg\lceil &\frac{\log(M/\delta)}{G^\star_e} + C_{1,e}\sqrt{\log(M/\delta)\log(2M/\beta)} \\ &+ C_{2,e}\log(2M/\beta) + C_{3,e}\Bigg\rceil\end{aligned}$$

for a call comparing pair e. Maximizing over all pairs that may be requested gives a budget available before sorting:

$$\mathbb{P}\Big\{N > N_0 + M \max_e U_e\Big\} \le \beta.$$

The realized-path bound in the theorem is usually smaller because it sums the bounds for the pairs actually requested. If a pairwise bound depends on earlier calls, a precomputed budget must also be uniform over those possible histories. Other sorters enter through their own deterministic call bounds or through a separate high-probability bound on their number of calls; an expected call count alone does not suffice here.

Proof of Theorem 3.2. Index the newly evaluated pairs by $e_k = (i_k, j_k)$, $k = 1, \dots, K$, where $K \le M$. Their tolerances are chosen before their tests begin. The conditional error-spending argument above gives

$$\mathbb{P}(\text{some comparison is incorrect}) \le \mathbb{E}\left[\sum_{k=1}^{K} \alpha_{e_k}\right] \le \delta.$$

If no comparison is incorrect, the actual and exact-comparator sorters follow the same execution and return the same ordering.

Whether a test is executed is determined before it begins. The conditional pairwise cost bound gives

$$\mathbb{E}[N] \le \mathbb{E}\left[\sum_{k=1}^{K} \left\{\frac{\log(1/\alpha_{e_k})}{G_{e_k}} + C_{e_k}\right\}\right].$$

Here C_{e_k} includes any fixed estimation cost for that pair. Repeated requests use the previously declared winner and incur no additional judgments. Since $K \le M$, bounding the sum by M times its largest summand yields

$$\mathbb{E}[N] \le M\,\mathbb{E}\left[\max_{(i,j)\in\mathcal{E}_{\text{eval}}} \left\{\frac{\log(1/\alpha_{ij})}{G_{ij}} + C_{ij}\right\}\right],$$

which proves Eq. 18.

Let $V = \max_{\mathcal{E}}\{\log(1/\alpha_{ij})/G_{ij} + C_{ij}\}$. Index the at most M requests and give a cached request cost zero. The event that a request occurs is known before its evaluation begins. Its conditional expected cost is at most V, including for a cached request. The tower property therefore gives $\mathbb{E}[N - N_0] \le MV$, proving Eq. 18 without assuming independence across calls.

For the upper-tail conclusion, let U be the maximum over $\mathcal{E}$ of the right side of Eq. 15, with the pair's justified rate and constants and with β replaced by β/M. Each new evaluation exceeds its own bound with conditional probability at most β/M. A union bound over at most M requests shows $N \le N_0 + MU$ with probability at least $1 - \beta$. Intersecting this event with the no-error event gives probability at least $1 - \delta - \beta$. All bounds are conditional on the pilot; integrating them gives unconditional versions, with the expectation also applied to any pilot-dependent right side.

B.2 A CENTRAL LIMIT THEOREM FOR TOTAL SAMPLE COST

Theorem B.2 (Ranking sample-cost fluctuations)**.** *Fix a finite model collection and a comparison-based sorter satisfying the interface in Section 3.2, with at most M calls on every execution. Fix its initial order and internal random seed. Suppose every pair it may request satisfies the matched-oracle assumptions of Theorem 3.1, after exchanging model labels when necessary. Each call starts from unit evidence, uses only its own independent fresh evaluation stream, and has tolerance $\alpha_k = c_k\delta$, where $c_k > 0$ and $\sum_{k=1}^{M} c_k \le 1$ are fixed.*

Let $e_1, \dots, e_m$ be the sequence of comparisons made with exact pairwise winners, including repetitions. With a fixed pilot cost N_0, as $\delta \downarrow 0$,

$$\frac{N_\delta - N_0 - B_\delta \sum_{k=1}^{m} 1/G^\star_{e_k}}{\sqrt{B_\delta}} \Rightarrow \mathcal{N}\left(0, \sum_{k=1}^{m} \frac{v^\star_{e_k}}{(G^\star_{e_k})^3}\right), \qquad B_\delta = \log(1/\delta), \tag{69}$$

where $v^\star_e$ is the correct-direction variance in Eq. 44.

Proof. The actual comparison sequence is random because later pairs depend on earlier returned winners. We therefore first prove the limit for the fixed noiseless sequence and then show that the actual execution coincides with that reference with probability tending to one.

Step 1: construct independent costs on the reference path. Force the sorter to receive the true winner at each call, but run an independent oracle evaluation to measure the cost of that call. Let τ^{ref}_k be this stopping time for pair e_k at tolerance $c_k\delta$, and define

$$N^{\text{ref}}_\delta = N_0 + \sum_{k=1}^{m} \tau^{\text{ref}}_k.$$

A fresh stream is assigned to each call, even if a pair appears more than once. The sequence $e_1, \dots, e_m$ is fixed, so these reference stopping times are independent. This assertion is made for the reference construction, not for the actual data-dependent path.

Step 2: sum the pairwise limits. The threshold for call k is

$$\log\{1/(c_k\delta)\} = B_\delta + \log(1/c_k).$$

The second term is fixed as δ decreases. It changes the centering by a constant and changes the ratio of the two square-root scales by a quantity tending to zero. Theorem A.2 therefore gives

$$\frac{\tau_k^{\mathrm{ref}} - B_\delta/G^\star_{e_k}}{\sqrt{B_\delta}} \;\Rightarrow\; \mathcal{N}\left(0, \frac{v^\star_{e_k}}{(G^\star_{e_k})^3}\right).$$

Independence gives joint convergence to independent normal variables. Since m is fixed, summing these variables proves Eq. 69 for N_δ^{ref} by the continuous-mapping theorem (Billingsley, 1999).

Step 3: transfer the limit to the actual execution. Run the actual sorter with the same streams as the reference at corresponding calls until the two paths separate. They can separate only when a reference evaluation returns an incorrect winner. By pairwise validity,

$$\mathbb{P}\{\text{some reference return is incorrect}\} \le \sum_{k=1}^{m} c_k\delta \le \delta.$$

Otherwise, the requested pairs, stopping times, and final costs are identical. Thus

$$\mathbb{P}\{N_\delta \neq N_\delta^{\mathrm{ref}}\} \le \delta \longrightarrow 0.$$

This is enough to transfer the distributional limit: the probabilities of any event involving the two normalized costs differ by at most δ. No bound on the cost after an incorrect return is needed for this step, and we do not infer convergence of expectations from it. □

For a randomized sorter, the theorem conditions on its seed. Different seeds may give different leading sums $\sum_k 1/G^\star_{e_k}$, so an unconditional Gaussian limit with deterministic centering need not follow. One may likewise condition on a pilot that fixes the sorter and allocation, provided the reference path and positive weights are then fixed and the evaluation streams remain independent. Sharing LLM learning across pairs is compatible with Theorem B.1 when its conditional bounds hold, but does not by itself satisfy the independence assumption of this central limit theorem.

B.3 Derivation of the comparison-error allocation

For a permutation σ, larger $\sigma(i)$ means a higher rank for M_i. Recall that our agreement objective is

$$Q(\sigma) = \sum_{i<j}\{\mu_{ij}\mathbf{1}[\sigma(j) > \sigma(i)] + (1-\mu_{ij})\mathbf{1}[\sigma(i) > \sigma(j)]\}.$$

Reversing a single pair's relative order changes its contribution by $\Delta_{ij} = |2\mu_{ij} - 1|$. A sorting error may also change other relative orders; a budget on direct comparison errors is consequently not a bound on Q-regret.

The candidate-pair allocation in Algorithm 1. Given a finite candidate set, positive estimates $\widehat{G}_{ij}$, and nonnegative estimates $\widehat{\Delta}_{ij}$ fixed before ranking, solve

$$\begin{aligned} \min_{0<\alpha_{ij}\le\bar{\alpha}} \quad & \sum_{(i,j)\in\mathcal{E}} \frac{\log(1/\alpha_{ij})}{\widehat{G}_{ij}} \\ \text{subject to} \quad & \sum_{(i,j)\in\mathcal{E}} \widehat{\Delta}_{ij}\alpha_{ij} \le \varepsilon, \qquad \sum_{(i,j)\in\mathcal{E}} \alpha_{ij} \le \delta. \end{aligned} \tag{70}$$

Take $\varepsilon, \delta > 0$. Strict feasibility follows by choosing every tolerance sufficiently small; the objective is strictly convex and diverges at any zero tolerance. The derivative of the Lagrangian for pair (i, j)

is $-1/(\widehat{G}_{ij}\alpha_{ij}) + \xi\widehat{\Delta}_{ij} + \zeta$. Minimizing on the capped interval gives Eq. 17. The nonnegative multipliers satisfy

$$\xi\Big(\sum_{\mathcal{E}} \widehat{\Delta}_{ij}\alpha_{ij} - \varepsilon\Big) = 0, \qquad \zeta\Big(\sum_{\mathcal{E}} \alpha_{ij} - \delta\Big) = 0.$$

These conditions are necessary and sufficient (Vandenberghe & Boyd, 2004). If a denominator vanishes, the capped solution is $\bar{\alpha}$. Rate estimates must be strictly positive, so a practical implementation floors nonpositive estimates before optimization. A convex solver can solve Eq. 70 directly; no knowledge of the eventual comparison path is needed.

With only the global budget and inactive caps, $\alpha_{ij} = \delta\widehat{G}_{ij}^{-1}/\sum_{\mathcal{E}}\widehat{G}_{ij}^{-1}$. Equal rates give uniform allocation. Poor estimates can make the allocation inefficient, but $\sum\alpha_{ij} \le \delta$ and caching still control the probability of any incorrect comparison. The gap-weighted constraint involves estimates and cannot be interpreted as a true-gap budget without additional estimation control. The program models the cost of all candidates; it need not minimize actual cost for an adaptive ranker that requests only some of them.

Call-list formulations. A difficult comparison costs more samples, but reversing a nearly tied pair has a smaller direct effect on Q. To formalize this tradeoff, first fix a list of m comparison calls $e_1, \dots, e_m$. Write $G_{e_k} > 0$ for the evidence rate used to model call k's cost, and $\Delta_{ij} = |2\mu_{ij} - 1|$ for a pair's preference gap. Each occurrence of a repeated pair receives its own tolerance α_k.

The cost model is $\log(1/\alpha_k)/G_{e_k}$, and the direct expected error cost is at most $\Delta_{e_k}\alpha_k$. For a budget $\varepsilon > 0$ and cap $\bar{\alpha} \in (0, 1)$, consider

$$\min_{0<\alpha_k\le\bar{\alpha}} \sum_{k=1}^{m} \frac{\log(1/\alpha_k)}{G_{e_k}} \quad \text{subject to} \quad \sum_{k=1}^{m} \Delta_{e_k}\alpha_k \le \varepsilon. \tag{71}$$

This is a direct-error surrogate, not a constraint on final ranking quality: a wrong comparison can change both other relative orders and the subsequent comparison path. Its solution is

$$\alpha_k = \min\left\{\bar{\alpha}, \frac{1}{\xi\Delta_{e_k}G_{e_k}}\right\}, \tag{72}$$

where $\xi > 0$ exhausts an active budget; otherwise every tolerance equals its cap. Adding the global constraint $\sum_k \alpha_k \le \delta$ gives

$$\alpha_k = \min\left\{\bar{\alpha}, \frac{1}{G_{e_k}(\xi\Delta_{e_k} + \zeta)}\right\}, \qquad \xi, \zeta \ge 0, \tag{73}$$

with the multipliers chosen to satisfy complementary slackness, and $1/0 = +\infty$. Derivation below establishes both rules using convex optimality conditions (Vandenberghe & Boyd, 2004). The global constraint, not the direct-error constraint alone, supplies the error spending required by Theorem B.1.

We index the allocation by calls rather than by distinct pairs. Fix the list $e_1, \dots, e_m$ with positive gaps Δ_{e_k} and rates G_{e_k}. Each call has its own tolerance α_k, even when a pair repeats. Linearity of expectation bounds the expected direct comparison-error cost by $\sum_k \Delta_{e_k}\alpha_k$, without requiring independent errors. As above, this is a surrogate cost, not the final ranking-quality loss.

Step 1: A Single direct-error budget. The objective in Eq. 71 is decreasing in every tolerance and strictly convex, since its diagonal second derivatives are

$$\frac{1}{G_{e_k}\alpha_k^2} > 0.$$

It diverges when any tolerance tends to zero, so an optimum is attained with all tolerances positive. Positive budgets admit strictly feasible tolerances by taking them sufficiently small. The convex Lagrange-multiplier conditions are therefore necessary and sufficient (Vandenberghe & Boyd, 2004).

If $\varepsilon \ge \bar{\alpha}\sum_k \Delta_{e_k}$, every tolerance can equal its cap, and monotonicity makes this optimal. Otherwise the direct-error budget is active. Introduce its multiplier $\xi > 0$ and minimize the Lagrangian

Algorithm 1 Ranking with allocated pairwise evaluations

Require: Ranker $\mathcal{A}$, candidate pairs $\mathcal{E}$, budgets ε, δ, cap $\bar{\alpha}$
1: Initialize an empty cache and run $\mathcal{A}$.
2: **while** $\mathcal{A}$ requests a pair (M_i, M_j) **do**
3: Estimate gaps and rates for pair (M_i, M_j); set tolerances by Eq. 17.
4: **if** the unordered pair is not cached **then**
5: Run `PairEval`(M_i, M_j, α_{ij}) with new judgments; save its winner.
6: **end if**
7: Return the winner to $\mathcal{A}$ in the requested orientation.
8: **end while**
9: **return** the ordering produced by $\mathcal{A}$.

separately over $0 < \alpha_k \le \bar{\alpha}$. The part depending on call k is

$$\frac{\log(1/\alpha_k)}{G_{e_k}} + \xi \Delta_{e_k} \alpha_k.$$

Its derivative is $-1/(G_{e_k}\alpha_k) + \xi\Delta_{e_k}$. The unconstrained minimizer is $1/(\xi\Delta_{e_k}G_{e_k})$; imposing the upper cap gives

$$\alpha_k = \min\left\{\bar{\alpha}, \frac{1}{\xi\Delta_{e_k}G_{e_k}}\right\}.$$

This proves Eq. 72 without introducing a separate multiplier for each cap.

To choose ξ, sum $\Delta_{e_k}\alpha_k$ over calls. This sum is continuous in ξ, equals the capped budget for sufficiently small ξ, and strictly decreases once any tolerance leaves its cap. It tends to zero as $\xi \to \infty$. There is therefore a unique ξ exhausting an active budget. Strict convexity gives a unique allocation.

Step 2: Add a global comparison-error budget. Assign a multiplier $\zeta \ge 0$ to $\sum_k \alpha_k \le \delta$. The per-call derivative becomes

$$-\frac{1}{G_{e_k}\alpha_k} + \xi\Delta_{e_k} + \zeta.$$

Minimizing over the same capped domain gives Eq. 73. The multipliers also satisfy

$$\xi\left(\sum_{k=1}^{m}\Delta_{e_k}\alpha_k - \varepsilon\right) = 0, \qquad \zeta\left(\sum_{k=1}^{m}\alpha_k - \delta\right) = 0.$$

The same strict-feasibility argument makes these conditions sufficient. The allocation remains unique, although the multipliers need not be. If both multipliers vanish, the decreasing objective places every tolerance at its cap; this is also the limiting interpretation of the displayed allocation formula.

What the optimization does and does not establish. Additive expected costs independent of the tolerances do not change this optimization for a fixed call list. The pairwise analysis motivates its leading logarithmic objective; the rule is not claimed to minimize the exact stopping cost or a high-probability quantile. For an adaptive sorter, the relevant spending constraint must hold along every possible execution, as in Theorem B.1. In particular, a tolerance assigned once to a pair cannot be spent again without accounting for each fresh call.

C PROMPTS

In this section, we present the prompt used for prediction using LLM estimator.

LLM Estimator Prompt

You are helping evaluate a pairwise comparison of two AI models (Model A vs. Model B).

Your task: from the list of remaining questions below, identify the top {K} questions that are most likely to **discriminate** between the two models. For each question, estimate "p_{win}": the probability that **Model A** produces a clearly better response than **Model B**.

- $p_{\text{win}} \approx 0.0 \rightarrow$ you are confident **Model B** will clearly win
- $p_{\text{win}} \approx 0.5 \rightarrow$ you expect a **tie** (both models perform similarly)
- $p_{\text{win}} \approx 1.0 \rightarrow$ you are confident **Model A** will clearly win

Prefer questions where p_win is far from 0.5 (strongly discriminative in either direction).

{History}

Here are the remaining questions (indexed 0 to {N-1}):

{Questions}

Respond with a JSON object containing a "candidates" list. Each item in the list must have:

- "index": integer index of the question
- "p_{win}": float between 0 and 1 (probability that Model A wins)

Return exactly $\{K\}$ candidates. Do not include any other text.

D DATASET VISUALIZATION

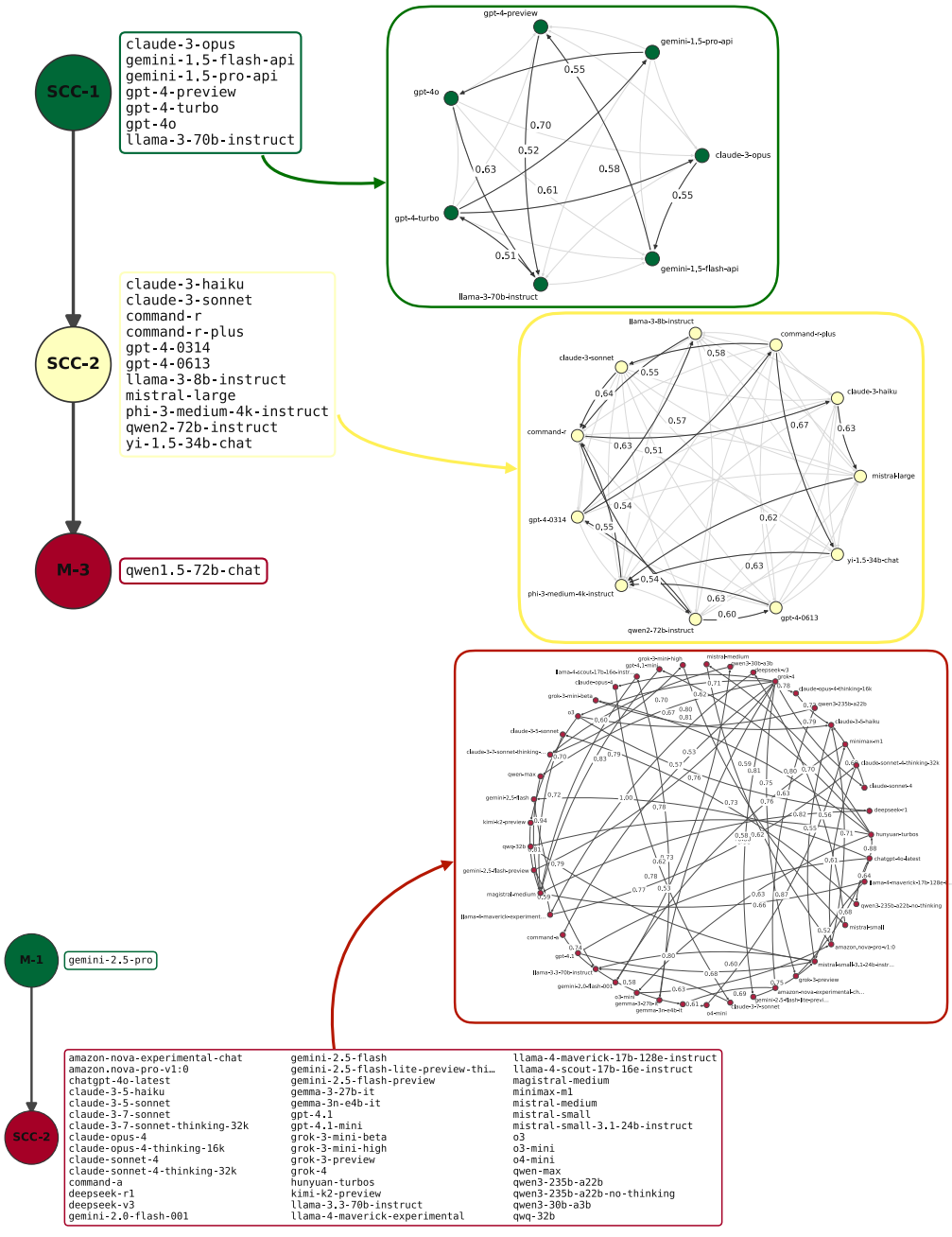


Figure 3: Directed preference graph for the PPE (top) and LM Arena (bottom) datasets using LLM as nodes and human preference as edge direction. The resultant graph shows that there are many strongly connected components and an optimal linear ordering of LLMs doesn't exist.

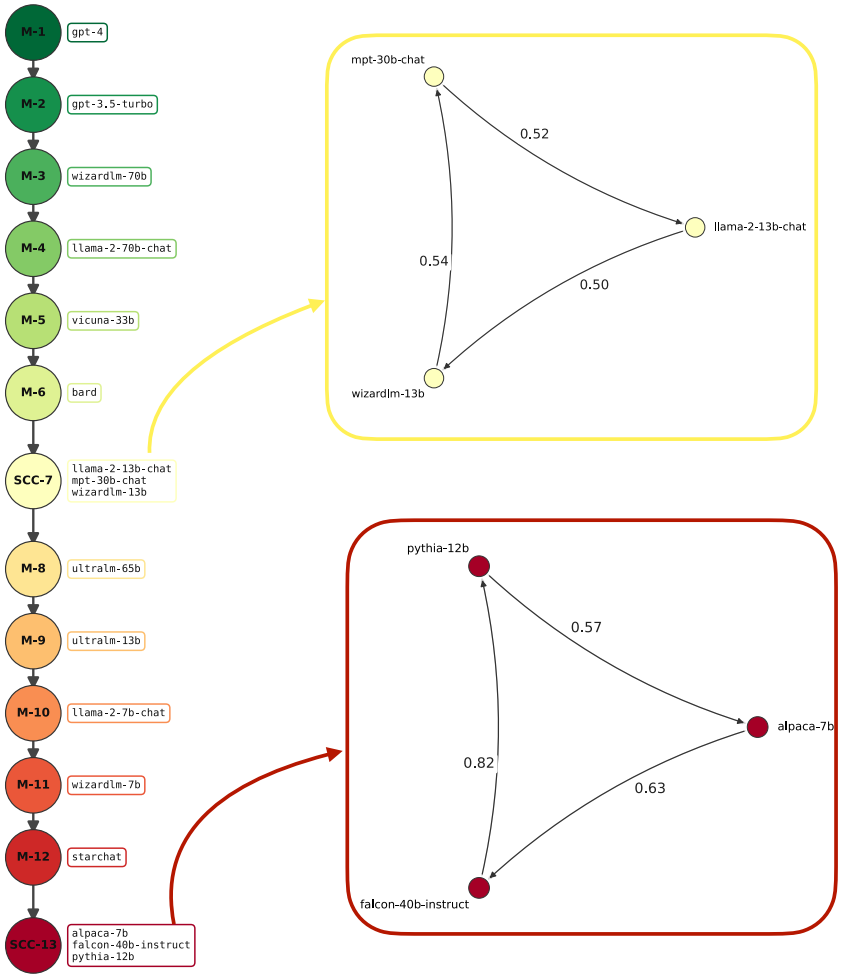


Figure 4: Directed preference graph for the Ultrafeedback dataset. The resultant graph appears much more linear than PPE and LM Arena graphs. This allows the ranking algorithms to converge faster using only 8% of the total evaluation samples.

E EXPERIMENTS

E.1 IMPLEMENTATION DETAILS

In this section, we provide the implementation details for different baselines and variants of our proposed algorithm reported in the paper. For the quick sort (doubling α) experiments, we start with $\alpha = 0.01$. We use $\bar{\alpha} = 0.99$. In adaptive α methods, we use 20 samples for the pilot phase to estimate $\hat{\Delta}_{ij}$ and $\hat{G}_{ij}$ and these samples are accounted for in the overall sample used. For baselines that use ϵ-greedy sampling, we set $\epsilon = 0.5$. For LLM-based preference probability prediction, we ask the LLM to return $K = 10$ candidates and their corresponding $\pi(x)$ scores, we set $\kappa = 0.001$ and use uniform probability over the remaining prompts.

E.2 ADDITIONAL EXPERIMENTS

Ablations with different LLMs. In this experiment, we replace Gemini 3.6 Flash with an open-source LLM, Gemma 4 26B, for prompt selection. As shown in Table 3, the results are comparable to those in Table 2 in both ranking quality and sample usage, with Gemma 4 26B even outperforming Gemini 3.6 Flash on LMArena. This demonstrates that the choice of prompt-selection LLM has minimal impact on ranking performance, and that even smaller open-source models can effectively select prompts for efficient evaluation.

Table 3: We report the quality of ranking (agreement with pairwise data) while using an open-source Gemma 4 26B model for prompt selection. We observe that the overall results are very close to the fraction of samples used when using Gemini-based LLM in Table 2.

	LMArena		PPE		Ultrafeedback	
Method	Quality (↑)	Samples (↓)	Quality (↑)	Samples (↓)	Quality (↑)	Samples (↓)
`Active Ranking` **[AR]**	59.8±0.5	72.9%	57.4±0.8	51.8%	65.4±3.3	64.8%
`Active Testing` (w/ QS) **[ATQ]**	60.6±0.5	19.0%	59.2±0.1	30.9%	75.0±0.1	19.9%
`Active Testing` (w/ MS) **[ATM]**	61.0±0.2	16.8%	59.0±0.5	29.8%	**75.1**±0.0	19.9%
`Quick Sort` (const. α) **[QS]**	**61.5**±0.3	15.5%	**59.4**±0.3	29.6%	**75.1**±0.0	14.4%
`Quick Sort` (doubling α) **[QSD]**	61.4±0.3	15.1%	58.8±0.4	32.7%	74.8±0.2	16.2%
`Merge Sort` (const. α) **[MS]**	61.4±0.4	14.5%	59.2±0.4	29.1%	**75.1**±0.0	13.2%
`Quick Sort` (adaptive α) **[QSA]** (Ours)	61.2±0.2	10.8%	58.2±0.8	**21.9%**	74.8±0.1	9.5%
`Merge Sort` (adaptive α) **[MSA]** (Ours)	61.4±0.1	**9.4%**	58.6±0.3	22.4%	74.5±0.3	**9.1%**
`Expected (MC)` **[EX]**	50.0 ±0.0	-	50.0 ±0.0	-	50.1 ±0.1	-
`Bradley-Terry` **[BT]**	62.6 ±0.0	100%	59.8 ±0.0	100%	75.0 ±0.0	100%
`HodgeRank` **[HG]**	62.5 ±0.0	100%	59.8 ±0.0	100%	75.0 ±0.0	100%
`Oracle` **[OE]**	63.6 ±0.0	100%	60.8 ±0.0	100%	75.2 ±0.0	100%

Ranking Visualization. In this experiment, we select the top 10 models (to keep computation tractable) from each dataset and perform exhaustive search over 10! permutations to find the linear ordering. In Figure 5, we compare the ranking produced by our method (adaptive merge sort) with the ranking obtained from exhaustive search. On Ultrafeedback and LMArena datasets, we observe that the ranking generated closely matches the one generated by exhaustive search achieving a Kendall's $\tau = 0.91$. On Ultrafeedback, we observe that the errors concentrate around ranks 6-8, which is expected as a transitive preference among these models is not supported by the data as shown in Figure 4.

Among the three datasets, PPE is where our ranking deviates most from the exhaustive optimum (Kendall $\tau = 0.51$). PPE also has the fewest annotations per model pair (~ 78), which suggests the deviations are likely due to the limited number of samples available rather than the ranking algorithm itself. To investigate this hypothesis, we tested every pair against the null hypothesis that its win rate is 0.5, using an exact two-sided binomial test. Only 22 out of 45 model pairs (for 10 models) reject the null at $\alpha = 0.05$. Since a ranking outputs a linear ordering, the remaining 23 pairs must still be ordered, with no evidence favouring either direction. Our method (adaptive merge sort) produces a linear ordering of 10 models (shown in Figure 5 (right)), which implies a pairwise decision for 45 possible pairs. We evaluate these pairwise decisions against the ground truth in PPE, and find 10 wrong decisions. However, 9 out of the 10 are among those that even the offline fixed budget test failed to distinguish the winner. Overall, this shows that the apparent deviation from the ranking

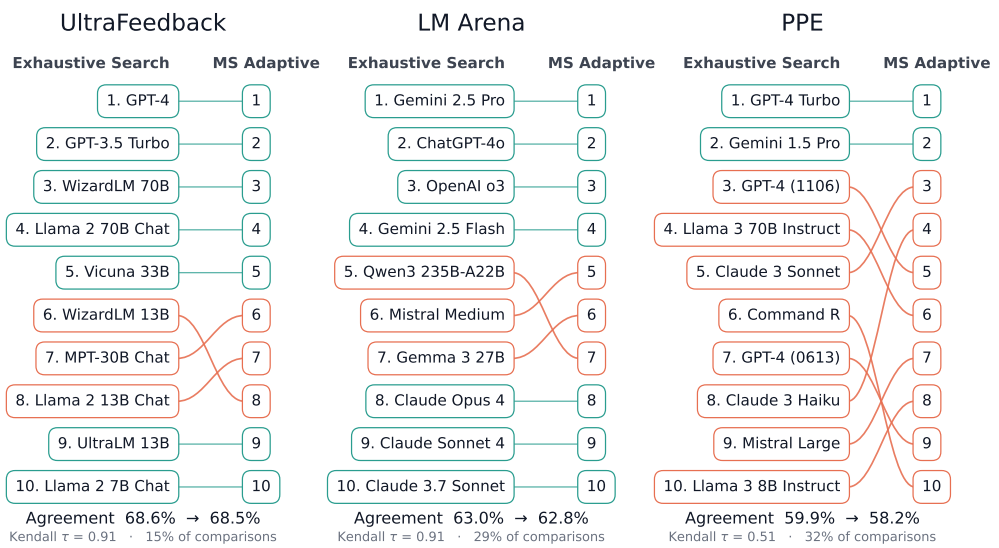


Figure 5: Visualization of the ranking obtained by exhaustive search and adaptive merge sort (MS adaptive) across Ultrafeedback, LM Arena, and PPE datasets.

Table 4: Quality of ranking (agreement with pairwise data) and fraction of total samples used, with uniform random vs. LLM-guided sampling inside each pairwise e-process, with standard error over 3 runs. A higher quality score and lower fraction of samples used is better.

		LMArena		PPE		Ultrafeedback	
Method	**Sampling**	Quality (↑)	Samples (↓)	Quality (↑)	Samples (↓)	Quality (↑)	Samples (↓)
`Quick Sort` (const. α) **[QS]**	Random	62.0±0.3	15.6%±0.6	58.7±0.5	28.4%±0.6	**75.0**±0.0	14.5%±0.8
	LLM	61.9±0.1	15.0%±0.7	58.4±0.3	29.7%±0.1	74.9±0.1	11.5%±0.8
`Quick Sort` (doubling α) **[QSD]**	Random	62.0±0.3	13.6%±1.3	57.7±0.4	30.7%±1.7	74.7±0.3	14.5%±0.6
	LLM	61.3±0.1	15.9%±0.4	58.4±0.2	31.3%±1.1	**75.0**±0.1	15.4%±0.2
`Merge Sort` (const. α) **[MS]**	Random	**62.1**±0.1	15.2%±0.6	57.9±0.5	28.5%±1.5	74.9±0.1	14.6%±0.6
	LLM	60.8±0.2	14.4%±0.5	59.2±0.2	28.1%±0.8	**75.0**±0.1	12.2%±0.8
`Quick Sort` (adaptive α) **[QSA]**	Random	61.3±0.2	13.4%±0.8	58.6±0.4	27.1%±0.5	74.8±0.2	11.9%±0.7
	LLM	61.0±0.3	**9.7%**±0.8	**59.3**±0.1	21.8%±1.3	74.7±0.2	**6.8%**±0.5
`Merge Sort` (adaptive α) **[MSA]**	Random	60.6±0.3	12.9%±0.3	58.8±0.3	25.9%±0.6	74.8±0.1	11.7%±0.6
	LLM	60.1±0.5	**9.7%**±0.8	58.3±0.4	**21.7%**±1.3	74.5±0.3	7.6%±0.9

derived from exhaustive search is an artifact of the limited data available in the PPE dataset and not a limitation of the proposed algorithm.

Synthetic Results. We evaluate our approach in a synthetic setting to exhaustively simulate performance across different win rate gaps (Δ). For a given gap, we generate preference probabilities from a normal distribution $\pi(x) \sim \mathcal{N}(\mu, \sigma)$, where $\mu = (1 + \Delta)/2$, for 10K samples. Using these probabilities, we sample pairwise evaluation labels $y \sim \text{Bern}(\pi(x))$ and run our proposed e-process-based pairwise evaluation with both random and active sampling, recording the stopping time (number of samples required to determine a winner). Across 50 trials per gap, active sampling uses $\sim$27% fewer samples compared to random sampling, with significant gains especially when win rate gap is small where distinguishing between models is hardest. These results showcase that active sampling provides the greatest benefit precisely in the regime where sample efficiency matters most.

Sampling Ablations. In this experiment, we perform ablations of all sorting based ranking methods by using either random or LLM-based sampling. We report the results in Table 4. We observe that LLM-based sampling provides gains in the number of samples in the majority of settings, with the gains being most pronounced in adaptive α methods. We hypothesize that these gains are more prominent because for higher α we need fewer samples to conclude hypothesis testing and the LLM-based prediction is effective in identifying discriminative examples. We also observe that adaptive α

methods showcase sample count gains irrespective of the underlying sampling method. This implies that the gains are an artifact of the derived error allocation strategy (Section 3.2) that are further amplified using the LLM-based sampling method.