ACTIVE EVALUATION FRAMEWORK FOR RELIABLE MODEL COMPARISON

RankBALD: Ranking-Aligned Active Evaluation for Language Models

Bonaventure F. P. Dossou

PhD Candidate · McGill University & Mila Québec AI Institute

Accepted at the **Conference on Language Modeling (COLM) 2026 main conference**.

CORE QUESTION

Which model is better?

And how can we answer that reliably without evaluating every model on every benchmark item?

$$Z^r_{s,t} = \mathbb{1}[\theta^r_s > \theta^r_t], \qquad Y^r_{s,t,j} = (U^r_{sj}, U^r_{tj}) \in \{0,1\}^2$$

INTUITION FIRST

Which model should we use, and which question should we ask next?

In practice, evaluation is a model-selection problem: among many LLMs, which one is best, which one should we use, and how confident are we in that ranking? Under a limited budget, the next benchmark item should help recover the ordering, not merely reduce uncertainty about an isolated score.

Current belief about capability

The two leading models are close. The real risk is not estimating every score perfectly; it is choosing the wrong winner.



What we care about: which model is best, and the final ordering A vs B vs C vs D.

CANDIDATE Q_1

Easy recall



Almost everyone answers correctly: little help for choosing the winner.

CANDIDATE Q_2

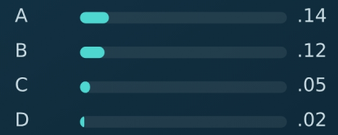
Boundary item



Rank-defining: plausible outcomes can separate A from B.

CANDIDATE Q_3

Very hard reasoning



Almost everyone fails: again, little help for the ranking.

RankBALD starts from this intuition: evaluate to recover the model ranking, not merely to reduce uncertainty about isolated scores.

FROM INTUITION TO A MEASUREMENT PROBLEM

The setup: many models, many questions, one limited evaluation budget

Before choosing an acquisition rule, we need to formalize what is being measured. We have a fixed pool of candidate LLMs, a fixed bank of benchmark items, and a limited number of responses we can afford to observe.

Candidate LLMs

The object of interest is not one isolated score. It is the ordering of models under the capability we care about.

- **A** — **Model A** — strong but uncertain — ? rank
- **B** — **Model B** — close competitor — ? rank
- **C** — **Model C** — different strengths — ? rank
- **D** — **Model D** — lower prior estimate — ? rank

→

Question / item bank

Items probe different abilities. Some are too easy, some too hard, and some are maximally informative for distinguishing close models.

- **Q1 · commonsense** — Most models likely answer correctly.
- **Q2 · math reasoning** — Separates the top two models.
- **Q3 · factuality** — Good for absolute score, not ranking.
- **Q4 · multilingual QA** — Tests generalization beyond English-centric settings.
- **Q5 · hard reasoning** — High value if the leading models disagree.
- **Q6 · safety / truthfulness** — Important capability dimension.

→

Active evaluation decision

At each step, choose the next item whose observed responses would most improve the ranking decision.

WHAT DO WE CARE ABOUT?
Which model is best for the target capability, and how stable is the full ranking?

WHAT CAN WE QUERY?
A small subset of item responses under a fixed budget.

WHAT MUST BE INFERRED?
Latent model capability from sparse correct / incorrect responses.

Core active-evaluation question: Which item should we ask next to verify the ranking efficiently?

This motivates IRT

IRT gives the measurement layer: it links item difficulty and model capability to response probability.

THE MEASUREMENT FRAMEWORK

What is Item Response Theory?

IRT is a probabilistic framework for inferring an unobserved capability from a pattern of correct and incorrect responses. It was developed for educational testing; here, students become language models and exam questions become benchmark items.

From observed responses to hidden capability

We do not observe a model's “true” capability directly. We observe only how it responds to a subset of benchmark items.

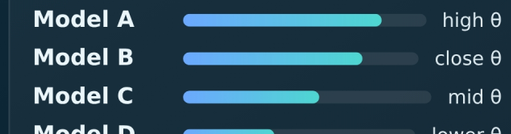
	Arithmetic	Logic	Coding	Math	Multilingual
Model A	✓	✓	✓	?	✓
Model B	✓	✓	✗	✓	?
Model C	✗	✓	?	✗	✓
Model D	?	✗	✓	✗	✗

Observed

Sparse correct / incorrect responses across heterogeneous benchmark items.

→

Inferred by IRT



Item Response Theory (IRT) explains observed responses through latent model capability and item properties, instead of treating every question as equally informative and relying only on raw accuracy.

Educational testing

Student · exam item · latent student ability

LLM evaluation

Language model · benchmark item · latent model capability

1PL / Rasch model

- Models item difficulty
- Assumes shared discrimination
- Simple and highly interpretable
- Less expressive for heterogeneous benchmarks

2PL IRT

- Models difficulty and discrimination
- Allows items to differ in informativeness
- Better captures benchmark heterogeneity
- Suitable for active item selection

Why 2PL here? Modern benchmark items are not equally informative. Some sharply separate nearby models, while others add little evidence about their relative capability. The next slide formalizes this with item difficulty and discrimination.

MEASUREMENT MODEL ONLY

2PL IRT models whether a model answers an item correctly

Before we can ask which benchmark item to evaluate next, we need a probabilistic link between **models**, **items**, and observed correctness. 2PL IRT provides that measurement layer; the active evaluator is introduced later as our contribution.

2PL IRT in one sentence

Each LLM has an unobserved ability, each benchmark item has a difficulty and a discrimination, and the model is more likely to be correct when ability exceeds difficulty.

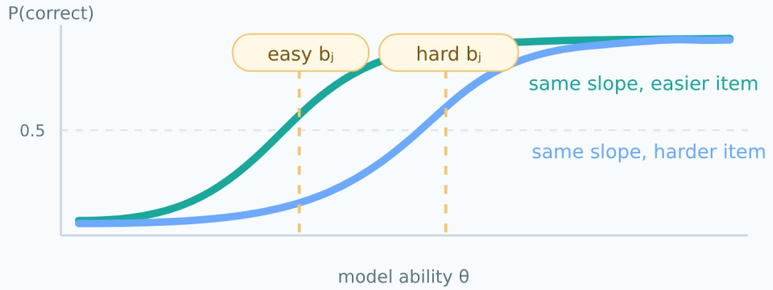
$$P(U_{kj}=1 \mid \theta_k, a_j, b_j) = \sigma(a_j(\theta_k - b_j))$$

θ_k	latent capability of model **k**; higher means more likely to answer difficult items correctly.
b_j	difficulty/location of item **j**; at $\theta_k=b_j$, the item is around the decision boundary.
a_j	discrimination/slope of item **j**; higher means the item sharply separates nearby capabilities.

Separation: IRT estimates response probabilities. It does not decide which item to ask next. Our active evaluation framework adds that decision rule.

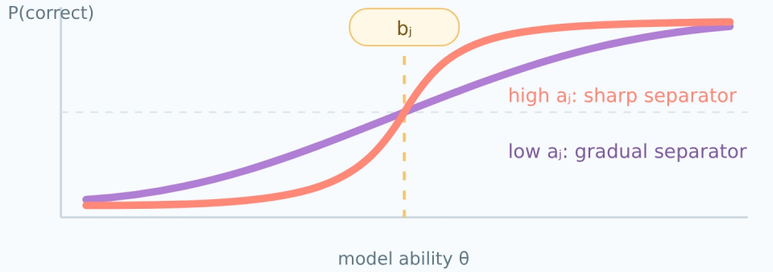
Difficulty b_j shifts the curve

Harder items require higher ability to reach the same probability of correctness.



Discrimination a_j changes the slope

Highly discriminative items are most informative near their difficulty boundary.



WHAT OTHERS HAVE DONE

Efficient evaluation has mostly been framed as measuring ability with fewer items

STATIC EVALUATION

Use the full benchmark

Every model answers every item; accuracy is aggregated uniformly.

Strength: simple and exhaustive.
Cost: spends equally on obvious and discriminative questions.

OFFLINE SUBSETS

TinyBenchmarks & MetaBench

Select a compact fixed subset once, before observing a new evaluation run.

Polo et al. (2024); Kipnis et al. (2025). Efficient, but cannot adapt after seeing responses.

ADAPTIVE IRT

Sequential ability estimation

Use previous answers to choose questions informative about latent capability.

Rodriguez et al.; Lalor et al.; Song et al. The acquisition target remains ability precision.

CLOSEST BASELINE

Fluid Benchmarking

Select items using Fisher information under a calibrated IRT response model.

Hofmann et al. (2025). Strong adaptive baseline, and the direct motivation for our question.

Shift in objective: instead of asking only “how accurately can we estimate each θ?”, we ask “how reliably can we recover their ordering?”

WHY THE EXISTING OBJECTIVE IS NOT ENOUGH

Fisher information is well matched to ability precision, not necessarily to ranking

$$\mathcal{I}_j(\theta_k) = a_j^2\, p_{kj}(\theta_k)\big(1 - p_{kj}(\theta_k)\big)$$

$$\mathcal{A}^r_{\text{Fisher}}(q_j) \propto \sum_{k=1}^{K} \mathbb{E}_{\theta_k \sim p^r_E}[\mathcal{I}_j(\theta_k)]$$

$$\text{Var}(\hat{\theta}_k) \gtrsim \left(\lambda + \sum_{j \in Q^*_r} \mathcal{I}_j(\theta_k)\right)^{-1}$$

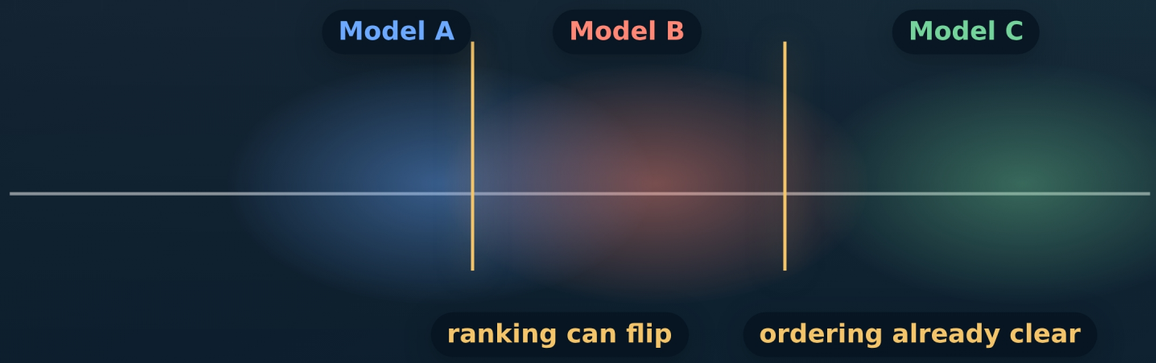
Fisher favors items with high local likelihood curvature, reducing the marginal uncertainty of each capability estimate.

Ability-estimation target

Make each posterior $p^r_E(\theta_k)$ narrower. An item is useful when it improves local precision, even if the model is already safely above or below every competitor.

Ranking target

Resolve uncertain comparisons $\theta_s > \theta_t$. An item matters when its possible outcomes can change the ordering of nearby competitors.



Important: Fisher is not incorrect; it is optimized for a different downstream objective (which is to estimate each model's latent capability as accurately as possible).

OUR WORK: PROBLEM FORMULATION

We introduce a novel active evaluation framework whose target is the ranking, not only θ precision

The active evaluator is the sequential decision component inside the framework, not a part of IRT. It sees a partial response matrix, updates beliefs using the same 2PL model, then chooses the next benchmark item to ask every fixed model.

Partial answers
Only selected columns of U are observed

→

IRT posterior
Belief over each model's θ

→

Ranking belief
Uncertainty over who is above whom

→

Next item
Choose the item with highest acquisition value

$$p_E^r(\theta_k) \propto p(U_{k,Q_r^*} \mid \theta_k)\, p(\theta_k) \;\approx\; \mathcal{N}(\hat{\theta}_k^r, \sigma_{k,r}^2)$$

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

Design choice: if the downstream decision is "which model is best?", the acquisition function should reduce uncertainty about the ordering.

What is held fixed across methods?

MODELS
Population, frozen weights, and observed responses

ITEM BANK
Questions, labels, and candidate pool

IRT CALIBRATION
Same item difficulties b and discriminations a

LIKELIHOOD
Same 2PL correctness model

INFERENCE
Same prior, MAP estimate, and Laplace posterior

PROTOCOL
Same seed items, budget, and stopping rule

Only one component changes: how the evaluator scores each candidate item.

THE RIGHT UNCERTAINTY OBJECT

Ranking is determined by pairwise variables, not by isolated ability estimates

For every model pair, define the binary ordering variable $Z_{a,b} = 1[\theta_a > \theta_b]$. We summarize uncertainty in that binary decision with entropy $H(\cdot)$: high when either model could win, low when the order is nearly settled.

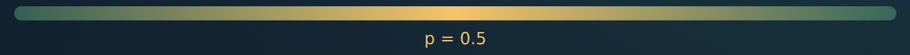
WHAT DOES H MEAN?

Let $p_{a,b} = P(\theta_a > \theta_b)$. This is the posterior probability that model A's latent ability is higher than model B's, computed from posterior samples after the observed item responses. Binary entropy converts that belief into one uncertainty score:

$$H(Z_{a,b}) = -p_{a,b} \log p_{a,b} - (1-p_{a,b}) \log(1-p_{a,b})$$

HOW TO READ IT

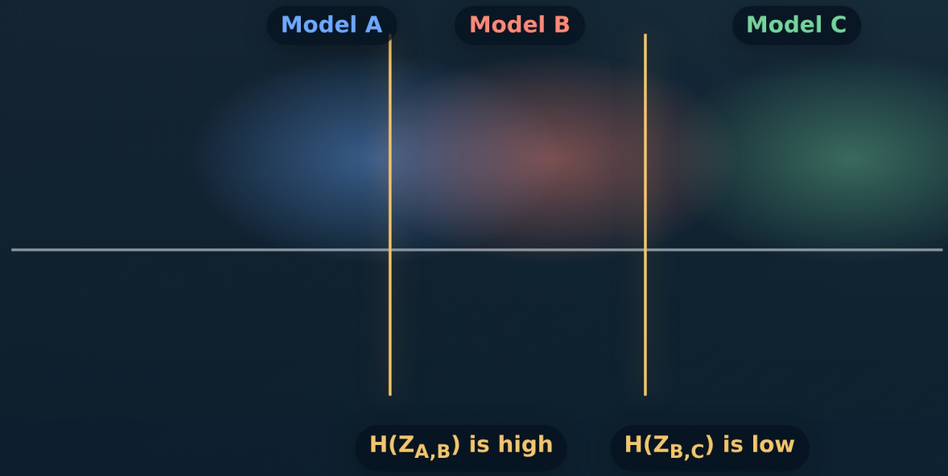
Entropy is largest when $p_{a,b} = 0.5$: the winner is genuinely unresolved. It approaches zero as $p_{a,b}$ approaches 0 or 1.



$$\pi^* = \operatorname{argsort}_k(-\theta_k), \qquad \theta_{\pi(1)} \geq \cdots \geq \theta_{\pi(K)}$$

$$Z^r_{s,t} = \mathbb{1}[\theta^r_s > \theta^r_t], \qquad Y^r_{s,t,j} = (U^r_{sj}, U^r_{tj}) \in \{0,1\}^2$$

Close competitors matter most: their ordering has high entropy and can be changed by plausible item outcomes.



Objective alignment: select the item that is expected to tell us the most about these ordering variables.

DESIGN PROGRESSION · 1 OF 3

BALD asks which response is informative; DiffBALD asks whether the item is informative where ability is uncertain

The first step adapts generic Bayesian disagreement to active evaluation. The next step recognizes a 2PL-specific failure mode: a response can look uncertain yet come from an item whose difficulty is poorly aligned with the current ability posterior.

FOUNDATION

BALD

Select an item whose observed correctness is expected to reduce epistemic uncertainty about a model's latent ability.

$$\mathrm{BALD}_{k,j} = I(U_{k,j};\, \theta_k)$$

What remains missing? BALD measures disagreement, but does not explicitly favor items whose difficulty lies in the posterior region where the 2PL response curve is most sensitive.

→

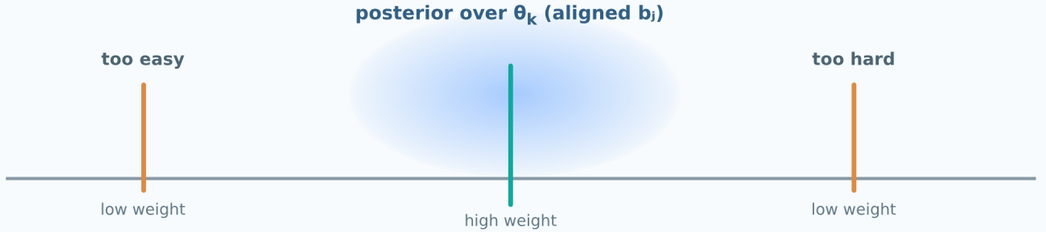
ADDS DIFFICULTY ALIGNMENT

DiffBALD

Weight BALD by how closely item difficulty b_j overlaps the posterior region of θ_k.

Item relevance on the 2PL ability scale

The aligned item probes the region in which the model's ability is still plausible.



$$\mathrm{DiffBALD}_{k,j} = \bar{\pi}_{k,j} \cdot \mathrm{BALD}_{k,j}$$

WHAT WAS FIXED? DiffBALD keeps BALD's information-gain principle, but discourages queries that are informative only in the abstract and poorly located on the IRT ability scale.

DESIGN PROGRESSION · 2 OF 3

DiffBALD improves where we query; VarBALD adds an explicit preference for posterior contraction

Difficulty alignment does not directly say how much the remaining uncertainty in θ will shrink. VarBALD adds a one-step look-ahead using posterior-averaged Fisher information.

DIFFICULTY-AWARE

DiffBALD

Prioritizes items near the model's plausible ability region, while retaining BALD's response-level information objective.

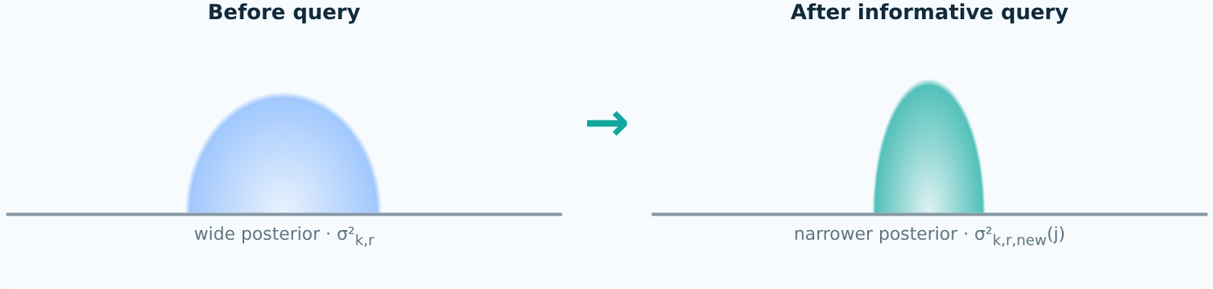
$$\bar{\pi}_{k,j} \cdot I(U_{k,j}; \theta_k)$$

What remains missing? Two aligned items can have similar information scores while implying different expected reductions in the width of the ability posterior.

ADDS VARIANCE REDUCTION

VarBALD

Prefer the candidate expected to produce the larger one-step reduction $\Delta\sigma^2$ in posterior ability variance.



$$\text{VarBALD} = \text{DiffBALD} \cdot [1 + \alpha\, \Delta\sigma^2/(\sigma^2 + \varepsilon)]$$

WHAT WAS FIXED? VarBALD makes stability explicit: among difficulty-aligned items, it rewards those expected to contract the scalar ability posterior most.

DESIGN PROGRESSION · 3 OF 3

VarBALD estimates each θ more precisely; RankBALD changes the target to the decision we actually care about: who ranks above whom

Smaller marginal variance is useful, but ranking error occurs at pairwise boundaries. The largest conceptual step is therefore not another scalar-precision proxy—it is information gain about the ordering variable itself.

SCALAR PRECISION

VarBALD

Optimizes posterior contraction for each model ability θ_k, one model at a time.

target: reduce $\mathrm{Var}(\theta_k)$

What remains missing? A model can be estimated more precisely without resolving whether it is above or below a close competitor. Ranking is a joint, pairwise decision.

→

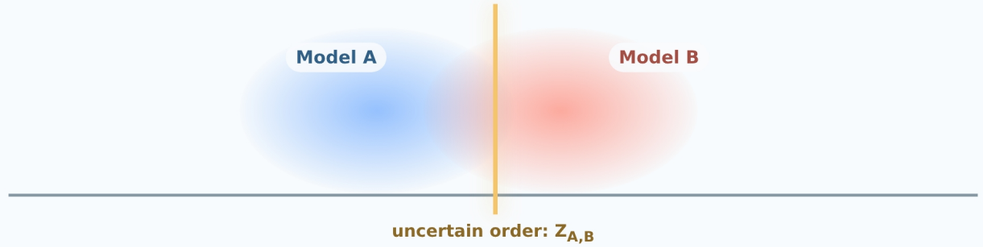
CHANGES THE INFERENCE TARGET

RankBALD

Score each candidate by the expected information its joint outcomes provide about uncertain pairwise orderings $Z_{s,t}$.

The decision boundary that matters

When two posteriors overlap, the next item should resolve the winner—not merely narrow both curves.



target: $I(Z_{s,t}; Y_{s,t,j})$

WHAT WAS FIXED? RankBALD aligns acquisition with the downstream evaluation decision. It inherits difficulty alignment and posterior calibration, but asks which item best resolves uncertain rank boundaries.

DESIGN INTUITION · WHY RANKING-AWARE ACQUISITION MATTERS

RankBALD chooses the question whose outcomes are expected to clarify the final ordering

At this stage, the evaluator has many possible benchmark items left. The key question is not only whether an item is uncertain, but whether its possible outcomes would change what we believe about the ranking.

Current state

The evaluator is confident about the bottom model, but uncertain about the top two.

?	**Model A** could be rank 1 or 2	close
?	**Model B** could be rank 1 or 2	close
3	**Model C** clearly below A/B	stable

Goal: spend the next query on the item that most reduces uncertainty about who should be ranked above whom.

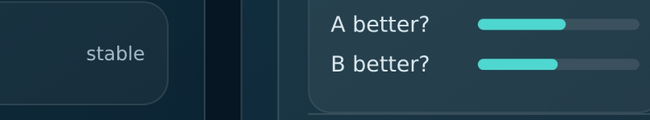
Two candidate questions

CANDIDATE ITEM 1

General confidence check

Changes ability estimates slightly, but leaves the top-two ordering nearly unchanged.

POSSIBLE OUTCOMES



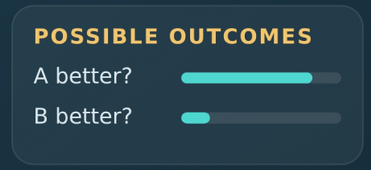
× Ordering remains ambiguous; useful for measurement, weak for model selection.

CANDIDATE ITEM 2

Boundary item

Possible responses separate the two models whose ordering is uncertain.

POSSIBLE OUTCOMES



✓ Ordering uncertainty is reduced; the query improves the final ranking.

RankBALD's choice

RankBALD evaluates candidate items by the expected information they provide about pairwise ordering variables.

↓

1	**Model A** ordering uncertainty reduced	clearer
2	**Model B** ordering uncertainty reduced	clearer
3	**Model C** unchanged but already stable	stable

Key implication: the acquisition function is aligned with the actual decision: reliable model ranking.

MAIN CONTRIBUTION

RankBALD scores an item by the ordering uncertainty its outcomes can resolve

$$Z^r_{s,t} = \mathbb{1}[\theta^r_s > \theta^r_t], \qquad Y^r_{s,t,j} = (U^r_{sj}, U^r_{tj}) \in \{0,1\}^2$$

$$I(Z^r_{s,t}; Y^r_{s,t,j} \mid p^r_E) = H(Z^r_{s,t}) - \sum_{y \in \{0,1\}^2} p(y)\, H(Z^r_{s,t} \mid Y = y)$$

$$\omega(d) = 1 - \frac{d-1}{K-2}, \quad d = |\rho_r(s) - \rho_r(t)|; \qquad \sigma_{\pi,r} = \mathrm{median}_k \,\mathrm{Std}[\theta_k]$$

$$\mathrm{RankBALD}^r_j = \frac{\sum_{(s,t)\in\mathcal{P}} \omega(|s-t|)\, \bar{\pi}^r_{js} \bar{\pi}^r_{jt}\, I\left(Z^r_{s,t}; Y^r_{s,t,j} \mid p^r_E\right)}{\sum_{(s,t)\in\mathcal{P}} \omega(|s-t|)\, \bar{\pi}^r_{js} \bar{\pi}^r_{jt} + \varepsilon}$$

Pairwise MI

Measures how much joint responses reveal about who outranks whom.

Difficulty alignment

Downweights questions that are implausibly easy or hard for both models.

Rank proximity

Prioritizes neighboring competitors where a ranking error is consequential.

In other words, RankBALD actively selects the benchmark questions that best distinguish competing language models, leading to a more accurate and reliable final ranking.

CONTROLLED COMPARISON

We compare acquisition objectives under the same data, inference pipeline, and evaluation protocol

102
models used to calibrate separate 2PL item banks

6
evaluated model families with 61–94 checkpoints each

6+5
Open LLM Leaderboard benchmarks plus AfroBench generalization

4
budgets: 10, 50, 100, and 500 items

Main benchmark suite

ARC Challenge · GSM8K · HellaSwag · MMLU · TruthfulQA · WinoGrande

Evaluated checkpoints and post-trained variants are excluded from calibration to prevent leakage.

Beyond English-centric evaluation: AfroBench is included as an additional generalization suite to test whether the same evaluation logic carries to African-language benchmarks.

Methods

Random-IRT · Fluid · BALD · DiffBALD · VarBALD · RankBALD

The item bank, calibrated parameters, model responses, prior, MAP/Laplace inference, initialization, and evaluation budget are identical across all methods, including baselines. Only the acquisition score differs.

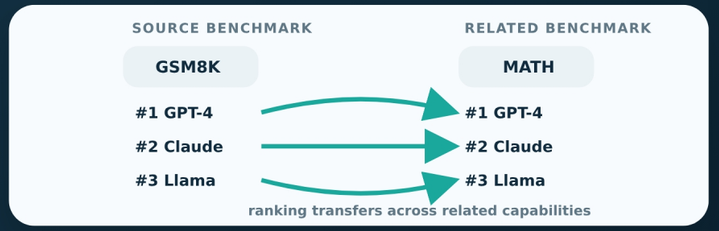
WHAT COUNTS AS A GOOD ADAPTIVE BENCHMARK?

Three complementary views of evaluation reliability

CROSS-BENCHMARK CONSISTENCY

Validity

Rank distance ↓



$$\text{Validity}(B_1, B_2) = \frac{1}{K} \sum_{k=1}^{K} |\text{rank}_{B_1}(\mathcal{M}_k) - \text{rank}_{B_2}(\mathcal{M}_k)|$$

Does the selected subset preserve relative model order across related capabilities?

CHECKPOINT STABILITY

Variance

Total variation ↓



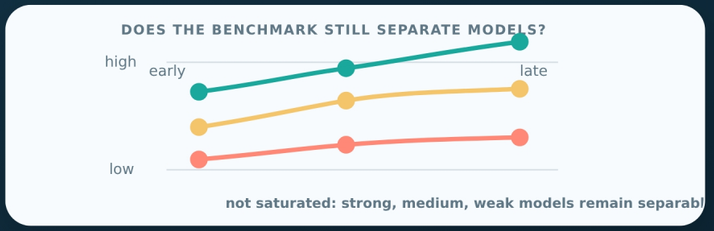
$$\text{TV}(\mathcal{M}_k, Q) = \frac{T}{T-1} \cdot \frac{\sum_{t=1}^{T-1} |x_k^{t+1} - x_k^t|}{|x_k^T - x_k^1|}$$

Do model trajectories remain smooth rather than oscillating as checkpoints improve?

DISCRIMINATION OVER TRAINING

Saturation

Rank correlation ↑



$$\text{Saturation} = \left|\rho_{\text{Spearman}}\left(\{1, \ldots, T\}, \{x_k^1(Q), \ldots, x_k^T(Q)\}\right)\right|$$

Does the benchmark remain monotonic and discriminative across the capability range?

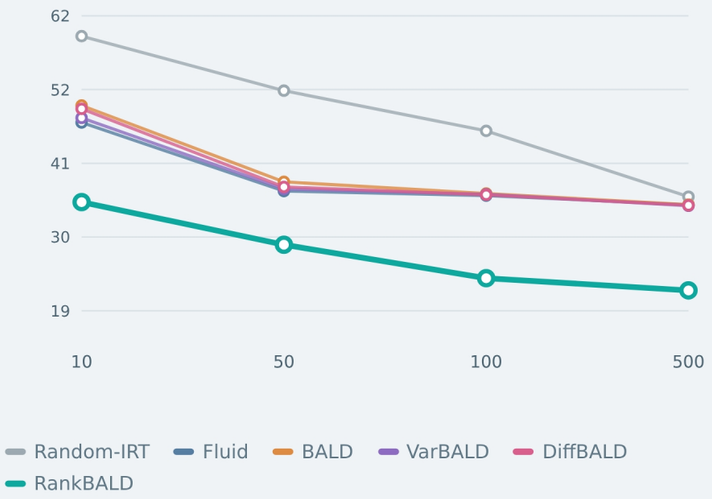
MAIN RESULT: ACROSS ITEM BUDGETS

Does RankBALD recover reliable rankings with fewer benchmark items?

QUESTION **How quickly does each acquisition rule reach stable, valid rankings as the item budget increases?**

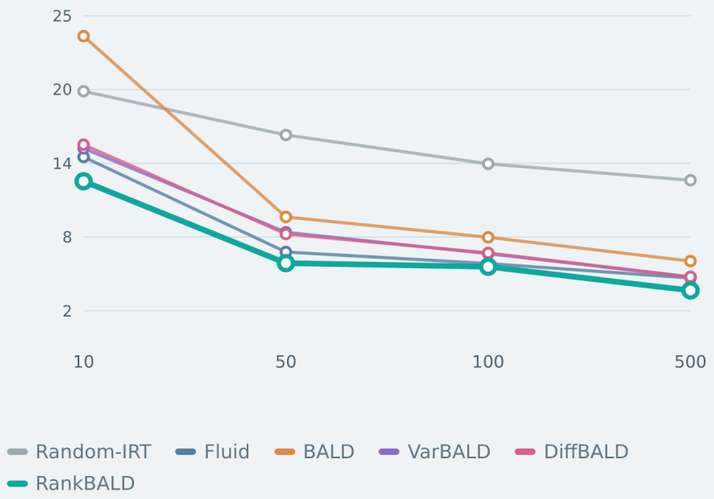
Validity

Rank distance · lower is better



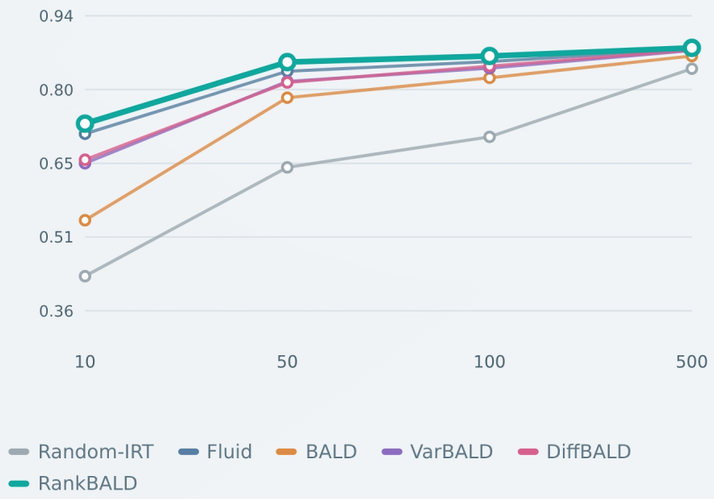
Variance

Total variation · lower is better



Saturation

Rank correlation · higher is better



33.2%
lower validity distance than Fluid at 100 items

4.1%
lower total variation than Fluid at 100 items

+0.011
higher saturation than Fluid at 100 items

Answer: RankBALD reaches the desired ordering signal earlier: the ranking-aware objective improves validity and stability before the full benchmark is observed.

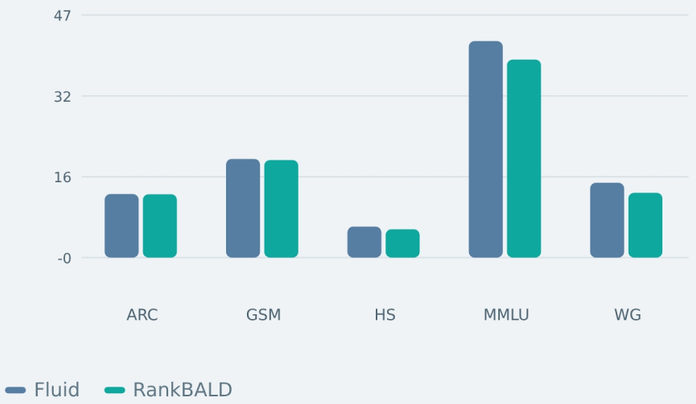
GRANULARITY 1: BENCHMARK

Are the gains broad across tasks, or driven by one benchmark?

QUESTION **Does the advantage persist across reasoning, knowledge, commonsense, and inference benchmarks?**

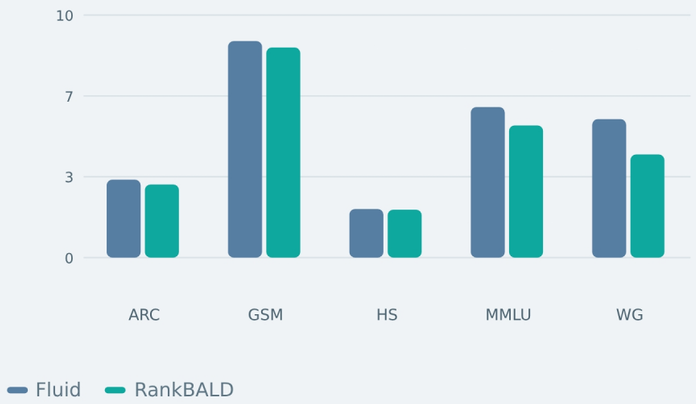
Validity at 100 items

RankBALD vs Fluid · lower is better



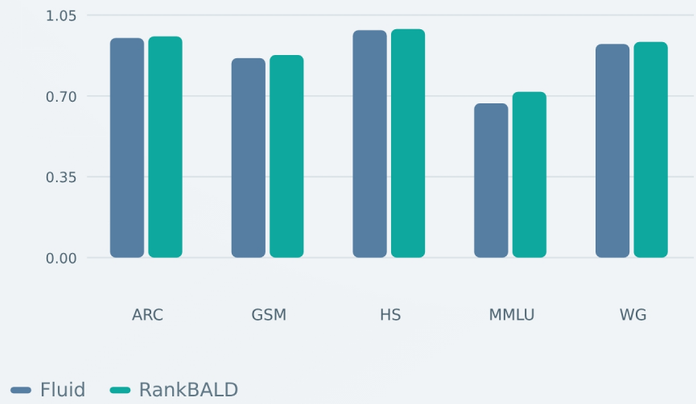
Variance at 100 items

RankBALD vs Fluid · lower is better



Saturation at 100 items

RankBALD vs Fluid · higher is better



Largest practical gains: WinoGrande variance drops 25.5%; MMLU validity improves 8.5% and saturation rises by 0.050.

Answer: The improvements are not a single-benchmark artifact; the largest gains appear on different metrics across different tasks, which supports robustness rather than overfitting to one benchmark.

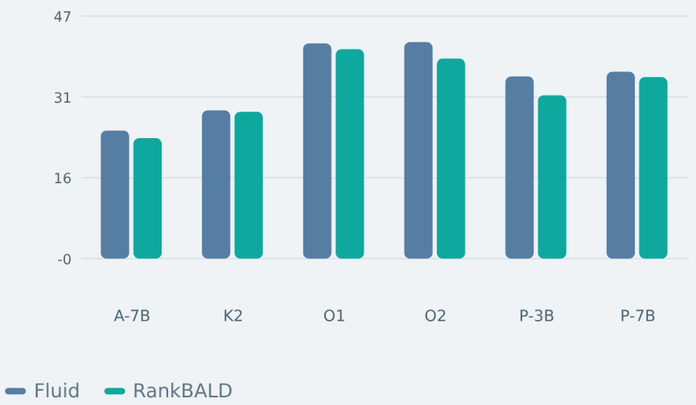
GRANULARITY 2: MODEL FAMILY

Does RankBALD work consistently across different LLM families?

QUESTION **Is the ranking-aware objective robust across model families, rather than tuned to one group of checkpoints?**

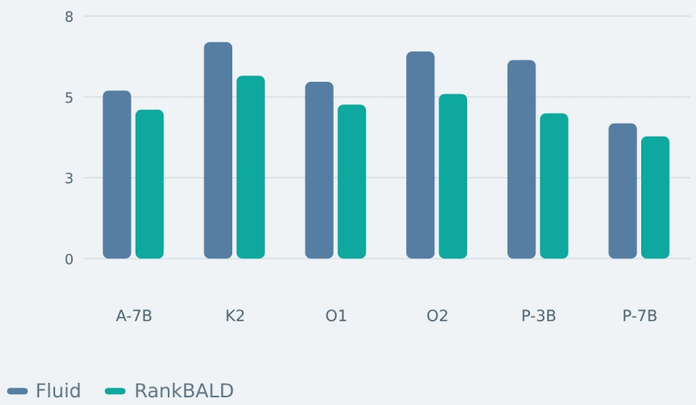
Validity at 100 items

RankBALD vs Fluid · lower is better



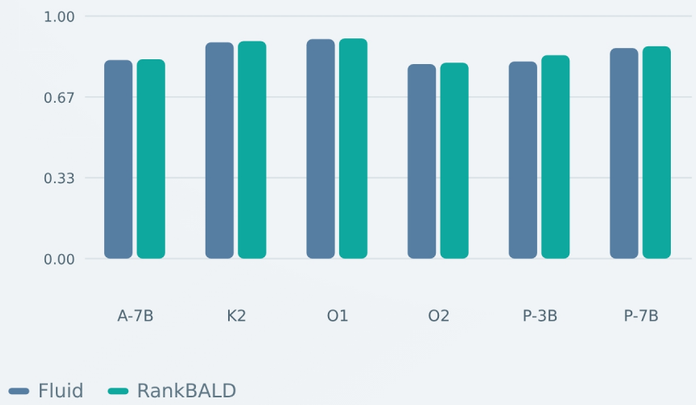
Variance at 100 items

RankBALD vs Fluid · lower is better



Saturation at 100 items

RankBALD vs Fluid · higher is better



5 / 6
families best on validity

6 / 6
families best on variance

6 / 6
families best on saturation

Answer: RankBALD is best or tied-best for almost every family-level comparison: 5/6 families on validity and 6/6 families on both variance and saturation.

GRANULARITY 3: UNCERTAINTY AND SIGNIFICANCE

Are the improvements statistically credible, or just noise?

QUESTION **Do paired comparisons against Fluid support the claim that RankBALD improves ranking reliability?**

Paired comparisons at 100 items

The comparison is made over six model-family observations, using the same item bank, responses, inference pipeline, and budget as every other method.

Validity: 24.23 vs 36.25

95% CI reported over six model-family observations; paired comparison with Fluid: $p = 0.021$.

Variance: 5.05 vs 6.06

RankBALD produces significantly smoother evaluation trajectories than Fluid: $p = 0.005$.

Saturation: .861 vs .850

The gap with Fluid is smaller versus Fluid ($p = 0.066$), but RankBALD is significant versus every BALD variant ($p \leq 0.006$).

Interpretation

VALIDITY

The largest and cleanest effect: RankBALD substantially reduces cross-benchmark rank distance.

VARIANCE

The strongest significance signal: ranking trajectories become smoother across checkpoints.

SATURATION

A smaller gap against Fluid, but still consistently better than the BALD-family alternatives.

Answer: Yes for the primary reliability metrics: the paired evidence is strongest for validity and variance, while saturation is directionally consistent and clearly separates RankBALD from the BALD variants.

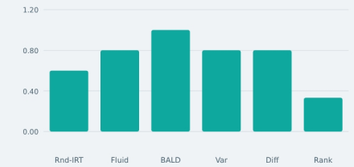
BEYOND ENGLISH-CENTRIC EVALUATION

Does ranking-aware active evaluation transfer beyond English-centric benchmarks?

QUESTION **When the benchmark setting is multilingual, sparse, and cost-sensitive, does the same objective alignment still help?**

Validity

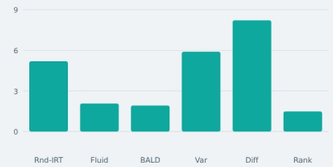
100 items · lower is better



Validity

Variance

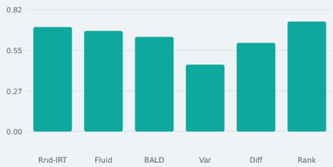
100 items · lower is better



Variance

Saturation

100 items · higher is better



Saturation

Per-benchmark RankBALD at 100 items

	AfriGSM	AfriMMLU	AfriXNLI	Belebele	SIB-200
Validity ↓	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**
Variance ↓	**1.500**	**1.500**	**1.500**	**1.500**	**1.500**
Saturation ↑	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**

RankBALD achieves the lowest cross-budget rank distance (0.333), lowest total variation (1.455), and highest rank correlation (0.740).

Generalization: objective alignment remains useful in sparse, multilingual, and cost-sensitive evaluation settings.

Answer: Yes. The AfroBench results suggest that ranking-aware acquisition remains useful when evaluation data are scarcer and language coverage is broader.

WHAT THIS CHANGES

Benchmark construction is an inference-and-decision problem

Practical implications

- Spend evaluation budget on items that distinguish close competitors.
- Improve model-selection reliability without changing the underlying response model.
- Support cheaper recurring evaluations, model audits, and benchmark refreshes.
- Define acquisition objectives around the decision practitioners need to make.

Current limitations

- Requires pre-fitted item parameters and a fixed item bank.
- Uses a Laplace-style posterior approximation with calibrated variance clipping.
- Pairwise scoring scales quadratically with the number of models.
- AfroBench contains few models and only one multi-checkpoint family.

Next step: scalable near-neighbor pair selection and online item calibration for evolving benchmarks.

TAKEAWAYS

Optimize evaluation for the decision you actually need to make

The central message is objective alignment: if the downstream task is model ranking, the acquisition function should directly reduce uncertainty about the ranking.

1 · Measurement is not selection

Fisher-style objectives estimate latent abilities well, but ranking models correctly is a different downstream goal.

2 · Ranking uncertainty is pairwise

Rank errors happen at boundaries between competing models, so the acquisition target should reflect pairwise orderings.

3 · RankBALD aligns acquisition with the goal

It prioritizes benchmark items that best distinguish competing language models and improve the final ranking.

4 · The pattern is robust

We observe gains across budgets, benchmarks, model families, significance tests, and AfroBench generalization.

RankBALD: Ranking-Aligned Active Evaluation for Language Models
Bonaventure F. P. Dossou and Jackie C. K. Cheung · Accepted at COLM 2026 main conference.

Next step

Extend active evaluation beyond static leaderboards: multilingual item banks, capability-specific ranking, and deployment-aware evaluation budgets.

Thanks for your attention. All questions are welcome.

COMPUTATIONAL DETAIL

Estimating pairwise mutual information by posterior sampling

BACKUP

$$\widehat{p}(z,y) = \frac{1}{S}\sum_{m=1}^{S} \mathbb{1}[Z^{(m)} = z]\, p\Big(y \mid \theta_s^{(m)}, \theta_t^{(m)}\Big)$$

$$I(Z_{s,t}^r; Y_{s,t,j}^r \mid p_E^r) = H(Z_{s,t}^r) - \sum_{y\in\{0,1\}^2} p(y)\, H(Z_{s,t}^r \mid Y = y)$$

For each posterior draw, the ordering Z is deterministic while the four joint response outcomes remain probabilistic under the two Bernoulli response models.

Per candidate item

- Draw posterior samples for every model ability.
- Form the pairwise order variable for each sample.
- Compute probabilities of (0,0), (0,1), (1,0), and (1,1).
- Accumulate the joint distribution p(Z,Y).
- Compute MI and aggregate across weighted model pairs.

The current implementation uses 64 Monte Carlo samples.

POSTERIOR MECHANICS

Difficulty alignment and one-step variance reduction

BACKUP

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

Under a Gaussian posterior, the expected difficulty kernel has a closed form: uncertainty broadens the region of useful item difficulty while attenuating the peak.

$$\sigma^2_{k,r,\text{new}}(j) = \left(\sigma^{-2}_{k,r} + \bar{I}^r_{kj}\right)^{-1}, \qquad \Delta\sigma^2_{kjr} = \sigma^2_{k,r} - \sigma^2_{k,r,\text{new}}(j)$$

Local precision is additive: current posterior precision plus expected item Fisher information gives the one-step lookahead variance.

Empirical lesson: difficulty and variance awareness help, but the decisive gain comes from changing the target from θ uncertainty to ordering uncertainty.

MAIN NUMERIC SUMMARY

Budget = 100: all methods across all three metrics

BACKUP

Method	Validity ↓	Variance ↓	Saturation ↑	$p_{validity}$	$p_{variance}$	$p_{saturation}$
Random-IRT	45.67 ± 3.64	13.77 ± 1.49	0.702 ± .024	.005	.002	.0004
Fluid	36.25 ± 2.81	6.06 ± .40	0.850 ± .017	.021	.005	.066
BALD	36.55 ± 4.03	8.09 ± 1.13	0.818 ± .021	.017	.022	.006
VarBALD	36.43 ± 2.85	6.82 ± .70	0.837 ± .019	.015	.013	.001
DiffBALD	36.37 ± 2.76	6.88 ± .68	0.841 ± .019	.013	.010	.002
RankBALD	**24.23 ± 2.65**	**5.05 ± .27**	**0.861 ± .017**	**—**	**—**	**—**

The LM-level significance aggregation differs from the overall budget table aggregation; therefore the RankBALD variance mean here (5.05) is not the same statistic as the 5.81 budget-table value.