

RankBALD: Ranking-Aligned Active Evaluation for Language Models

Bonaventure F. P. Dossou^{♡†} Jackie Chi Kit Cheung^{♡†‡}

[♡]Mila Québec AI Institute [†]McGill University [‡]Canada CIFAR AI Chair, Mila

Abstract

Language model evaluation faces a structural challenge: as performance gaps between state-of-the-art models narrow, static evaluation sets require increasingly large budgets to yield stable, reliable model comparisons. Adaptive evaluation can improve measurement efficiency by sequentially selecting informative test items, but existing approaches primarily focus on estimating latent parameters that describe model capabilities. This focus overlooks the fact that experimenters may have diverse downstream evaluation goals, such as to produce a relative ordering of models. We introduce a framework for adaptive evaluation in the Item Response Theory (IRT) setting that is compatible with both capability assessment and model-ranking objectives. Within this framework, ranking reliability depends on resolving uncertainty near the boundaries of pairwise orderings rather than minimizing overall posterior variance. To operationalize this insight, we draw on Bayesian active learning (BAL), a family of methods that select examples by maximizing expected information gain under a posterior belief. BAL was originally developed for information-efficient parameter identification. We adapt BAL to the IRT evaluation setting, giving rise to what we call *active evaluation*: sequential item selection driven by principled uncertainty reduction over model abilities and their relative ordering. We develop a family of acquisition criteria that increasingly align with the ranking objective, culminating in **RankBALD**, which directly maximizes the mutual information between candidate test items and the pairwise model-ordering variables. Experiments across six English benchmarks from the Open LLM Leaderboard and five African-language benchmarks show that, compared with ability-estimation baselines, ranking-aligned acquisition consistently improves ranking validity and reduces evaluation variance. These results demonstrate that aligning item selection with the downstream evaluation objective substantially improves the efficiency and reliability of active evaluation.

1 Introduction

Large language models (LLMs) are compared using benchmarks that guide deployment decisions, leaderboard governance, and scientific progress. In these settings, the primary objective is often to determine reliable relative rankings among models rather than absolute performance. However, static test sets struggle to distinguish models with similar performance. As performance gaps between state-of-the-art (SOTA) systems narrow, obtaining stable rankings requires increasingly large evaluation budgets. This motivates *adaptive evaluation*, where test items are selected sequentially to maximize measurement efficiency.

Item Response Theory (IRT) (Lord, 1952; van der Linden & Hambleton, 1997; DeMars, 2010) provides a principled framework for modeling latent ability while accounting for item difficulty and discrimination. Originally developed for computerized adaptive testing (Meijer & Nering, 1999; Chang, 2015), IRT has more recently been adopted for the adaptive evaluation of machine learning systems, where benchmark items are selected dynamically

to maximize the informativeness of model responses (Rodriguez et al., 2021; Hofmann et al., 2025; Lalor et al., 2024; Song et al., 2025). In this literature, item selection strategies are primarily designed to improve the precision of ability estimation, typically using statistical informativeness measures such as Fisher information that prioritize items most sensitive to the current ability estimate. While effective for measuring individual model capability, this objective does not directly address the reliability of relative model rankings, which often constitute the primary goal of modern evaluation. One recent example is Fluid Language Model Benchmarking (Hofmann et al., 2025), which applies Fisher-information-based selection to construct adaptive evaluation sets for language models.

The sequential nature of adaptive evaluation mirrors a problem in machine learning: Bayesian active learning (BAL), where the goal is to efficiently identify unknown parameters by sequentially selecting the most informative observations under a constrained budget (Settles, 2009; Cohn et al., 1996). BAL methods, such as Bayesian Active Learning by Disagreement (BALD) (Houlsby et al., 2011; Gal et al., 2017), select examples by maximizing the mutual information between model parameters and predicted outputs, targeting the reduction of epistemic uncertainty. Within a two-parameter logistic (2PL) IRT model (Lord, 1952; Birnbaum, 1969), a standard probabilistic response model that characterizes item responses as a function of model ability, item difficulty, and item discrimination, BAL provides a natural framework for selecting informative evaluation items under posterior uncertainty. In particular, it offers a principled way to define acquisition objectives in terms of expected information gain, thereby explicitly tying evaluation procedures to the quantities one aims to infer.

The central claim of this work is that adaptive evaluation should be aligned with the decision it is intended to support. When the primary objective is reliable model ranking under limited budgets, acquisition strategies should directly target ordering uncertainty rather than treating ranking as a by-product of ability estimation. To operationalize this principle, we develop a sequence of Bayesian acquisition strategies that increasingly align with the ranking objective. We introduce the *active evaluation* setting (see Section 3.1) and provide a direct extension of BALD. We then introduce *DiffBALD*, which incorporates difficulty-aware weighting to prioritize items whose difficulty aligns with the region of posterior ability uncertainty. Building on this, *VarBALD* further augments the acquisition objective with an explicit preference for items that reduce posterior variance in ability estimates, improving evaluation stability under limited query budgets. Finally, we propose **RankBALD**, an acquisition criterion that directly targets uncertainty in pairwise model ordering by maximizing mutual information over ranking variables. This progression reflects a shift from generic uncertainty reduction toward acquisition strategies explicitly aligned with the downstream evaluation goal.

We evaluate our approach across six Open LLM Leaderboard benchmarks (Beeching et al., 2023): ARC Challenge (Clark et al., 2018), GSM8K (Cobbe et al., 2021), HellaSwag (Zellers et al., 2019), MMLU (Hendrycks et al., 2021), TruthfulQA (Lin et al., 2022), and WinoGrande (Sakaguchi et al., 2021). Across 2,802 model-benchmark combinations, RankBALD achieves the strongest ranking validity, lowest model variance, and fastest saturation among all tested acquisition strategies. Moreover, we observe the same performance pattern on five African-language benchmarks (Ojo et al., 2025), indicating that ranking-aligned acquisition improves evaluation reliability beyond English-centric settings (Appendix A.2). These empirical findings highlight three main contributions of this work: **(a)** we introduce *active evaluation*, an objective-aligned Bayesian evaluation design problem, demonstrating that acquisition criteria should be defined with respect to the downstream evaluation goal rather than treated as generic uncertainty-reduction heuristics; **(b)** we introduce difficulty-aware and variance-aware extensions of BALD that improve posterior uncertainty reduction and evaluation stability under limited query budgets; **(c)** we propose **RankBALD**, an acquisition objective that directly maximizes mutual information over pairwise ordering variables, thereby aligning active item selection with the goal of reliable model ranking.

2 Background: Measurement and active evaluation

2.1 Item Response Theory for Language Model Evaluation

IRT (Lord, 1952; van der Linden & Hambleton, 1997; DeMars, 2010) provides a principled statistical framework for estimating latent ability from observed responses to evaluation items. Originally developed in psychometrics, IRT has recently been adopted in NLP for model evaluation (Rodriguez et al., 2021; Lalor et al., 2024; Song et al., 2025; Hofmann et al., 2025), enabling more reliable comparisons of models by accounting for variation in item difficulty and discriminative power. In the context of LLM evaluation, benchmark items vary widely in difficulty and informativeness. Standard metrics such as accuracy implicitly treat all items as equally informative (Rodriguez et al., 2021), which can mislead comparisons when models differ only slightly in performance. IRT addresses this limitation by explicitly modeling how informative each item is for distinguishing between models. Consider a population $\mathcal{M} = \{\mathcal{M}_1, \dots, \mathcal{M}_K\}$ of LLMs, evaluated on a benchmark $\mathcal{Q} = \{q_j\}_{j=1}^N$. Each interaction yields a binary outcome indicating whether the model correctly answers an item. Let $U_{kj} \in \{0, 1\}$ denote the response of model $\mathcal{M}_k$ on item q_j . Following recent adaptive evaluation work (Hofmann et al., 2025), we model responses using the two-parameter logistic (2PL) item response theory model (Lord, 1952; Birnbaum, 1969). Under this framework, each model $\mathcal{M}_k$ is characterized by a latent ability parameter θ_k , while each item q_j is characterized by a difficulty parameter b_j and a positive discrimination parameter a_j . The probability that model $\mathcal{M}_k$ responds correctly to item q_j is then given by $p(U_{kj} = 1) = \sigma(a_j(\theta_k - b_j))$, where σ denotes the logistic function. In this setting, the latent ability θ_k can be interpreted as the overall capability of model $\mathcal{M}_k$ on the benchmark, while the difficulty parameter b_j reflects how challenging an item is across models. The discrimination parameter a_j quantifies an item’s sensitivity to differences in model capability, determining how sharply it separates stronger models from weaker ones. This provides a more nuanced view of evaluation than aggregate accuracy, identifying which items are most informative for comparing models. Together, these parameters characterize both where an item is informative and how sharply it discriminates, enabling principled reasoning about item selection. The Fisher information contributed by item q_j for a model with ability θ_k is: $I(\theta_k, a_j, b_j) = a_j^2 \sigma(a_j(\theta_k - b_j)) (1 - \sigma(a_j(\theta_k - b_j)))$, which is maximized when item difficulty is close to the model’s ability. Items that are far too easy or too difficult therefore provide little information about capability, highlighting that not all benchmark items contribute equally to model comparison and motivating the need for active item selection.

2.2 Bayesian Active Learning

Active learning aims to reduce annotation costs by selecting, at each step, the most informative examples under a limited budget (Settles, 2009; Cohn et al., 1996). Rather than relying on random sampling, the learner evaluates a pool of candidates and queries those expected to provide the greatest improvement. In NLP, this corresponds to selecting examples whose annotations would most improve model performance on downstream tasks. In the BAL setting, this is formalized by maintaining a posterior distribution over model parameters, capturing epistemic uncertainty, and selecting examples that maximize expected information gain, i.e., those whose labels are most informative about the model (Houlsby et al., 2011; Gal & Ghahramani, 2016; Gal et al., 2017; Kirsch et al., 2019). A widely used BAL criterion is BALD (Houlsby et al., 2011; Gal et al., 2017). For an input $x \in \mathcal{X}$ with predicted output y , BALD selects

$$x^* = \arg \max_{x \in \mathcal{X}} I(y; \omega \mid \mathcal{D}) = H[y \mid \mathcal{D}] - \mathbb{E}_{p(\omega \mid \mathcal{D})} [H[y \mid \omega]], \quad (1)$$

where ω denotes model parameters, $\mathcal{D}$ the training data, and $H(\cdot)$ entropy. This objective favors examples whose predictions are uncertain under the posterior predictive distribution but confident under individual parameter settings, i.e., where plausible models disagree. As a result, BALD isolates epistemic uncertainty and prioritizes examples that are most informative for refining the posterior over ω . We adapt BALD for active evaluation in Section 4.1.

3 Active Evaluation Framework

3.1 Sequential Active Evaluation

We introduce *active evaluation*, a form of *adaptive evaluation* in which benchmark items are selected sequentially based on the model’s previous responses. At each step, the evaluator updates its belief about the model’s capabilities and selects the next item to reduce uncertainty within a constrained evaluation budget. This parallels active learning, but differs in that the objective is not to improve a predictive model, but to efficiently characterize and compare a set of models. The goal is therefore to construct an evaluation set that efficiently supports the downstream evaluation objective under a limited budget. Active evaluation proceeds in a sequence of *acquisition rounds*. In each round, previously collected responses are used to update the evaluator’s belief about the model’s abilities, and this updated belief guides the selection of the next evaluation item. Starting from an empty evaluation set $Q_0^* = \emptyset$, an external evaluator E iteratively constructs a benchmark by selecting informative items from a benchmark $\mathcal{Q}$. At acquisition round r , each LLM $\mathcal{M}_k$ in a population $\mathcal{M}$ is evaluated on the current item set Q_{r-1}^* . Using the observed responses, E updates its belief about model abilities by estimating latent parameters $\{\hat{\theta}_k^r\}_{k=1}^K$, yielding an updated posterior belief state p_E^r . Based on this belief, the evaluator assigns each unasked item $q_j \in \mathcal{Q} \setminus Q_{r-1}^*$, an *acquisition score* $\mathcal{A}^r(q_j)$, which quantifies the expected utility of evaluating on that item under the current experimental objective. The next evaluation item is then selected as

$$q_j^* = \arg \max_{q_j \in \mathcal{Q} \setminus Q_{r-1}^*} \mathcal{A}^r(q_j)$$

This sequential procedure continues until the evaluation budget is exhausted, yielding a final benchmark $Q^* = Q_R^*$ with calibrated ability estimates $\{\hat{\theta}_k^R\}$. The evaluator’s objective $\mathcal{G}$ takes one of two forms: (a) **ability estimation**, targeting accurate recovery of $\{\theta_k\}$, or (b) **ranking**, targeting reliable recovery of the ordering π such that $\theta_{\pi(1)} \geq \dots \geq \theta_{\pi(K)}$. While correct ranking follows from perfect ability recovery, budget-constrained benchmarking leaves estimates uncertain, and improvements in marginal estimation precision do not necessarily translate into more reliable pairwise ordering. Existing approaches, such as *Fluid Benchmarking*, select items by maximizing Fisher information with respect to latent abilities, which, via the Cramér-Rao bound (Cramér, 1999; Rao, 1992), quantifies estimation precision but provides no direct control over pairwise discrimination. This mismatch motivates our primary contribution: a Bayesian acquisition criterion that directly targets ranking reliability, introduced in Section 4.

3.2 Probabilistic Belief over Abilities

We formalize the probabilistic belief that the evaluator E maintains over latent model abilities $\{\theta_k\}$ throughout the active evaluation process, given a selected set of benchmark items. This belief induces corresponding uncertainty over relative model ordering, since ranking outcomes are determined by pairwise comparisons among the latent abilities, and underlies the acquisition mechanisms introduced in Section 4. We consider a population of LLMs $\mathcal{M}$, with each LLM parameterized by fixed weights w_k . Given a benchmark $\mathcal{Q}$ with ground-truth labels $\{y_j\}_{j=1}^N$, each model produces a binary correctness response:

$$U_{kj} = \begin{cases} 1 & \text{if } \mathcal{M}_k(q_j) = y_j \\ 0 & \text{otherwise} \end{cases}, \quad U = (U_{kj}) \in \{0, 1\}^{K \times N} \quad (2)$$

To obtain a calibrated representation of the model’s capability, responses are mapped into a latent ability space using a 2PL IRT model. This representation provides the probabilistic belief state required for active item selection:

$$U_{kj} \sim \text{Bernoulli}(p_{kj}(\theta)), \quad p_{kj}(\theta) := p(U_{kj} = 1 \mid \theta) = \sigma(a_j(\theta - b_j)) \quad (3)$$

We use the same item parameters a_j and b_j as Hofmann et al. (2025), treating them as precomputed and fixed. At acquisition round r , E estimates model abilities θ_k from the

subset of already acquired items $Q_r^* \subseteq Q$ by maximizing the posterior likelihood under the fixed item parameters:

$$\hat{\theta}_k^r = \arg \max_{\theta} \prod_{j \in Q_r^*} [p_{kj}(\theta)]^{U_{kj}} [1 - p_{kj}(\theta)]^{1-U_{kj}}$$

We use the Maximum A Posteriori (MAP) estimation (Birnbaum, 1969) to determine $\hat{\theta}_k^r$. Let $U_{k,Q_r^*} = \{U_{kj} : j \in Q_r^*\}$ be the set of observed responses. Consequently, the likelihood of the correctness of the observed responses under a candidate ability value θ is:

$$p(U_{k,Q_r^*} | \theta) = \prod_{j \in Q_r^*} p_{kj}(\theta)^{U_{kj}} (1 - p_{kj}(\theta))^{1-U_{kj}}$$

The evaluator's posterior belief, over model abilities, at round r is therefore defined as:

$$p_E^r(\theta) \stackrel{\text{def}}{=} p(\theta | U_{k,Q_r^*}) \propto p(U_{k,Q_r^*} | \theta) p(\theta)$$

The ability θ_k is treated as a latent random variable with posterior distribution $p_E^r(\theta)$. In practice, we approximate this posterior using a Laplace approximation centered at the MAP estimate (see Appendix A.6 for details).

4 Ability-Aware Acquisition Functions

Given the posterior belief $p_E^r(\theta)$ defined in the previous section, E must select a new question q_j to refine $p_E^r(\theta)$. We build acquisition objectives of increasing alignment with the ranking evaluation goal: uncertainty in response predictions (BALD), difficulty-aligned uncertainty (DiffBALD), reduction of posterior ability variance (VarBALD), and uncertainty in pairwise model ordering (RankBALD). Appendix A.1 gives a step-by-step derivation bridging BALD's original notation to each of these criteria in turn, explicitly separating what is inherited from BALD, what is required by the shift from training to evaluation, and what is novel to RankBALD.

4.1 BALD Acquisition for Active Evaluation

We extend the BALD acquisition principle introduced in Section 2 to the active evaluation setting (more details in Appendix A.1), where the acquisition aims to reduce uncertainty about model abilities under an IRT response model. For a candidate item q_j , the BALD acquisition score is defined as

$$\text{BALD}_{kj}^r = I(U_{kj}; \theta_k) = H(\mathbb{E}_{p_E^r(\theta)}[p_{kj}(\theta)]) - \mathbb{E}_{p_E^r(\theta)}[H(p_{kj}(\theta))],$$

where $H(\cdot)$ denotes Bernoulli entropy. Since $p_E^r(\theta)$ has no closed form, these expectations are approximated via Monte Carlo (MC) sampling (see Appendices A.6, A.8). Let

$$\theta_k^{(t,r)} \sim p_E^r(\theta), \quad p_{kj}^{(t,r)} = p_{kj}(\theta_k^{(t,r)}), \quad \bar{p}_{kj}^r = \frac{1}{T} \sum_{t=1}^T p_{kj}^{(t,r)},$$

where $T = 64$ denotes the number of MC samples. This yields the estimator

$$\text{BALD}_{kj}^r \approx H(\bar{p}_{kj}^r) - \frac{1}{T} \sum_{t=1}^T H(p_{kj}^{(t,r)})$$

4.2 Difficulty-Aware BALD Acquisition (DiffBALD)

Under the 2PL model, informative items are those whose difficulties b_j align with the region of posterior uncertainty, where response outcomes are most sensitive to ability. This

motivates a difficulty-aware BALD acquisition criterion. To incorporate this alignment in a simple and computationally tractable way, we introduce a smooth weighting centered at the item difficulty, which serves as a proxy for the region where item informativeness is highest under the 2PL model. For an ability value θ , we define

$$\pi_{kj}(\theta) = \exp\left(-\frac{1}{2}(b_j - \theta)^2\right), \quad \bar{\pi}_{kj}^r = \frac{1}{T} \sum_{t=1}^T \pi_{kj}(\theta_k^{(t,r)})$$

We define the DiffBALD acquisition function as: $\mathcal{A}^r(q_j) = \text{DiffBALD}_{kj}^r = \bar{\pi}_{kj}^r \cdot \text{BALD}_{kj}^r$.

4.3 Variance-Aware BALD (VarBALD)

While DiffBALD targets mutual information between ability and correctness outcomes, it does not explicitly account for residual posterior uncertainty in the ability parameter itself. Under the Laplace approximation, this uncertainty is captured by the posterior variance $\sigma_{k,r}^2$, given by the inverse curvature of the log-posterior (Bishop, 2006; Murphy, 2012a):

$$\sigma_{k,r}^2 = \left(-\frac{\partial^2}{\partial \theta^2} \log p_E^r(\theta) \Big|_{\theta=\hat{\theta}_k^r} \right)^{-1}$$

To better align acquisition with the reduction of this uncertainty, we augment DiffBALD with an explicit variance-reduction term. For each candidate item q_j , we define the posterior-averaged Fisher information

$$\bar{I}_{kj}^r = \mathbb{E}_{\theta \sim p_E^r} \left[a_j^2 p_{kj}(\theta) (1 - p_{kj}(\theta)) \right],$$

with expectation approximated via posterior sampling. Under a one-step lookahead, observing correctness yields the updated posterior variance

$$\sigma_{k,r,\text{new}}^2(j) = \left(\frac{1}{\sigma_{k,r}^2} + \bar{I}_{kj}^r \right)^{-1},$$

and expected reduction $\Delta \sigma_{kjr}^2 = \sigma_{k,r}^2 - \sigma_{k,r,\text{new}}^2(j)$. Consequently, we define the VarBALD acquisition function as

$$\mathcal{A}^r(q_j) = \text{VarBALD}_{kj}^r = \bar{\pi}_{kj}^r \cdot \text{BALD}_{kj}^r \cdot \left(1 + \alpha \frac{\Delta \sigma_{kjr}^2}{\sigma_{k,r}^2 + \varepsilon} \right),$$

where $\alpha \geq 0$ controls the strength of variance reduction and $\varepsilon > 0$ ensures numerical stability. When $\alpha = 0$, the criterion reduces to DiffBALD.

4.4 RankBALD

We now introduce our primary contribution: an acquisition criterion explicitly aligned with the downstream ranking objective (see Section 3.1). RankBALD redirects acquisition toward expected information gain about whether one model outperforms another. To formalize this objective, we consider a pair of models (s, t) with latent abilities θ_s and θ_t . The ranking outcome for this pair is determined by the binary order variable

$$Z_{s,t}^r = \mathbb{1}[\theta_s^r > \theta_t^r],$$

which is revealed indirectly through their correct responses on candidate items. Item selection is therefore driven by maximization of the expected information gained about pairwise orderings. Ideally, one would measure the mutual information between the ordering variable and the latent abilities themselves, i.e.,

$$I(Z_{s,t}^r; \theta_s, \theta_t \mid p_E^r)$$

However, the latent abilities are not directly observable. Instead, information about the ordering must be obtained through observed correctness outcomes. For a candidate item q_j , let

$$Y_{s,t,j}^r = (U_{sj}^r, U_{tj}^r) \in \{0, 1\}^2$$

denote the joint correctness outcomes of the two models. We therefore select the next item according to

$$q_j^* = \arg \max_{q_j \in \mathcal{Q} \setminus \mathcal{Q}_{r-1}^*} \sum_{(s,t) \in \mathcal{P}} I(Z_{s,t}^r; Y_{s,t,j}^r \mid p_E^r),$$

where $\mathcal{P} = \{(s, t) : 1 \leq s < t \leq K\}$ denotes the set of ordered model pairs. Each mutual information term is estimated via MC sampling from $p_E^r(\theta)$. As in DiffBALD and VarBALD, we incorporate difficulty-aware weighting to discourage comparisons driven by uninformative items. Specifically, we define the kernel

$$\pi_{kj}^r(\theta) = \exp\left(-\frac{1}{2} \frac{(b_j - \theta)^2}{\sigma_{\pi,r}^2}\right), \quad \bar{\pi}_{kj}^r = \mathbb{E}_{\theta \sim p_E^r}[\pi_{kj}^r(\theta)],$$

where the bandwidth is calibrated to the posterior scale via

$$\sigma_{\pi,r} = \text{median}_k \left(\text{Std}_{\theta \sim p_E^r}[\theta_k] \right)$$

Because ranking instability arises primarily from small ability gaps, we further emphasize comparisons between nearby models. Let ξ denote the ranking induced by the checkpoint enumeration, obtained by indexing models in training order, which serves as a fixed proxy for the ability ordering of checkpoints within a family. We define the distance between models s and t as their rank distance:

$$\omega(d) = 1 - \frac{d-1}{(K-1)-1}, \quad d = |\xi(s) - \xi(t)|,$$

so that adjacent models receive maximal weight and distant comparisons are deemphasized. Because ξ is fixed, the weights ω are computed once and held constant across rounds. Combining these components, we define the RankBALD acquisition function as

$$\mathcal{A}^r(q_j) = \text{RankBALD}_j^r = \frac{\sum_{(s,t) \in \mathcal{P}} \omega(|\xi(s) - \xi(t)|) \bar{\pi}_{js}^r \bar{\pi}_{jt}^r I(Z_{s,t}^r; Y_{s,t,j}^r \mid p_E^r)}{\sum_{(s,t) \in \mathcal{P}} \omega(|\xi(s) - \xi(t)|) \bar{\pi}_{js}^r \bar{\pi}_{jt}^r + \varepsilon},$$

where $\varepsilon > 0$ ensures numerical stability. The resulting criterion combines the difficulty kernel of DiffBALD, now with a posterior-calibrated bandwidth, with a ranking-specific pairwise information objective, yielding an acquisition score that is sensitive to epistemic uncertainty, difficulty alignment, and rank-boundary resolution.

5 Experimental Setup and Evaluation Measures

We evaluate whether acquisition strategies aligned with the evaluation objective improve the efficiency and reliability of active evaluation. To enable a controlled comparison, we adopt the experimental protocol of Hofmann et al. (2025) and replace only the acquisition component with our proposed ones, keeping all other elements unchanged.

Open LLM Leaderboard Benchmarks: The primary experimental suite comprises six language models with publicly available checkpoints (2.8B-65B parameters), each with between 61 and 94 checkpoints, yielding 2,802 checkpoint–benchmark combinations. Each checkpoint is evaluated on six Open LLM Leaderboard benchmarks: ARC Challenge, GSM8K, HellaSwag, MMLU, TruthfulQA, and WinoGrande. In total, the dataset contains over 13 million item-level evaluations. Item parameters are calibrated using separate 2PL IRT models trained on responses from 102 language models drawn from the Open LLM Leaderboard. Evaluated models and their post-trained variants are excluded from calibration to prevent data leakage. The six language models were selected based on the availability of publicly released per-checkpoint evaluation results on the Open LLM Leaderboard, which is a strict

requirement of our experimental protocol: computing the variance and saturation metrics requires a trajectory of ability estimates across training checkpoints. Models such as Qwen and most other frontier systems do not release intermediate checkpoints with leaderboard evaluations and are therefore incompatible with this setting.

AfroBench (African Language Evaluation): To assess generalization beyond English-centric benchmarks, we additionally evaluate our approach on AfroBench (Ojo et al., 2025), a benchmark suite covering multiple African languages. This setting is particularly relevant as it reflects low-resource and cost-sensitive (Dossou et al., 2026) evaluation scenarios, where annotation is often limited and expensive. In such contexts, evaluation efficiency becomes critical, and methods that can reliably characterize model performance with fewer items are especially valuable. AfroBench therefore provides a complementary testbed for assessing whether the benefits of ranking-aligned acquisition extend to multilingual and resource-constrained settings. Details of the adaptation are provided in Appendix A.2.

Benchmark quality is evaluated using the three metrics introduced in Hofmann et al. (2025). For each model $\mathcal{M}_k$, let $x_k^t(Q)$ denote its performance (e.g. accuracy) at checkpoint $t \in \{1, \dots, T\}$ on benchmark Q , where T is the total number of checkpoints:

- **Validity** captures cross-benchmark consistency as the average rank distance across the K models, $\text{Validity}(B_1, B_2) = \frac{1}{K} \sum_{k=1}^K |\text{rank}_{B_1}(\mathcal{M}_k) - \text{rank}_{B_2}(\mathcal{M}_k)|$, evaluated on two capability pairs: ARC Challenge vs. MMLU (knowledge and reasoning) and HellaSwag vs. WinoGrande (commonsense reasoning).
- **Variance** captures checkpoint stability via normalized total variation: $\text{TV}(\mathcal{M}_k, Q) = \frac{T}{T-1} \frac{\sum_{t=1}^{T-1} |x_k^{t+1} - x_k^t|}{|x_k^T - x_k^1|}$, where lower values indicate smoother, more stable trajectories and hence higher evaluation quality.
- **Saturation** is the absolute Spearman rank correlation between the ordered checkpoint sequence $\{1, \dots, T\}$ and the corresponding performance values $\{x_k^1(Q), \dots, x_k^T(Q)\}$, with values close to 1 indicating the benchmark remains discriminative across the considered capability range.

6 Results and Analysis

RankBALD substantially improves evaluation reliability across budgets, benchmarks, and model families (Tables 1, 2, and 3), consistently achieving the lowest rank distance and total variation while attaining the highest rank correlation across nearly all conditions, with the largest gains in the low-budget regime. These results indicate that directly targeting pairwise ordering uncertainty accelerates recovery of the global ranking more efficiently than methods that reduce marginal posterior uncertainty in scalar abilities. We note that the progression $\text{BALD} \rightarrow \text{DiffBALD} \rightarrow \text{VarBALD} \rightarrow \text{RankBALD}$ constitutes a component-wise analysis: comparing DiffBALD against BALD isolates the contribution of difficulty-aware weighting, comparing VarBALD against DiffBALD isolates variance reduction, and comparing RankBALD against VarBALD isolates the effect of replacing scalar ability mutual information with pairwise ordering mutual information, which yields the largest single gain across all conditions.

Across evaluation budgets, Fluid Benchmarking and response-level BALD variants exhibit similar convergence behavior, with Fluid achieving competitive total variation at very small budgets. This is expected, since under the 2PL model BALD is closely related to the item information that Fluid maximizes, so the two criteria tend to select similar items. In contrast, RankBALD stabilizes pairwise ordering early (around 50 items), enabling reliable ranking with fewer evaluation items and yielding lower variance at moderate and large budgets. This indicates that resolving ordering uncertainty early leads to more stable evaluation trajectories. The similar performance of DiffBALD and VarBALD suggests that, once difficulty alignment is incorporated, most of the relevant information gain is already captured by the BALD objective, and explicit variance reduction provides limited additional benefit under the Laplace approximation. The performance gap with RankBALD, therefore, reflects not an improved ability to estimate, but the explicit targeting of pairwise

Method	Validity (rank dist. ↓)				Variance (total var. ↓)				Saturation (rank corr. ↑)			
	10	50	100	500	10	50	100	500	10	50	100	500
Random-IRT	59.45	51.52	45.67	36.09	19.38	16.00	13.77	12.50	0.428	0.642	0.702	0.836
Fluid Benchmarking	46.89	36.91	36.25	34.92	14.30	6.95	6.06	4.94	0.708	0.831	0.850	0.875
BALD	49.36	38.25	36.55	34.94	23.64	9.66	8.09	6.25	0.538	0.779	0.818	0.861
VarBALD	47.56	37.22	36.43	34.77	14.96	8.48	6.82	5.02	0.650	0.811	0.837	0.872
DiffBALD	48.88	37.51	36.37	34.83	15.25	8.34	6.88	5.03	0.657	0.809	0.841	0.873
RankBALD	35.30	29.10	24.23	22.45	12.42	6.09	5.81	3.98	0.728	0.849	0.861	0.877

Table 1: Performance across acquisition strategies as the number of items per benchmark increases. RankBALD yields the strongest gains across validity, variance, and saturation, especially in the low-budget regime.

Method	Validity (rank dist. ↓)					Variance (total var. ↓)					Saturation (rank corr. ↑)				
	ARC	GSM	HS	MMLU	WG	ARC	GSM	HS	MMLU	WG	ARC	GSM	HS	MMLU	WG
Random-IRT	25.48	36.55	15.61	41.39	26.35	7.92	17.50	6.48	22.00	10.44	0.821	0.713	0.819	0.555	0.780
Fluid	12.44	19.31	6.05	42.38	14.64	3.27	9.09	2.04	6.32	5.81	0.951	0.864	0.985	0.668	0.925
BALD	14.55	24.17	6.60	45.80	21.92	3.58	11.85	2.26	7.98	9.90	0.933	0.834	0.979	0.616	0.863
VarBALD	12.97	21.11	6.38	42.87	17.27	3.49	10.03	2.16	6.50	7.85	0.941	0.853	0.982	0.651	0.909
DiffBALD	12.92	21.07	6.36	42.83	16.66	3.45	10.38	2.15	6.69	7.86	0.944	0.855	0.983	0.651	0.918
RankBALD	12.39	19.10	5.53	38.77	12.68	3.07	8.82	2.01	5.55	4.33	0.958	0.877	0.990	0.718	0.934

Table 2: Benchmark-specific evaluation at 100 items. Validity and variance are reported as rank distance and total variation (↓), while saturation is reported as rank correlation (↑). Benchmarks include ARC Challenge (ARC), GSM8K (GSM), HellaSwag (HS), MMLU, and WinoGrande (WG). TruthfulQA is excluded, following the Fluid Benchmarking protocol.

ordering uncertainty. These patterns hold across benchmarks and model families (Tables 2 and 3), indicating that ordering-aligned acquisition provides consistent advantages across tasks and model scales. While the Open LLM Leaderboard does not include any African-language benchmark, we complement our main results with an adaptation of the pipeline to AfroBench in Appendix A.2, using item-level data obtained directly from the AfroBench authors (Ojo et al., 2025). These results are consistent with our main findings: RankBALD outperforms all baselines across validity, variance, and saturation, further supporting the generality of our approach beyond standard English-centric benchmarks.

These results highlight a fundamental distinction between scalar ability estimation and ranking recovery: reducing marginal posterior uncertainty does not necessarily resolve uncertainty near pairwise-ordering boundaries, which determine ranking correctness. By explicitly targeting ordering uncertainty, RankBALD achieves more reliable model ranking without modifying the response model or inference procedure, showing that substantial gains in evaluation efficiency arise from aligning the acquisition objective with the downstream decision. A promising direction for future work is to further concentrate acquisition on comparisons among closely competing models, where ordering uncertainty is most consequential.

7 Conclusion

We argued that adaptive evaluation should be aligned with the decision it is intended to support. In particular, ranking is determined by pairwise ordering boundaries rather than the precision of individual ability estimates, and reducing the latter does not necessarily resolve the former. Building on this insight, we introduced a family of Bayesian acquisition criteria for IRT-based active evaluation. DiffBALD and VarBALD extend BALD to the IRT setting with difficulty-aware weighting and variance-reduction components, while RankBALD directly targets pairwise ordering uncertainty by maximizing mutual information about whether one model outperforms another. Across 2,802 checkpoint-benchmark combinations on six Open LLM Leaderboard benchmarks, and in an additional adaptation to AfroBench spanning five African-language tasks, RankBALD consistently achieves the strongest ranking validity, lowest variance, and fastest saturation, without modifying the underlying response model or inference procedure. More broadly, these results suggest that

Method	Validity (rank dist. ↓)						Variance (total var. ↓)						Saturation (rank corr. ↑)					
	A-7B	K2-65B	O1-7B	O2-7B	P-3B	P-7B	A-7B	K2-65B	O1-7B	O2-7B	P-3B	P-7B	A-7B	K2-65B	O1-7B	O2-7B	P-3B	P-7B
Random-IRT	39.74	31.77	55.58	44.48	51.95	51.10	11.47	15.29	9.37	13.50	19.99	12.98	0.672	0.660	0.775	0.630	0.713	0.759
Fluid	24.75	28.65	41.61	41.86	35.22	36.13	5.54	7.14	5.83	6.83	6.55	4.46	0.819	0.892	0.905	0.802	0.813	0.868
BALD	27.34	28.53	49.47	50.30	36.63	39.72	5.77	13.20	6.43	9.25	7.33	6.57	0.788	0.838	0.886	0.734	0.811	0.853
VarBALD	26.69	27.58	42.83	41.92	35.17	38.31	5.47	9.41	6.05	8.09	7.06	4.86	0.802	0.864	0.895	0.786	0.813	0.863
DiffBALD	26.88	27.85	43.01	41.27	35.90	37.75	5.45	9.36	6.17	8.08	7.22	4.99	0.810	0.871	0.898	0.790	0.814	0.865
RankBALD	23.28	28.38	40.48	38.67	31.58	35.09	4.91	6.03	5.08	5.43	4.79	4.03	0.822	0.897	0.908	0.808	0.839	0.876

Table 3: Language-model-specific evaluation at 100 items. Validity and variance are reported as rank distance and total variation ($\downarrow$), while saturation is reported as rank correlation ($\uparrow$). Model abbreviations are: A-7B = Amber-7B, K2-65B = K2-65B, O1-7B = OLMo1-7B, O2-7B = OLMo2-7B, P-3B = Pythia-2.8B, and P-7B = Pythia-6.9B.

benchmark construction should be viewed as an inference problem over model capabilities, where the acquisition objective is defined with respect to the intended evaluation decision. The gains demonstrated here arise purely from this alignment, pointing toward a general principle: active evaluation methods should be designed around what practitioners actually want to know, not only around what is easiest to quantify.

8 Limitations

Several limitations of the current work are worth noting. First, our approach relies on pre-fitted IRT parameters, which assumes a fixed item bank and may not generalize to zero-shot evaluation settings where item parameters are unavailable. Second, the Laplace approximation used to represent posterior uncertainty may underestimate tail uncertainty in early acquisition rounds when few items have been observed. Third, item selection in our framework is sequential and greedy, selecting one item at a time without accounting for batch-level diversity or redundancy. A natural extension is to move toward batch acquisition, drawing on BatchBALD (Kirsch et al., 2019), which accounts for the joint information gain of a set of items rather than summing individual gains greedily. Adapting this to the ranking objective, selecting a batch of items that jointly maximize mutual information over pairwise ordering variables, is a principled direction for future work, connecting directly to settings where evaluation items must be selected in advance for a fixed evaluation round.

9 Code Release

Our source code is publicly available at https://github.com/bonaventuredossou/al_rankbald.

References

- Edward Beeching, Cl  mentine Fourier, Nathan Habib, Sheon Han, Nathan Lambert, Nazneen Rajani, Omar Sanseviero, Lewis Tunstall, and Thomas Wolf. Open LLM leaderboard. <https://huggingface.co/spaces/open-llm-leaderboard-old/open-llm-leaderboard>, 2023.
- Allan Birnbaum. Statistical theory for logistic mental test models with a prior distribution of ability. *Journal of Mathematical Psychology*, 6(2):258–276, 1969.
- Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, New York, 2006. ISBN 978-0-387-31073-2.
- Hua-Hua Chang. Psychometrics behind computerized adaptive testing. *Psychometrika*, 80(1):1–20, 2015.
- Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try ARC, the AI2 reasoning challenge. *arXiv Preprint arXiv:1803.05457*, 2018.

- Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv Preprint arXiv:2110.14168*, 2021.
- David A. Cohn, Zoubin Ghahramani, and Michael I. Jordan. Active learning with statistical models. *Journal of Artificial Intelligence Research*, 4:129–145, 1996.
- Harald Cramér. *Mathematical Methods of Statistics (PMS-9)*. Princeton University Press, 1999. ISBN 9780691005478. URL <http://www.jstor.org/stable/j.ctt1bpm9r4>.
- Christine DeMars. *Item Response Theory*. Oxford University Press, 2010.
- Bonaventure F. P. Dossou, Ines Arous, Audrey Durand, and Jackie Chi Kit Cheung. Active learning with non-uniform costs for African natural language processing. In Vera Demberg, Kentaro Inui, and Lluís Marquez (eds.), *Findings of the Association for Computational Linguistics: EACL 2026*, pp. 6644–6656, Rabat, Morocco, March 2026. Association for Computational Linguistics. ISBN 979-8-89176-386-9. doi: 10.18653/v1/2026.findings-eacl.349. URL <https://aclanthology.org/2026.findings-eacl.349/>.
- Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the 33rd International Conference on Machine Learning*, pp. 1050–1059, 2016.
- Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep bayesian active learning with image data. In *International Conference on Machine Learning*, pp. 1183–1192. PMLR, 2017.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.
- Valentin Hofmann, David Heineman, Ian Magnusson, Kyle Lo, Jesse Dodge, Maarten Sap, Pang Wei Koh, Chun Wang, Hannaneh Hajishirzi, and Noah A. Smith. Fluid language model benchmarking. In *Second Conference on Language Modeling*, 2025.
- Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. Bayesian active learning for classification and preference learning. In *Advances in Neural Information Processing Systems*, pp. 1476–1484, 2011.
- Alex Kipnis, Konstantinos Voudouris, Luca M. Schulze Buschoff, and Eric Schulz. metabench - a sparse benchmark of reasoning and knowledge in large language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=4T33izzFpK>.
- Andreas Kirsch, Joost Van Amersfoort, and Yarin Gal. BatchBALD: Efficient and diverse batch acquisition for deep bayesian active learning. In *Advances in Neural Information Processing Systems*, volume 32, 2019.
- John P. Lalor, Pedro Rodriguez, João Sedoc, and Jose Hernandez-Orallo. Item response theory for natural language processing. In Mohsen Mesgar and Sharid Loáiciga (eds.), *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Tutorial Abstracts*, pp. 9–13, St. Julian’s, Malta, March 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.eacl-tutorials.2. URL <https://aclanthology.org/2024.eacl-tutorials.2/>.
- Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3214–3252, 2022.
- Frederic Lord. A theory of test scores. *Psychometric Monographs*, 1952.
- David J. C. MacKay. Bayesian interpolation. *Neural Computation*, 4(3):415–447, 1992.
- Rob R. Meijer and Michael L. Nering. Computerized adaptive testing: Overview and introduction. *Applied Psychological Measurement*, 23(3):187–194, 1999.

Kevin P. Murphy. *Machine Learning: A Probabilistic Perspective*. MIT Press, Cambridge, MA, 2012a. ISBN 978-0-262-01802-9.

Kevin P. Murphy. *Machine Learning: A Probabilistic Perspective*. MIT Press, 2012b.

Radford M. Neal. *Bayesian Learning for Neural Networks*, volume 118 of *Lecture Notes in Statistics*. Springer, 1996.

Jessica Ojo, Odunayo Ogundepo, Akintunde Oladipo, Kelechi Ogueji, Jimmy Lin, Pontus Stenetorp, and David Ifeoluwa Adelani. AfroBench: How good are large language models on African languages? In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 19048–19095, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.976. URL <https://aclanthology.org/2025.findings-acl.976/>.

Felipe Maia Polo, Lucas Weber, Leshem Choshen, Yuekai Sun, Gongjun Xu, and Mikhail Yurochkin. tinybenchmarks: evaluating llms with fewer examples. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

C. Radhakrishna Rao. *Information and the Accuracy Attainable in the Estimation of Statistical Parameters*, pp. 235–247. Springer New York, New York, NY, 1992. ISBN 978-1-4612-0919-5. doi: 10.1007/978-1-4612-0919-5_16. URL https://doi.org/10.1007/978-1-4612-0919-5_16.

Pedro Rodriguez, Joe Barrow, Alexander Hoyle, John P. Lalor, Robin Jia, and Jordan Boyd-Graber. Evaluation examples are not equally informative: How should that change NLP leaderboards? In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 4486–4503, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.346. URL <https://aclanthology.org/2021.acl-long.346/>.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. WinoGrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9): 99–106, 2021.

Burr Settles. Active learning literature survey. 2009.

Wei Song, Zhenya Huang, Cheng Cheng, Weibo Gao, Bihan Xu, GuanHao Zhao, Fei Wang, and Runze Wu. IRT-router: Effective and interpretable multi-LLM routing via item response theory. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15629–15644, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.761. URL <https://aclanthology.org/2025.acl-long.761/>.

Wim J. van der Linden and Ronald K. Hambleton (eds.). *Handbook of Modern Item Response Theory*. Springer, 1997. doi: 10.1007/978-1-4757-2691-6. URL <https://doi.org/10.1007/978-1-4757-2691-6>.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: Can a machine really finish your sentence? In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4791–4800, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1472. URL <https://aclanthology.org/P19-1472/>.

A Appendix

A.1 From BALD to RankBALD: A Step-by-Step Derivation

This section provides a self-contained derivation bridging the original BALD formulation (Houlsby et al., 2011) (Section 2) to RankBALD (Section 4.4), intended for a reader already familiar with BALD but not with the intermediate reasoning behind its adaptation here. We proceed in six steps: (0) restating BALD in its original setting, (1) mapping every BALD symbol onto its active-evaluation counterpart, (2) re-deriving BALD_{kj}^r as a substitution, and (3-5) introducing DiffBALD, VarBALD, and RankBALD as successive, explicitly isolated additions, each accompanied by the exact estimator used in our implementation.

In classical active learning, once a label y is obtained for x , the pair (x, y) is added to $\mathcal{D}$ and the model is *retrained*: ω itself changes. In active evaluation, this does not happen. The evaluated models $\mathcal{M}_k$ are frozen throughout the entire procedure; their weights w_k never change. What updates after each acquisition round is not any model, but the *evaluator’s belief* $p_E^r(\theta)$ about the models’ fixed, already-determined abilities. Every quantity introduced below should be read with this distinction in mind: we are not training a model to solve a task; we are refining an external observer’s estimate of a fixed target.

Step 0: BALD in its original setting: Houlsby et al. (2011) consider a single model being trained, with parameters ω , a labeled set $\mathcal{D}$ collected so far, and a pool of unlabeled candidates $x \in \mathcal{X}$. Querying x would reveal a predicted output y , with $y \perp \mathcal{D} \mid \omega$: given the parameters, the prediction depends only on the input, not on the training history. The learner maintains a posterior $p(\omega \mid \mathcal{D})$ and selects

$$x^* = \arg \max_{x \in \mathcal{X}} I(y; \omega \mid \mathcal{D}) = H[y \mid \mathcal{D}] - \mathbb{E}_{p(\omega \mid \mathcal{D})} [H[y \mid \omega]]$$

The first term is the entropy of the *posterior-marginalized* prediction: how uncertain is y once we average over everything we currently believe about ω . The second term is the *expected* entropy conditioned on a specific, single value of ω : how uncertain is y once we commit to one particular hypothesis. Their difference is large exactly where plausible hypotheses disagree with one another.

Step 1: mapping every symbol: Table 4 gives the full correspondence; we justify each row in turn.

Table 4: Notational correspondence between BALD and active evaluation.

BALD (Houlsby et al., 2011)	Active evaluation (this paper)
$x \in \mathcal{X}$ (candidate input)	$q_j \in \mathcal{Q}$ (candidate item)
y (predicted output)	U_{kj} (model k ’s response to item j)
ω (model parameters)	θ_k (model ability)
$\mathcal{D}$ (training data)	$\mathcal{Q}_r^*$ (items acquired by round r)
$p(\omega \mid \mathcal{D})$ (posterior)	$p_E^r(\theta)$ (evaluator’s belief)
$I(y; \omega \mid \mathcal{D})$	$\text{BALD}_{kj}^r = I(U_{kj}; \theta_k)$

Structure-preserving substitutions: $x \rightarrow q_j$ and $\mathcal{D} \rightarrow \mathcal{Q}_r^*$ are literal: both are “the next thing to query” and “everything queried so far,” respectively. Likewise $p(\omega \mid \mathcal{D}) \rightarrow p_E^r(\theta)$: both are a belief over an unknown, conditioned on what has been observed.

Parameters to ability: a forced dimensionality reduction: In BALD, ω is the full parameter vector of the model being trained, potentially of very high dimension. In active evaluation we are not recovering weights; we are summarizing capability. The 2PL model (Section 2) already supplies a principled scalar summary, θ_k , so the object of inference collapses from ω to one real number per model. This is forced by the evaluation goal, not a free choice.

Predicted output to response: same role, different form: y and U_{kj} play the same role in their respective objectives: both are the observable outcome whose distribution depends on the

input and the parameter, and both satisfy the same conditional-independence structure relative to the accumulated data, i.e., $U_{kj} \perp Q_r^* \mid \theta_k$, mirroring $y \perp \mathcal{D} \mid \omega$ above. This shared independence structure is exactly what makes $H[U_{kj} \mid Q_r^*]$ and $H[U_{kj} \mid \theta_k]$ meaningfully different quantities, just as it does for y in the original formulation. What differs is their *form*: y 's distribution $p(y \mid x, \omega)$ is an arbitrary function computed by the trained network, with no closed form in general. U_{kj} 's distribution is fixed and known in closed form by construction:

$$U_{kj} \sim \text{Bernoulli}(p_{kj}(\theta)), \quad p_{kj}(\theta) := p(U_{kj} = 1 \mid \theta) = \sigma(a_j(\theta - b_j))$$

(Eq. 3). We write $p_{kj}(\theta)$ as shorthand for this conditional probability and reuse it in every acquisition criterion below.

The model index k : a structural addition with no counterpart: BALD is written for one model in training. Active evaluation tracks a population of K models at once, so every quantity below carries an index k that simply has no analogue in the original formulation, since BALD never needed to handle more than one model simultaneously.

Step 2: re-deriving BALD_{kj}^r : Substituting every mapped symbol from Step 1 into the Step 0 objective gives

$$q_j^* = \arg \max_{q_j \in \mathcal{Q} \setminus Q_r^*} I(U_{kj}; \theta_k \mid Q_r^*) = H[U_{kj} \mid Q_r^*] - \mathbb{E}_{p_E^r(\theta)} [H[U_{kj} \mid \theta]].$$

Because U_{kj} is Bernoulli with parameter $p_{kj}(\theta)$, both terms admit closed forms. The predictive entropy of the posterior-marginalized prediction is the entropy of the mixture probability, $H[U_{kj} \mid Q_r^*] = H(\mathbb{E}_{p_E^r(\theta)} [p_{kj}(\theta)])$; the expected conditional entropy is the expectation of the entropy of each individual Bernoulli, $\mathbb{E}_{p_E^r(\theta)} [H[U_{kj} \mid \theta]] = \mathbb{E}_{p_E^r(\theta)} [H(p_{kj}(\theta))]$. Substituting both gives exactly the paper's Section 4.1 objective:

$$\text{BALD}_{kj}^r = H(\mathbb{E}_{p_E^r(\theta)} [p_{kj}(\theta)]) - \mathbb{E}_{p_E^r(\theta)} [H(p_{kj}(\theta))],$$

approximated via the same Monte Carlo sampling from the posterior that BALD itself uses (Appendices A.6, A.8). No new modeling assumption has been introduced in this step: every term is a direct substitution of a BALD term via Step 1's correspondence, followed by the standard Bernoulli-entropy identity.

A remark on the first term. One might expect the predictive entropy to be written $H[p_{kj}(\theta)]$, but this is not a scalar: θ is a random variable under p_E^r , so $p_{kj}(\theta)$ is itself random and $H[p_{kj}(\theta)]$ is a random quantity rather than a number. The predictive entropy is by definition the entropy of the posterior predictive distribution, $p(U_{kj} = 1 \mid Q_r^*) = \int p_{kj}(\theta) p_E^r(\theta) d\theta = \bar{p}_{kj}^r$, so the marginalization is forced by the definition of the term rather than being an additional Bayesian modeling choice. Note that the two terms of BALD_{kj}^r contain the same object $p_{kj}(\theta)$ and differ only in whether H is applied before or after the expectation over p_E^r . Their difference is non-negative because H is concave, and vanishes exactly when $p_{kj}(\theta)$ is constant over the posterior. If p_E^r were replaced by a point mass at the MAP, both terms would collapse to $H(p_{kj}(\hat{\theta}_k^r))$ and the criterion would be identically zero: the entire epistemic signal comes from the order in which H and $\mathbb{E}$ are applied.

Step 3: DiffBALD, the first new component: BALD_{kj}^r is already sensitive to where an item sits on the 2PL difficulty scale, since the response probability is most responsive to ability when b_j is near the region of posterior mass. Expanding the entropy about $\bar{p}_{kj}^r$ and applying the delta method gives, for small posterior variance, $\text{BALD}_{kj}^r \approx \frac{1}{2} \sigma_{k,r}^2 a_j^2 p_{kj}(\mu)(1 - p_{kj}(\mu))$, where μ denotes the posterior mean: the item information of Section 2 scaled by $\sigma_{k,r}^2$, which is maximized at $b_j = \mu$. This difficulty preference is inherited from the IRT layer introduced in Step 1 and is absent from generic BALD, where no notion of item difficulty exists. What

it does not provide is control over the width of that preference: the peak in b_j has extent of order $1/a_j$, set by the item's own discrimination. We therefore introduce a difficulty kernel, with no BALD counterpart, that imposes a bandwidth on the same axis:

$$\pi_{kj}(\theta) = \exp\left(-\frac{1}{2}(b_j - \theta)^2\right), \quad \bar{\pi}_{kj}^r = \frac{1}{T} \sum_{t=1}^T \pi_{kj}(\theta_k^{(t,r)}),$$

$$\text{DiffBALD}_{kj}^r = \bar{\pi}_{kj}^r \cdot \text{BALD}_{kj}^r$$

This kernel is a design choice rather than a derived quantity: it is chosen for its shape, peaked at $\theta = b_j$ and smoothly decaying, not obtained from an optimality argument. What is derivable is its expectation under a Gaussian posterior $\theta \sim \mathcal{N}(\mu, \sigma^2)$. Combining the exponents of $\pi_{kj}(\theta)$ and the Gaussian density and completing the square in θ gives, after simplification,

$$\mathbb{E}_\theta[\pi_{kj}(\theta)] = \frac{1}{\sqrt{1 + \sigma^2}} \exp\left(-\frac{(b_j - \mu)^2}{2(1 + \sigma^2)}\right),$$

the closed form used in Appendix A.7 to justify the clipping bounds $\sigma_{\min}^2, \sigma_{\max}^2$. In our implementation, this expectation is nonetheless approximated by Monte Carlo averaging over $\theta_k^{(t,r)}$ rather than evaluated in closed form, consistent with how every other posterior expectation in this paper is computed.

Step 4: VarBALD, addressing what DiffBALD still misses: DiffBALD's mutual-information term captures pointwise disagreement about a specific response but does not explicitly track the residual *width* of the posterior after that response is observed, a quantity BALD also never targets directly, since it relies on the MI term alone as its full measure of uncertainty. Unlike the DiffBALD kernel, both components introduced here are derivable from first principles rather than chosen by design.

Fisher information under the 2PL model: Write the Bernoulli log-likelihood of a single response, $\ell(\theta) = U_{kj} \log p_{kj}(\theta) + (1 - U_{kj}) \log(1 - p_{kj}(\theta))$. Using the logistic derivative identity $\sigma'(z) = \sigma(z)(1 - \sigma(z))$ and the chain rule through $p_{kj}(\theta) = \sigma(a_j(\theta - b_j))$, one obtains $\ell'(\theta) = a_j(U_{kj} - p_{kj}(\theta))$ and, differentiating once more, $\ell''(\theta) = -a_j^2 p_{kj}(\theta)(1 - p_{kj}(\theta))$. Since this does not depend on U_{kj} , the observed and expected Fisher information coincide:

$$\mathcal{I}(\theta) = \mathbb{E}[-\ell''(\theta)] = a_j^2 p_{kj}(\theta)(1 - p_{kj}(\theta)),$$

which is exactly the item-information formula of Section 2. The posterior-averaged quantity used by VarBALD is this same expression, averaged over posterior samples of θ rather than evaluated at a point:

$$\sigma_{k,r}^2 = \text{Var}_{\theta \sim p_E^r}[\theta], \quad \bar{I}_{kj}^r = \frac{1}{T} \sum_{t=1}^T a_j^2 p_{kj}(\theta_k^{(t,r)})(1 - p_{kj}(\theta_k^{(t,r)}))$$

The one-step lookahead update: Under a local Gaussian approximation, the current log-posterior has curvature (precision) $1/\sigma_{k,r}^2$. If item j were hypothetically observed, the updated log-posterior is the current log-posterior plus the new log-likelihood term; taking the second derivative at the mode and using precision additivity of independent quadratic contributions gives updated precision $1/\sigma_{k,r}^2 + \bar{I}_{kj}^r$. Inverting yields

$$\sigma_{k,r,\text{new}}^2(j) = \left(\frac{1}{\sigma_{k,r}^2} + \bar{I}_{kj}^r\right)^{-1}, \quad \Delta\sigma_{kj}^2 = \sigma_{k,r}^2 - \sigma_{k,r,\text{new}}^2(j),$$

the same precision-additivity logic already used in Appendix A.6 to accumulate Fisher information over the full acquired set Q_r^* , applied here hypothetically to a single candidate item. Combining both derived quantities with the DiffBALD kernel from Step 3 gives

$$\text{VarBALD}_{kj}^r = \bar{\pi}_{kj}^r \cdot \text{BALD}_{kj}^r \cdot \left(1 + \alpha \frac{\Delta\sigma_{kj}^2}{\sigma_{k,r}^2 + \varepsilon}\right)$$

In our implementation, $\sigma_{k,r}^2$ is computed as the empirical variance of $\theta_k^{(t,r)}$ samples drawn from the entropy-adaptive broadened posterior (Appendix A.8), rather than the Laplace variance of Appendix A.6 directly. Since the broadened sampling distribution is a two-component Gaussian mixture with tail multiplier $\kappa_{k,r} \in [\kappa_{\min}, \kappa_{\max}]$ sharing the mode $\hat{\theta}_k^r$, its variance is $\frac{1}{2}(1 + \kappa_{k,r}^2)$ times the underlying Laplace variance, so $\sigma_{k,r}^2$ as used in VarBALD is systematically larger than the point-curvature estimate, reflecting the same broadened uncertainty used throughout posterior sampling, in the paper.

Step 5: RankBALD, a change in what is being estimated: Every criterion through VarBALD still targets the same *kind* of quantity BALD targets: information about a per-model scalar θ_k , the low-dimensional descendant of ω from Step 1. RankBALD changes the question itself, from “how much does item j tell me about θ_k ” to “how much does item j tell me about whether model s outperforms model t .” This requires an object with no BALD analogue at all: the pairwise order variable

$$Z_{s,t}^r = \mathbb{1}[\theta_s^r > \theta_t^r]$$

The ideal quantity, $I(Z_{s,t}^r; \theta_s, \theta_t \mid p_E^r)$, is inaccessible because θ_s, θ_t are latent, so we instead measure information gained through the *observable* joint responses $Y_{s,t,j}^r = (U_{s,j}^r, U_{t,j}^r) \in \{0, 1\}^2$. Unlike BALD_{kj}^r in Step 2, which exploits the closed-form Bernoulli entropy to compute its objective analytically from posterior samples of θ alone, the pairwise mutual information term is estimated by a *plug-in (empirical histogram) estimator* over fully sampled discrete outcomes. Concretely, for $t = 1, \dots, T$: draw $\theta_s^{(t,r)} \sim p_E^r(\theta_s)$ and $\theta_t^{(t,r)} \sim p_E^r(\theta_t)$ (independent per-model posteriors, as established in Step 1); compute

$$Z^{(t,r)} = \mathbb{1}[\theta_s^{(t,r)} > \theta_t^{(t,r)}];$$

compute the response probabilities under the 2PL model and, critically, *sample* binary responses from them rather than using the probabilities directly:

$$p_{s,j}^{(t,r)} = \sigma(a_j(\theta_s^{(t,r)} - b_j)), \quad U_{s,j}^{(t,r)} \sim \text{Bernoulli}(p_{s,j}^{(t,r)}),$$

and likewise for $U_{t,j}^{(t,r)}$, giving $Y_{s,t,j}^{(t,r)} = (U_{s,j}^{(t,r)}, U_{t,j}^{(t,r)})$ as a sampled realization. This yields T i.i.d. triples $(Z^{(t,r)}, Y_{s,t,j}^{(t,r)})$ over the $2 \times 4 = 8$ possible outcomes of $(Z, Y) \in \{0, 1\} \times \{0, 1\}^2$. The empirical joint distribution is formed by frequency count,

$$\hat{P}(Z = z, Y = y) = \frac{1}{T} \sum_{t=1}^T \mathbb{1}[Z^{(t,r)} = z, Y_{s,t,j}^{(t,r)} = y], \quad z \in \{0, 1\}, y \in \{0, 1\}^2,$$

with marginals $\hat{P}(Z = z) = \sum_y \hat{P}(z, y)$ and $\hat{P}(Y = y) = \sum_z \hat{P}(z, y)$, and mutual information is estimated by its definition applied directly to this empirical joint:

$$\hat{I}(Z_{s,t}^r; Y_{s,t,j}^r \mid p_E^r) = \max \left(0, \sum_z \sum_y \hat{P}(z, y) \log \frac{\hat{P}(z, y) + \varepsilon}{\hat{P}(z) \hat{P}(y) + \varepsilon} \right)$$

This plug-in estimator trades the analytic entropy decomposition used in Step 2 for a fully empirical one over the joint 8-outcome space, avoiding a more involved nested-conditioning derivation at the cost of additional Monte Carlo variance from sampling $U_{s,j}, U_{t,j}$ directly.

As in DiffBALD and VarBALD, we incorporate difficulty-aware weighting, now with a posterior-calibrated bandwidth $\sigma_{\pi,r}$ rather than the fixed bandwidth of Step 3:

$$\sigma_{\pi,r} = \text{median}_k \left(\text{Std}_{\theta \sim p_E^r}[\theta_k] \right), \quad \pi_{kj}^r(\theta) = \exp \left(-\frac{1}{2} \frac{(b_j - \theta)^2}{\sigma_{\pi,r}^2} \right), \quad \bar{\pi}_{kj}^r = \mathbb{E}_{\theta \sim p_E^r} [\pi_{kj}^r(\theta)]$$

Because ranking instability arises primarily from small ability gaps, we further weight pairs by their current rank distance, $\omega(d) = 1 - \frac{d-1}{(K-1)-1}$ with $d = |\tilde{\zeta}(s) - \tilde{\zeta}(t)|$, so adjacent

models receive maximal weight. Combining the plug-in MI estimator, the recalibrated difficulty kernel, and the rank-distance weight over all model pairs $\mathcal{P} = \{(s, t) : 1 \leq s < t \leq K\}$ gives the full RankBALD acquisition function:

$$\text{RankBALD}_j^r = \frac{\sum_{(s,t) \in \mathcal{P}} \omega(|\tilde{\zeta}(s) - \tilde{\zeta}(t)|) \tilde{\pi}_{js}^r \tilde{\pi}_{jt}^r \hat{I}(Z_{s,t}^r; Y_{s,t,j}^r | p_E^r)}{\sum_{(s,t) \in \mathcal{P}} \omega(|\tilde{\zeta}(s) - \tilde{\zeta}(t)|) \tilde{\pi}_{js}^r \tilde{\pi}_{jt}^r + \varepsilon}$$

Table 5 consolidates the six steps above.

Table 5: Provenance of components across the derivation (Section 4).

Component	Provenance
Mutual-information objective	Inherited from BALD (Step 0)
Entropy decomposition	Inherited from BALD (Step 0, 2)
Monte Carlo posterior sampling	Inherited from BALD (Step 0)
Scalar ability θ_k (vs. full ω)	Required by evaluation setting (Step 1)
Frozen models / evaluator-only updates	Required by evaluation setting
Round-indexed evaluator posterior $p_E^r(\theta)$	Required by evaluation setting (Step 1)
IRT response model (a_j, b_j)	Required by evaluation setting (Step 1)
Difficulty-alignment kernel $\pi_{kj}(\theta)$	Novel to RankBALD (Step 3)
Fisher information / lookahead variance	Novel to RankBALD (Step 4)
Pairwise ordering variable $Z_{s,t}^r$	Novel to RankBALD (Step 5)
Plug-in MI estimator over (Z, Y)	Novel to RankBALD (Step 5)
Rank-distance weighting $\omega(\cdot)$	Novel to RankBALD (Step 5)

A.2 AfroIRT-Bench: Adapting Fluid Benchmarking to African-Language Evaluation

Motivation: The Open LLM Leaderboard (Beeching et al., 2023), the de facto standard for comparing frontier language models, does not include any African-language benchmark in its evaluation suite. This makes it impossible to directly obtain the structured, per-item response data that the Fluid Benchmarking pipeline requires for a broad set of models in this domain. As an alternative, we contacted the authors of AfroBench (Ojo et al., 2025), a comprehensive multilingual benchmark covering African languages, to obtain item-level model response data. Due to the practical constraints of running large-scale evaluations, the authors were able to provide results for five models: gpt-5-2025-08-07, gemini-3-pro-preview, gemini-2.0-flash, gemma-3-27b-it, and Qwen3.5-27B. These five models form the basis of the experiments reported in this appendix.

Model families as pseudo-checkpoints: The original Fluid Benchmarking framework (Hofmann et al., 2025) was designed for language models evaluated across *training checkpoints*, where each checkpoint represents a snapshot of the same model at a different stage of training, producing a natural trajectory of improving ability scores. AfroBench, by contrast, consists of *final-checkpoint* evaluations of distinct frontier models. To faithfully adapt the pipeline to this setting, we group models into **model families** and treat members of each family as pseudo-checkpoints ordered by increasing capability. This mirrors the intuition behind training checkpoints: just as a model improves through training, larger or newer members of the same model family consistently outperform smaller or older ones. Concretely, we define three families:

- **Gemini/Gemma** ($S = 3$): gemma-3-27b-it $\rightarrow$ gemini-2.0-flash $\rightarrow$ gemini-3-pro-preview, ordered from weakest to strongest. Gemma is the open-source counterpart of the Gemini model family, making this a natural capability progression within a single lineage
- **GPT** ($S = 1$): gpt-5-2025-08-07 (single snapshot)
- **Qwen** ($S = 1$): Qwen3.5-27B (single snapshot)

Measure	Method	Items per benchmark				Benchmark (100 items)					LM family (100 items)		
		10	50	100	500	GSM	MMLU	XNLI	Bele	SIB	Gem/Gemma	GPT	Qwen
Validity <i>Rank dist. ↓</i>	Random-IRT	0.900	0.800	0.600	0.800	0.400	0.800	0.000	0.400	0.400	0.800	0.400	0.000
	Fluid Benchmarking	0.900	0.800	0.800	0.700	0.200	0.200	0.400	0.200	0.200	0.667	0.000	0.000
	BALD	0.700	0.600	1.000	1.200	0.800	0.600	0.200	1.200	0.200	1.467	0.000	0.400
	VarBALD	0.800	0.800	0.800	0.800	0.000	0.000	0.000	0.000	0.400	0.133	0.400	0.000
	DiffBALD	0.600	0.800	0.800	0.800	0.000	0.000	0.000	0.000	0.000	0.400	0.400	0.000
	RankBALD	0.667	0.333	0.333	0.333	0.000	0.000	0.000	0.000	0.000	0.133	—	—
Variance <i>Total var. ↓</i>	Random-IRT	5.076	5.076	5.076	5.076	1.500	1.500	1.500	2.814	1.500	1.763	—	—
	Fluid Benchmarking	2.024	2.024	2.024	2.024	1.500	1.500	1.500	5.441	1.500	2.288	—	—
	BALD	1.875	1.875	1.875	1.875	1.553	1.500	1.500	1.500	1.500	1.511	—	—
	VarBALD	5.768	5.768	5.768	5.768	1.500	1.500	1.500	4.831	1.500	2.166	—	—
	DiffBALD	8.028	8.028	8.028	8.028	1.500	1.500	1.500	4.652	1.500	2.130	—	—
	RankBALD	1.455	1.455	1.455	1.455	1.500	1.500	1.500	1.500	1.500	1.456	—	—
Saturation <i>Rank corr. ↑</i>	Random-IRT	0.703	0.703	0.703	0.703	1.000	1.000	1.000	0.500	1.000	0.900	—	—
	Fluid Benchmarking	0.677	0.677	0.677	0.677	1.000	1.000	1.000	0.500	1.000	0.900	—	—
	BALD	0.637	0.637	0.637	0.637	0.500	1.000	1.000	1.000	1.000	0.900	—	—
	VarBALD	0.450	0.450	0.450	0.450	1.000	1.000	1.000	0.500	1.000	0.900	—	—
	DiffBALD	0.597	0.597	0.597	0.597	1.000	1.000	1.000	0.500	1.000	0.900	—	—
	RankBALD	0.740	0.740	0.740	0.740	1.000	1.000	1.000	1.000	1.000	0.900	—	—

Table 6: Evaluation of acquisition strategies on AfroBench under three experimental axes: (1) varying the number of items per benchmark, (2) per-benchmark performance at a fixed evaluation size of 100 items, and (3) per-model-family performance at the same fixed size. Validity and variance are reported as rank distance and total variation, respectively ($\downarrow$ lower is better); saturation is reported as rank correlation ($\uparrow$ higher is better). Bold indicates the best value in each column. “—” denotes undefined values for singleton families ($S = 1$), where variance and saturation require at least two pseudo-checkpoints. **Model families:** Gemini/Gemma comprises gemma-3-27b-it $\rightarrow$ gemini-2.0-flash $\rightarrow$ gemini-3-pro-preview ($S = 3$, ordered by capability); GPT (gpt-5-2025-08-07) and Qwen (Qwen3.5-72B) are singletons ($S = 1$). **Benchmarks:** GSM = AfriGSM, MMLU = AfriMMLU, XNLI = AfriXNLI, Bele = Belebele, SIB = SIB-200.

Singleton families ($S = 1$) contribute to validity evaluation but yield undefined variance and saturation scores, since both metrics require a trajectory of at least two points. These cases are reported as “—” in Table 6.

Benchmarks: We evaluate across five African-language benchmarks: AfriGSM (GSM), AfriMMLU (MMLU), AfriXNLI (XNLI), Belebele (Bele), and SIB-200 (SIB), covering mathematical reasoning, knowledge, natural language inference, reading comprehension, and topic classification, respectively. Item-level IRT parameters (discrimination a and difficulty b) are estimated using a 2PL model fitted jointly across all five models, providing the item selection backbone for all active methods.

Results: Table 6 reports performance across three experimental axes: (i) varying the number of items per benchmark, (ii) per-benchmark performance at a fixed evaluation size of 100 items, and (iii) per-model-family performance at the same fixed size.

Validity: RankBALD achieves the lowest rank distance across all item counts (0.333 at 50, 100, and 500 items), all five benchmarks (0.000 on GSM, MMLU, XNLI, Bele, and SIB), and the Gemini/Gemma family (0.133, tied with VarBALD). This indicates that RankBALD’s shared acquisition strategy most faithfully preserves the relative ranking of model families. BALD, by contrast, degrades at larger item counts (1.200 at 500 items), suggesting it overfits to individual ability estimates rather than relative ordering.

Variance: RankBALD achieves the lowest total variation (1.455) across item counts, outperforming BALD (1.875), the previous best baseline. Lower variance indicates that the ability estimates produced by RankBALD are more stable as the number of selected items grows; a desirable property for reliable evaluation. Within the Gemini/Gemma family, RankBALD again achieves the lowest variance (1.456), ahead of BALD (1.511) and the random baselines (1.763-1.773).

Saturation: RankBALD achieves the highest rank correlation (0.740) across item counts, indicating that its ability estimates converge most monotonically toward the true full-benchmark score. This exceeds Random-IRT (0.703), the best baseline, by 0.037. Within the Gemini/Gemma family, all methods, including RankBALD, achieve a saturation score of 0.900; a ceiling effect attributable to the small number of pseudo-checkpoints ($S = 3$) rather than a limitation of RankBALD itself.

Consistency with main results: Despite the sparse evaluation setting, i.e., only five models grouped into three families, with a single multi-member family ($S = 3$), the AfroIRT-Bench results are fully consistent with the main paper findings. RankBALD achieves the lowest rank distance, lowest total variation, and highest rank correlation across virtually all experimental axes (Table 6), mirroring its dominant performance on the primary benchmarks. This consistency across two very different evaluation settings, a well-resourced English-centric leaderboard and a sparse African-language benchmark built from data obtained directly from benchmark authors, strengthens the case for RankBALD as a robust and general-purpose active evaluation strategy. The one exception is the saturation ceiling observed within the Gemini/Gemma family, which is an artifact of the limited number of pseudo-checkpoints available rather than a limitation of the method itself.

Table 7: Statistical significance of results at budget=100. Validity (rank distance ↓), variance (total variation ↓), and saturation (rank correlation ↑) are reported as mean ± SE with 95% CI and paired t-test p-values relative to RankBALD over 6 LM-level observations.

Method	Validity ↓			Variance ↓			Saturation ↑		
	Mean ± SE	95% CI	<i>p</i>	Mean ± SE	95% CI	<i>p</i>	Mean ± SE	95% CI	<i>p</i>
Random-IRT	45.67 ± 3.64	[38.5, 52.8]	0.005	13.77 ± 1.49	[10.9, 16.7]	0.002	0.702 ± 0.024	[0.655, 0.749]	0.0004
Fluid	36.25 ± 2.81	[30.7, 41.8]	0.021	6.06 ± 0.40	[5.3, 6.9]	0.005	0.850 ± 0.017	[0.817, 0.883]	0.066
BALD	36.55 ± 4.03	[28.6, 44.4]	0.017	8.09 ± 1.13	[5.9, 10.3]	0.022	0.818 ± 0.021	[0.778, 0.858]	0.006
VarBALD	36.43 ± 2.85	[30.9, 42.0]	0.015	6.82 ± 0.70	[5.5, 8.2]	0.013	0.837 ± 0.019	[0.800, 0.874]	0.001
DiffBALD	36.37 ± 2.76	[31.0, 41.8]	0.013	6.88 ± 0.68	[5.5, 8.2]	0.010	0.841 ± 0.019	[0.804, 0.878]	0.002
RankBALD	24.23 ± 2.65	[19.0, 29.5]	–	5.05 ± 0.27	[4.5, 5.6]	–	0.861 ± 0.017	[0.828, 0.894]	–

A.3 Statistical Significance of Results

We compute standard errors over 6 LM-level observations via paired t-test at budget=100, providing a measure of the reliability of the reported differences across model families. Results are reported as mean ± SE with 95% confidence intervals and significance relative to RankBALD.

RankBALD’s improvements are statistically significant on validity ($p = 0.021$ vs Fluid) and variance ($p = 0.005$ vs Fluid) at budget=100, confirming that the reported differences reflect genuine signal rather than random fluctuation across model families. On saturation, RankBALD improves significantly over all BALD variants ($p \leq 0.006$) but the gap with Fluid does not reach significance at this budget ($p = 0.066$), consistent with the smaller numerical difference reported in Table 1 (0.861 vs 0.850). These results indicate that the primary advantage of RankBALD lies in its ability to resolve pairwise ordering uncertainty more reliably, as captured by the validity metric, and to produce more stable evaluation trajectories, as captured by variance, rather than in improving monotonic convergence toward the full benchmark score.

A.4 Comparison with Offline Evaluation Methods

Two recent works on efficient LLM evaluation deserve explicit discussion: TinyBenchmarks (Polo et al., 2024) and MetaBench (Kipnis et al., 2025). Both are valuable contributions to the broader goal of reducing evaluation costs, and we cite them here to clarify the relationship to our work. TinyBenchmarks (Polo et al., 2024) selects a fixed small subset of benchmark items offline, prior to any model evaluation, using IRT-based item selection to identify the most informative items across a large pool. MetaBench (Kipnis et al., 2025) aggregates performance across multiple benchmarks to predict leaderboard scores from a small number of items. Both approaches are designed for the offline setting, where

item selection is performed once and applied uniformly across all models. Our setting is fundamentally different: items are selected sequentially and online, conditioned on observed model responses at each acquisition round. This distinction matters because online adaptive selection can exploit information gathered during evaluation to refine subsequent item choices, whereas offline methods must commit to a fixed subset before seeing any responses. Exploring whether elements of these offline approaches, particularly their item selection criteria, could be adapted to the sequential setting is an interesting direction for future work.

A.5 Sensitivity Analysis for Variance Clipping

Appendix A.7 provides a principled derivation of the clipping bounds $\sigma_{\min}^2 = 2$ and $\sigma_{\max}^2 = 6$, tying them to the IRT item discrimination scale rather than treating them as free hyperparameters. A full empirical sensitivity analysis sweeping $\sigma_{\min}^2 \in \{1, 2, 3\}$ and $\sigma_{\max}^2 \in \{4, 6, 9\}$ is left for future work and will be included in an extended version of this paper.

A.6 Posterior Inference

The ability θ_k is treated as a latent random variable with posterior distribution $p_E^r(\theta)$, which forms the probabilistic state used to guide active item selection. To guide active item selection, the evaluator must maintain a probabilistic belief over model abilities rather than a single point estimate. Bayesian acquisition criteria depend on expected reductions in posterior uncertainty, which requires explicit quantification of epistemic uncertainty in the ability estimate. We therefore approximate the posterior locally using a Laplace approximation around the MAP estimate:

$$p_E^r(\theta) \approx \mathcal{N}(\hat{\theta}_k^r, \sigma_{k,r}^2),$$

where the variance is given by the inverse curvature of the log-posterior:

$$\sigma_{k,r}^2 = \left(-\frac{\partial^2}{\partial \theta^2} \log p_E^r(\theta) \Big|_{\theta=\hat{\theta}_k^r} \right)^{-1}$$

Laplace approximations around the MAP estimate are standard in Bayesian generalized linear models, where posterior curvature is approximated using the observed Fisher information (MacKay, 1992; Murphy, 2012b). Under the 2PL likelihood, the negative curvature of the log-likelihood with respect to θ corresponds to

$$\mathcal{I}_k^r(\theta) = \sum_{j \in Q_k^*} a_j^2 p_{kj}(\theta)(1 - p_{kj}(\theta)),$$

yielding the practical estimator $\sigma_{k,r}^2 \approx (\mathcal{I}_k^r(\hat{\theta}_k^r) + \lambda)^{-1}$, where $\lambda > 0$ denotes a fixed prior precision that ensures numerical stability and provides a base level of uncertainty in the absence of sufficient observations. This value is then clipped to $[\sigma_{\min}^2, \sigma_{\max}^2]$ for further stability, providing a tractable uncertainty-aware summary of the evaluator’s posterior belief. We choose $\sigma_{\min}^2 = 2$ and $\sigma_{\max}^2 = 6$, with justification provided in Appendix A.7. Additionally, expectations under $p_E^r(\theta)$ are approximated with Monte Carlo samples from a Gaussian approximation centered at the MAP estimate. To improve robustness under posterior misspecification, we use a broadened sampling distribution; full details are given in Appendix A.8.

A.7 Justification of variance clipping

The Laplace variance $\sigma_k^{2(r)}$ is an inverse-curvature estimate and may become arbitrarily large when the observed items provide little Fisher information, or arbitrarily small when curvature becomes extreme. To ensure numerical stability and non-degenerate acquisition behavior, we clip $\sigma_k^{2(r)}$ to a bounded interval $[\sigma_{\min}^2, \sigma_{\max}^2]$.

We justify the specific choices $\sigma_{\min}^2 = 2$ and $\sigma_{\max}^2 = 6$ by analyzing: **(i)** when BALD collapses under overconfident posteriors and **(ii)** when multiplicative acquisition scores collapse under overly diffuse posteriors.

Lower bound: BALD measures epistemic disagreement through the mutual information between latent ability and future response:

$$\text{BALD}_{kj}^{(r)} = H(\mathbb{E}_{\theta}[p_{kj}(\theta)]) - \mathbb{E}_{\theta}[H(p_{kj}(\theta))], \quad p_{kj}(\theta) = \sigma(a_j(\theta - b_j))$$

This quantity is exactly zero if $p_{kj}(\theta)$ is (almost) constant over posterior samples. A necessary condition for non-degenerate BALD is therefore that posterior sampling induces nontrivial variation in the logit

$$\ell_{kj}(\theta) = a_j(\theta - b_j)$$

The logistic link saturates outside its natural scale and satisfies

$$\sigma'(x) = \sigma(x)(1 - \sigma(x)) \leq \frac{1}{4},$$

so meaningful variation in $\sigma(x)$ requires that posterior samples move $\ell_{kj}(\theta)$ by $\mathcal{O}(1)$ rather than remaining concentrated in a tiny neighborhood. A sufficient, item-scale condition is that for some non-negligible fraction of posterior mass,

$$|\ell_{kj}(\theta) - \ell_{kj}(\mu_k^{(r)})| = a_j|\theta - \mu_k^{(r)}| \gtrsim 1,$$

where $\mu_k^{(r)}$ denotes the posterior center (e.g., the MAP or posterior mean). This highlights the correct scaling: the *natural* ability displacement needed to avoid a near-constant predictive probability is $\approx 1/a_j$. To obtain a single global bound (independent of j) while remaining conservative, we enforce this condition at the typical discrimination scale

$$\bar{a} \stackrel{\text{def}}{=} \text{median}\{a_j : j \in Q_r^*\}$$

In our implementation, item discriminations are calibrated on the standard IRT scale (and optionally normalized), so $\bar{a}$ is close to 1. We then impose a dispersion requirement ensuring that Monte Carlo estimates of BALD do not collapse in practice:

$$\mathbb{P}\left(|\theta - \mu_k^{(r)}| \geq \frac{1}{\bar{a}}\right) \geq \rho,$$

with ρ chosen so that a substantial fraction of posterior samples explore both sides of the ability scale (we take $\rho \approx 0.48$, i.e., roughly half the samples). For a Gaussian posterior $\theta \sim \mathcal{N}(\mu, \sigma^2)$,

$$\mathbb{P}\left(|\theta - \mu| \geq \frac{1}{\bar{a}}\right) = 2\Phi\left(-\frac{1}{\bar{a}\sigma}\right),$$

where Φ is the standard normal CDF. Setting $\rho = 0.48$ yields

$$2\Phi\left(-\frac{1}{\bar{a}\sigma}\right) \geq 0.48 \iff \frac{1}{\bar{a}\sigma} \leq 0.706\dots \iff \sigma \gtrsim \frac{1.416}{\bar{a}} \iff \sigma^2 \gtrsim \frac{2.0}{\bar{a}^2}$$

In our calibrated setting, item discriminations lie on the natural IRT scale, with $\bar{a} \approx 1$. Under this scaling, the smallest variance satisfying the dispersion criterion is $\sigma^2 \approx 2$. We therefore set the lower clipping bound to $\sigma_{\min}^2 = 2$. This choice ensures that posterior sampling regularly explores the non-saturated region of the logistic link at the typical discrimination scale.

Specifically, a non-negligible fraction of posterior mass satisfies $|a_j(\theta - b_j)| \lesssim 1$, where predictive probabilities remain sensitive to changes in ability. As a result, epistemic disagreement is preserved, and BALD does not collapse in practice.

Upper bound: While the lower bound prevents overconfident posterior collapse, an upper bound is required to prevent degeneration driven by excessive posterior dispersion. Our acquisition scores are multiplicative:

$$S_{kj}^{(r)} = \tilde{\pi}_{kj}^{(r)} \cdot \text{BALD}_{kj}^{(r)}$$

Since responses are Bernoulli, $0 \leq H(p) \leq \log 2$, hence $0 \leq \text{BALD}_{kj}^{(r)} \leq \log 2$. Thus, if the expected difficulty weight $\tilde{\pi}_{kj}^{(r)}$ vanishes uniformly across items, all acquisition scores collapse regardless of epistemic disagreement. We use the difficulty-aware weighting

$$\pi_{kj}(\theta) = \exp\left(-\frac{1}{2}(b_j - \theta)^2\right), \quad \tilde{\pi}_{kj}^{(r)} = \mathbb{E}_{\theta \sim p_E^{(r)}}[\pi_{kj}(\theta)]$$

For $\theta \sim \mathcal{N}(\mu, \sigma^2)$, the expectation admits the closed form

$$\mathbb{E}_\theta[\pi_{kj}(\theta)] = \frac{1}{\sqrt{1 + \sigma^2}} \exp\left(-\frac{(b_j - \mu)^2}{2(1 + \sigma^2)}\right)$$

In particular, even under perfect alignment $b_j = \mu$, the best-case expected weight is

$$\max_{b_j} \mathbb{E}_\theta[\pi_{kj}(\theta)] = \frac{1}{\sqrt{1 + \sigma^2}}$$

Therefore, for all items j ,

$$S_{kj}^{(r)} \leq (\log 2) \frac{1}{\sqrt{1 + \sigma^2}},$$

so letting $\sigma^2 \rightarrow \infty$ forces $S_{kj}^{(r)} \rightarrow 0$ uniformly, rendering acquisition unstable. To choose a concrete upper bound, we tie it to the lower bound in a principled way by constraining the maximum attenuation factor of the multiplicative weight over the admissible variance range.

Specifically, we require that the best-case weight does not shrink by more than a factor $\sqrt{1 + \sigma_{\max}^2}$ relative to its maximum value 1 (attained at $\sigma^2 = 0$), and we set the variance range so that the upper bound is a fixed multiple of the lower bound: $\sigma_{\max}^2 = 3\sigma_{\min}^2$.

With $\sigma_{\min}^2 = 2$, this yields

$$\frac{1}{\sqrt{1 + \sigma_{\max}^2}} = \frac{1}{\sqrt{7}}$$

Equivalently, this enforces the explicit attenuation floor

$$\frac{1}{\sqrt{1 + \sigma_{\max}^2}} \geq \frac{1}{\sqrt{7}} \iff \sigma_{\max}^2 \leq 6,$$

ensuring that even in the best-aligned regime, the expected difficulty weight remains at least $1/\sqrt{7}$ and cannot uniformly suppress all acquisition scores. We therefore adopt the upper clipping bound $\sigma_{\max}^2 = 6$.

A.8 Entropy-Adaptive Posterior Sampling

Bayesian acquisition criteria require expectations with respect to the evaluator's posterior belief $p_E^r(\theta)$. In practice, these expectations are approximated via Monte Carlo sampling. While the Laplace approximation provides a tractable local Gaussian representation of the posterior, relying solely on this approximation can underestimate tail uncertainty, particularly when curvature is locally steep or the posterior is mildly misspecified. This can concentrate samples near the mode and miss informative regions, leading to unstable acquisition behavior under limited evaluation budgets. To obtain a more robust sampling distribution, we construct a Gaussian scale mixture centered at the MAP estimate. Let the

differential entropy of the Laplace approximation be $H_{k,r} = \frac{1}{2} \log(2\pi e \sigma_{k,r}^2)$. We map this entropy to a tail multiplier $\kappa_{k,r} = \text{clip}(1 + H_{k,r}, \kappa_{\min}, \kappa_{\max})$, and define two sampling scales $s_{1,k,r} = \sigma_{k,r}$, $s_{2,k,r} = \kappa_{k,r} \sigma_{k,r}$. Samples of θ_k are then drawn from the mixture

$$\theta_k^r \sim \frac{1}{2} \mathcal{N}(\hat{\theta}_k^r, s_{1,k,r}^2) + \frac{1}{2} \mathcal{N}(\hat{\theta}_k^r, s_{2,k,r}^2)$$

Gaussian scale mixtures are commonly used in Bayesian inference to encourage exploration beyond local curvature approximations (Neal, 1996). The resulting sampling distribution preserves local posterior structure while providing sufficient dispersion for stable acquisition under posterior uncertainty.