WHEN DOES A RANDOM SUBSET PRESERVE THE FULL-BENCHMARK MODEL RANKING?

Anonymous authors
Paper under double-blind review

ABSTRACT

Evaluating large language models (LLMs) is expensive, so a natural strategy is to evaluate each model on only a random subset of benchmark questions. But how many questions are enough? We show that subset size alone is a poor guide: at the same evaluation budget, the agreement between the subset ranking and the full-benchmark ranking can vary substantially across benchmarks. The missing factor is how informative benchmark questions are for distinguishing the particular models being compared. We characterize this using population-relative item discrimination and combine it with classical measurement theory to predict how closely a random subset recovers the full-benchmark model ranking. Across 114 benchmark instances from 11 suites and five evaluation budgets spanning 1%–20%, predicted and observed ranking agreement correlate at $r = 0.95$, with correlations ranging from $r = 0.91$ to $r = 0.96$ when evaluated separately at each budget. The predictor also estimates the subset size required for a target level of ranking agreement and the potential for further improvements from evaluating more questions. Finally, we show that achievable ranking agreement depends on the model population itself: when evaluated models become more similar, benchmarks can become substantially less informative, and discrimination estimates from a mismatched source population may not transfer. Overall, our results provide a simple way to assess when random sampling can recover full-benchmark rankings, how many questions are needed, and when benchmark properties limit how well a subset can recover the full ranking.

1 INTRODUCTION

Evaluating large language models is increasingly expensive. Modern benchmarks contain hundreds to thousands of questions, and evaluating many models across such benchmarks requires substantial inference computation. This has motivated work on efficient evaluation, which aims to estimate model performance using only a fraction of a benchmark (Maia Polo et al., 2024; Vivek et al., 2024; Kipnis et al., 2025; Hofmann et al., 2025; Bowyer et al., 2026; Rubinstein et al., 2026). The basic promise is simple: if a small subset of questions preserves the model ranking, much of the evaluation cost can be avoided.

But how large should this subset be?

Consider evaluating the same fraction of each benchmark. Across a collection of 114 benchmark instances, evaluating a random 10% of the questions produces full-benchmark ranking agreement ranging from 0.31 to 1.00 (Figure 1). Thus, the same evaluation budget can preserve a ranking well on one benchmark but poorly on another. The key question is: *what determines how much ranking information a benchmark provides about the population of models being evaluated?*

We show that a standard measurement property, *item discrimination* D (Lin et al., 2024), provides a useful explanation. We define D as the population-relative association between individual question correctness and overall model performance. Higher D means that questions more consistently distinguish stronger from weaker models, while lower D means that more questions are needed to recover their ordering. Building on classical measurement theory and the Spearman-Brown relationship (Spearman, 1910; Brown, 1910; Urbano et al., 2013), we obtain a predictor relating D and evaluation size k to full benchmark ranking agreement under random subsampling. Across 114

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

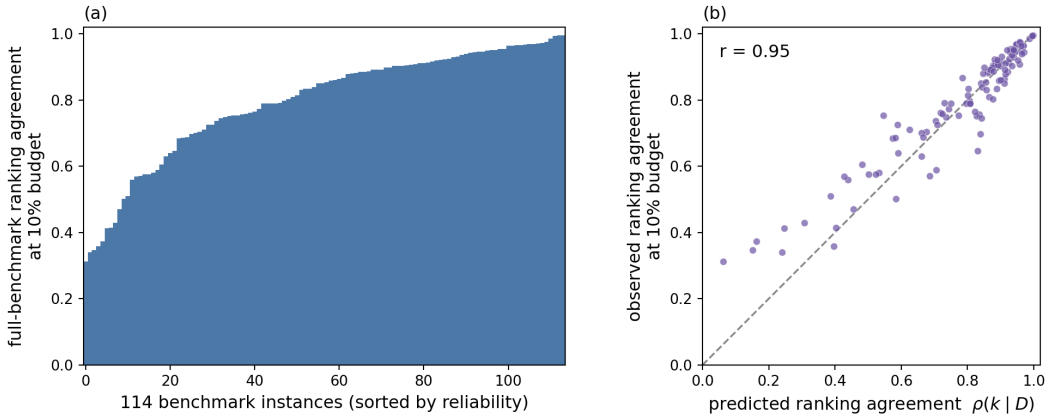


Figure 1: **Evaluation budget does not determine how well a subset preserves the full-benchmark ranking across benchmarks.** (a) At a 10% budget, agreement between the model ranking from a random subset and the full-benchmark ranking ranges from 0.31 to 1.00 across the 114 benchmark instances. Thus, the same evaluation budget can preserve the full-benchmark ranking very differently across benchmarks. (b) This variation is predicted by population-relative item discrimination D: observed agreement is plotted against $\rho(k \mid D)$, using only D and $k = \lceil 0.1N \rceil$ (dashed line: $y = x$).

.

benchmark instances from 11 suites and evaluation sizes from 1% to 20%, the predictor's estimates of full-benchmark ranking agreement correlate with observed agreement at Pearson $r = 0.95$.

The relationship gives us three practical ways to reason about evaluation. Given D and an evaluation size k, it predicts how closely a random subset will match the full-benchmark ranking. Given a target agreement level, it estimates the required number of questions. Finally, its full-budget value gives a rankability ceiling, indicating when additional questions are unlikely to provide substantial gains.

A further consideration is that D is population-relative: the same benchmark can have different discrimination for different model populations. Prior work has shown that benchmark-prediction methods can degrade when transferred across model populations (Zhang et al., 2025). We therefore study population dependence as a boundary condition, examining both transfer of D and the behavior of source-based question selection under population shift.

Our contributions are:

- **Predicting ranking agreement under random subsampling.** We develop and evaluate a Spearman-Brown based predictor that relates evaluation size and population-relative item discrimination to full benchmark ranking agreement under random subsampling. Across 114 benchmark instances from 11 suites and five evaluation sizes spanning 1%-20%, predicted and observed full-benchmark agreement correlate at Pearson $r = 0.95$, with mean absolute error 0.053.

- **A practical evaluation workflow.** We show how D can be estimated from a small source population and used to predict how closely the ranking from a random subset will match the full-benchmark ranking. The predictor can also be inverted to choose how many questions are needed for a desired level of ranking agreement. Finally, the resulting rankability ceiling indicates when evaluating additional questions is expected to provide only limited gains.

- **A population-relative view.** We show that how well a benchmark can distinguish and rank models depends on the model population being evaluated, and that source estimates of D transfer when source and target populations are well matched. We further identify population mismatch as a boundary condition for source-based evaluation methods.

2 Related Work

Efficient LLM evaluation. Most work on efficient evaluation estimates a model's benchmark *score* or missing model-item outcomes from a fraction of the questions. tinyBenchmarks fits a multidimensional IRT model on a population of reference LLMs to estimate benchmark scores from a small set of items (Maia Polo et al., 2024). Related approaches select anchor points, compress benchmarks psychometrically, select items using model disagreement or feature-based criteria, and extend IRT to multiple benchmarks and continuous metrics (Vivek et al., 2024; Kipnis et al., 2025; Rubinstein et al., 2026; Bowyer et al., 2026; Hofmann et al., 2025; Liao et al., 2025). Other methods exploit historical evaluations: CollabEval uses matrix completion as a control variate for efficient estimation of model means and pairwise differences (Fisch et al., 2026), while BayesAME uses information-based active selection and a stopping rule for score estimation (Cordero Encinar et al., 2026). Random sampling also remains a strong baseline: Zhang et al. (2025) find it competitive with more elaborate benchmark-prediction methods. These works primarily estimate scores or outcomes. We instead study when plain random subsampling reliably recovers the *ranking* of a model population.

Active testing and certified evaluation. Active testing selects which test points to label and corrects the resulting selection bias to estimate a *single* model's risk efficiently (Kossen et al., 2021), related approaches have been extended to LLMs and to stratified sampling (Berrada et al., 2025; Liu et al., 2026). Other methods use importance sampling, Bayesian quadrature, adaptive partitioning, or anytime-valid confidence sequences to reduce the cost or determine when evaluation can stop (Yilmaz et al., 2021; Huang et al., 2026; Wang et al., 2025; Hsu & Shekhar, 2026). These methods target efficient estimation of one model's score, whereas we ask when random sampling can recover a model ranking and what benchmark property determines the agreement.

Predicting ranking agreement. SubLIME (Saranathan et al., 2025) studies data-efficient ranking preservation through *subset selection*: it uses results from 5-20 anchor LLMs together with text-based dataset metrics to predict the rank correlation of candidate subsets with the full-benchmark ranking, and selects a subset expected to preserve that ranking. PSN-IRT (Zhou et al., 2026) instead fits a four-parameter IRT model to assess benchmark measurement quality and selects items by Fisher information, with rankings evaluated against external human-preference leaderboards. In information retrieval, test theory has also been used to relate test-collection reliability to ranking stability as the number of queries varies (Bodoff & Li, 2007; Urbano et al., 2013). Our question is different: rather than selecting which questions to evaluate, we characterize how closely a random subset can recover the full-benchmark ranking of a model population. Our closed-form predictor instead uses population-relative item discrimination to relate subset size to ranking agreement and its full-benchmark ceiling.

Classical test theory. Our formulation draws on classical test theory, which models observed scores as true score plus error and uses reliability to quantify measurement consistency (Lord & Novick, 1968). The Spearman-Brown formula gives reliability for a test lengthened by a factor k (Spearman, 1910; Brown, 1910); under our identification of D^2 with single-item reliability, $\rho(k \mid D)$ is the corresponding index of reliability. We do not claim this functional form as new. Rather, we evaluate its use for binary LLM correctness matrices with a population-relative item discrimination D, and test its relationship to full-benchmark ranking agreement across 114 benchmark instances, including its behavior under population shift.

3 Problem Formulation

We first define the evaluation setting and full-benchmark ranking agreement. We then introduce population-relative item discrimination D. (§3.2).

3.1 Evaluation Setting

A *benchmark* consists of N questions (items). For a model population $\mathcal{P}$, we record a binary correctness matrix Y, where $Y_{ij} \in \{0, 1\}$ indicates whether model i answers question j correctly. The

full-benchmark accuracy of model i is

$$a_i = \frac{1}{N} \sum_{j=1}^{N} Y_{ij}, \tag{1}$$

and the resulting ordering of models by a_i defines the full-benchmark reference ranking. In efficient evaluation, each model is evaluated on only a random subset $S \subseteq \{1, \ldots, N\}$ of k questions, with subset accuracy

$$\hat{a}_i^S = \frac{1}{k} \sum_{j \in S} Y_{ij}, \tag{2}$$

and models are ranked by these subset scores. We refer to the fraction k/N as the *evaluation budget*.

Let $\mathbf{a} = (a_1, \ldots, a_M)$ and $\hat{\mathbf{a}}^S = (\hat{a}_1^S, \ldots, \hat{a}_M^S)$ denote the vectors of full-benchmark and subset accuracies across the M models.

We use two ranking-agreement measures (Spearman, 1904). The primary measure is agreement with the full-benchmark ranking,

$$R_{\text{full}}(k) = \mathbb{E}_{S:|S|=k} \left[\text{Spearman} \left(\hat{\mathbf{a}}^S, \mathbf{a} \right) \right]. \tag{3}$$

where the expectation is over random subsets of k questions. This directly measures whether a reduced evaluation preserves the model ordering obtained from the full benchmark. We use *ranking reliability* to mean how consistently a reduced evaluation preserves this model ordering.

As a complementary measure, we compare the subset ranking with a disjoint held-out set. For a sampled subset S, let $\bar{S}$ denote the remaining $N - k$ questions:

$$R_{\text{held}}(k) = \mathbb{E}_{S:|S|=k} \left[\text{Spearman} \left(\hat{\mathbf{a}}^S, \hat{\mathbf{a}}^{\bar{S}} \right) \right]. \tag{4}$$

R_{held} measures agreement between two disjoint evaluations of the same population and provides a supporting check that the observed relationship is not specific to comparison with the full-benchmark reference.

We distinguish two model populations. *Source models* have full benchmark evaluations available and are used to estimate benchmark properties. *Target models* are the models whose ranking we seek efficiently. In the intended setting only the target models' k-question evaluations are required. In our experiments we additionally evaluate them on the remaining questions to measure ranking reliability.

3.2 Item Discrimination

Not all questions are equally informative for ranking models. A question is informative when models that perform well overall are more likely to answer it correctly. We quantify this using the *item discrimination* D_j, the point-biserial correlation (Kornbrot, 2005), over the model population, between correctness on question j and full-benchmark accuracy:

$$D_j = \text{Corr}_{i \in \mathcal{P}} \left(Y_{ij}, a_i \right). \tag{5}$$

We summarize discrimination at the benchmark level by averaging,

$$D(\mathcal{P}) = \frac{1}{N} \sum_{j=1}^{N} D_j. \tag{6}$$

Although we write D for brevity, it is a property of the benchmark-population pair, not of the benchmark alone, because each D_j is computed over the models in $\mathcal{P}$. We write $D_S = D(\mathcal{P}_S)$ and $D_T = D(\mathcal{P}_T)$ for the discrimination of a source and a target population. The same benchmark can have different discrimination for different populations. Consequently, D_S need not equal D_T, which is why source-target transfer is an empirical question (§5.4) rather than an assumption of the predictor.

4 RELIABILITY PREDICTOR

This section develops the predictor from classical measurement theory and the Spearman-Brown relationship (Spearman, 1910; Brown, 1910). The full derivation and assumptions are given in Appendix A.

For a random k-question evaluation, the predictor is

$$\rho(k \mid D) = \sqrt{\frac{kD^2}{kD^2 + 1 - D^2}}. \tag{7}$$

It follows the Spearman-Brown relationship: as more questions are evaluated, measurement noise averages out, while higher discrimination means fewer questions are needed for the same predicted agreement.

Two points on interpretation are important. First, the predictor has no fitted coefficients. D is measured directly, and no coefficient is fit to map (D, k) to observed reliability. Second, the derivation concerns the Pearson correlation between the reduced score and latent ability, whereas experiments evaluate the Spearman agreement between rankings from disjoint question sets. Their correspondence is therefore an empirical validation (§5.2) rather than an exact identity. As an illustration, to reach $\rho = 0.9$, HellaSwag ($D = 0.54$) requires about 11 questions, MMLU-Pro ($D = 0.29$) about 47, and GPQA ($D = 0.06$) on the order of a thousand, so evaluation cost is set not only by the desired reliability but by how informative the benchmark is for the population. This behavior is illustrated in Figure 5 in Appendix A.

Because $\rho(k \mid D)$ is monotonic in k, the predictor can be inverted to give the number of questions required for a desired reliability:

$$k^\star(\rho \mid D) = \frac{\rho^2}{1 - \rho^2} \cdot \frac{1 - D^2}{D^2}. \tag{8}$$

Evaluating the predictor at $k = N$ gives the predicted rankability ceiling $\rho_{\max}(D, N) = \rho(N \mid D)$. This is a prediction of the reliability attainable by the full benchmark, not its agreement with itself, which is trivially one. It is the model-predicted correlation between the full-benchmark measurement and latent ability. We compare it with an independent split-half estimate of rank reproducibility in §5.3.

4.1 PRACTICAL USE

The predictor can be used to plan evaluation before the target models are evaluated. Given a fixed evaluation size k, Equation 7 predicts the attainable reliability. For a desired reliability $\rho^\star$, Equation 8 gives the required evaluation size. The gap between the predicted ceiling and the reliability at budget k is the evaluation headroom,

$$H(k \mid D, N) = \rho_{\max}(D, N) - \rho(k \mid D). \tag{9}$$

Small headroom indicates that more questions are expected to add little ranking reliability, whereas large headroom indicates greater potential benefit from additional evaluation. In practice D is estimated from a small set of source models with full evaluations, so the target population is not required to choose the initial budget. How well this estimate transfers across populations is studied in §5.4. The overall procedure is summarized in Algorithm 1.

5 EXPERIMENTS

We test the reliability predictor and its practical use on a large, heterogeneous collection of LLM benchmarks. The setup is described in §5.1 and the findings follow in §5.2 through §5.4.

5.1 EXPERIMENTAL SETUP

Table 1 summarizes the benchmarks, model populations, budgets, and methods. Full details are in Appendix B.

Table 1: Summary of the experimental setup. Full benchmark and model details are in Appendix B.

Component	Value
Benchmarks	114 instances from 11 suites
Questions per instance	100–14,042 (median 250)
Models per instance	218–6,558 (median 425)
Model-question outcomes	$\sim 9.5 \times 10^7$
Evaluation budgets	$5\%,\ 10\%,\ 20\%$ (main), 1%–20% (Appendix D)
Reliability metric	Spearman rank correlation
Source-pool sizes M	$5, 10, 15, 20, 30, 50$ (main), 2–50 (Appendix E)
Sampling	Uniform random
Comparison methods	mRMR++, Anchor Points, TinyBenchmarks, metabench, AIPW, PPI++

Setup. We evaluate 114 benchmark instances from 11 suites. For each benchmark, we construct disjoint source and target populations. Source models have full evaluations for estimating D, while target models are ranked using random question subsets. Our main analyses use matched source and target populations, with population shift studied separately as a boundary condition (§5.4). We use R_{full} as the primary measure of ranking reliability. A disjoint held-out comparison is used only as a secondary robustness check. Split-half reproducibility provides an independent check of the predicted ceiling. Full construction details are in Appendix C and Appendix D.

Comparison methods. Random subsampling is the primary evaluation strategy. We additionally compare against representative item-selection and score-estimation methods (Table 1). Methods requiring a source population use fully evaluated source models, while random subsampling requires none.

5.2 Predicting Ranking Reliability with D and k

The variation at 10% in Figure 1 persists across budgets. The median full-benchmark ranking agreement of random subsampling is 0.75, 0.85, and 0.93 at 5%, 10%, and 20%, while substantial cross-instance variation remains at each budget (Figure 7 in Appendix D). Increasing the budget raises overall agreement but does not remove the differences in how informative a benchmark-population pair is. This motivates the central test of the predictor, whether D and k predict the reliability of random evaluation. We compare the prediction $\rho(k \mid D)$ (Equation 7) with empirical full-benchmark ranking agreement $R_{\text{full}}(k)$ across 114 instances and budgets from 1% to 20%.

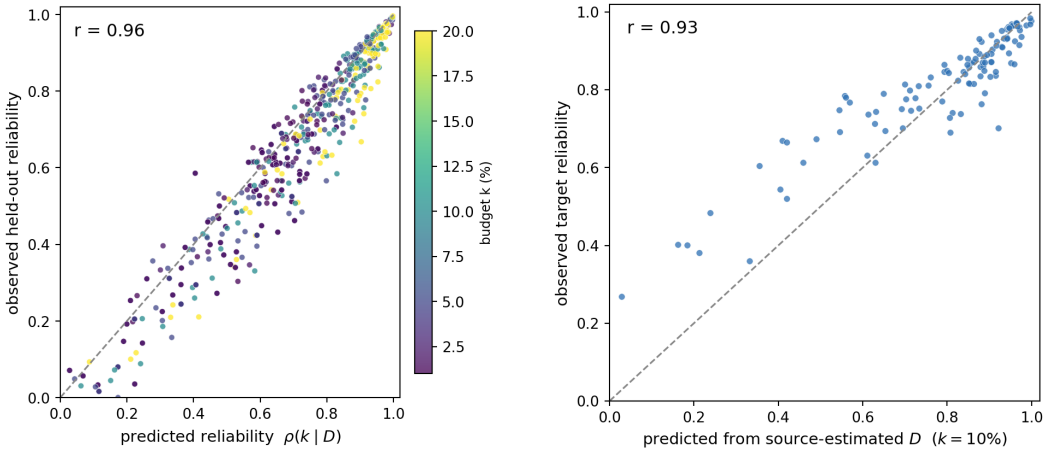
The predictor captures cross-benchmark variation. The correlation remains strong within each discrimination range (Table 2), showing that the predictor captures substantial variation in ranking reliability beyond evaluation size alone. As a secondary check, predicted and observed reliabilities track each other closely at held out reliability (Figure 2a). The overall Pearson correlation is $r = 0.96$ with MAE $= 0.060$, and the fit is comparable within all three discrimination ranges (Table 3). Benchmarks with different population-relative discrimination therefore exhibit systematically different ranking reliability at the same budget, and the predictor captures much of this variation using only D and k. The prediction is systematically optimistic, exceeding $R_{\text{held}}(k)$ by 0.048 on average. We therefore read the predictor as well calibrated for relative variation in reliability, with a small, characterized positive offset in its absolute level.

The relationship is not explained by budget alone. Both quantities increase with k, so their association could arise from a shared dependence on budget. However, the cross-benchmark correlation remains strong within each fixed budget ($r = 0.93$–0.97), and the partial correlation after controlling for k is $r = 0.956$ (Appendix D). To address the non-independence of instances, we also verify that the fit persists when aggregates and their constituent splits are not pooled (Appendix D). The relationship therefore reflects benchmark-level structure captured by D, not merely the effect of evaluating more questions.

D can be estimated before evaluating the target. The practical value of the predictor is that D can be estimated from source models before the target models are fully evaluated. Using D

Table 2: Agreement between the predicted reliability $\rho(k \mid D)$ and observed full-benchmark ranking agreement $R_{\text{full}}(k)$ across 114 instances and budgets from 1% to 20%. (Appendix D).

D range	n	Pearson r	MAE	bias (pred − obs)
high ($D > 0.27$)	360	0.93	0.036	$+0.007$
mid (0.14–0.27)	140	0.90	0.068	-0.017
low ($D < 0.14$)	70	0.87	0.111	-0.097
overall	570	0.95	0.053	-0.012



(a) (b)

Figure 2: Validation of the reliability predictor. (a) Predicted $\rho(k \mid D)$ versus observed held-out reliability across 114 instances and budgets from 1% to 20% ($r = 0.96$). (b) Source-population D predicts target-population reliability at 10% budget ($r = 0.93$).

estimated from a disjoint source population to predict the target population's ranking reliability at a 10% budget gives $r = 0.93$, close to the $r = 0.92$ obtained using the target population's own D. The two discrimination estimates themselves correlate at $r = 0.985$ across instances (Figure 2b, Appendix E). The source population need not be large. Estimating D from only 15 source models predicts held-out reliability at $r = 0.95$. The main requirement is thus not full evaluation of the target, but a source population that yields an informative estimate of D. How this estimate behaves when the populations differ substantially is examined in §5.4.

5.3 Evaluation Budget, Headroom, and Rankability

Beyond predicting reliability at a fixed budget, the predictor provides the budget required for a desired reliability and the maximum reliability predicted for a benchmark-population pair. We use these to characterize headroom and population-dependent rankability.

The required budget varies widely, and some targets are unreachable. Inverting the predictor (Equation 8) for $\rho^\star = 0.9$ gives a median required budget of 42 questions (15% of the benchmark), with an interquartile range of 7%–43% across pairs (Figure 3a). For 15% of pairs the predicted $k^\star$ exceeds the full benchmark size N, meaning $\rho^\star = 0.9$ lies above the predicted ceiling and cannot be reached by random evaluation under the model. For these pairs, the limitation is not simply evaluation budget. The predicted rankability ceiling itself lies below the target reliability.

The full-budget prediction defines a rankability ceiling. The predicted ceiling $\rho_{\max}(D, N) = \rho(N \mid D)$ agrees closely with an independent split-half estimate of rank reproducibility ($r = 0.94$, MAE $= 0.017$, Figure 3b), where the split-half correlation is placed on the reliability scale by the Spearman-Brown correction. Because the two halves share no items, this is an independent check

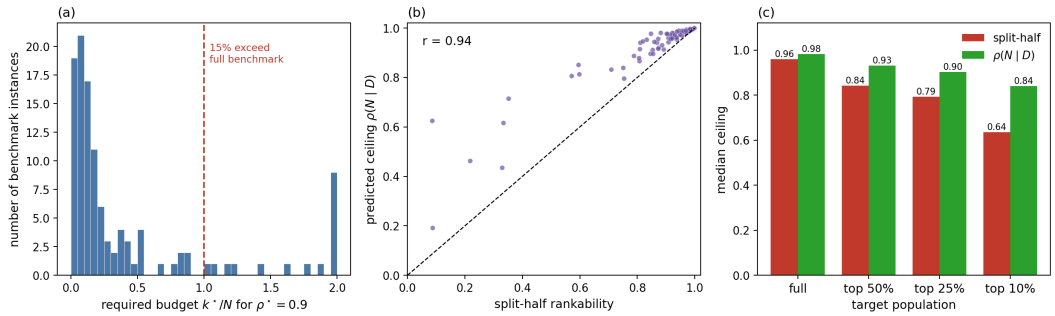


Figure 3: Consequences of the reliability predictor. (a) Required budget $k^\star$ for $\rho^\star = 0.9$ as a fraction of N. For 15% of pairs $k^\star > N$. (b) Predicted rankability ceiling $\rho_{\max}(D, N)$ versus an independent split-half rankability estimate. (c) Median ceiling as the target population narrows to stronger models.

rather than a restatement of the predictor, and it supports using $\rho_{\max}(D, N)$ as a practical rankability reference.

Headroom separates budget-limited from rankability-limited evaluation. At $k = 10\%$ the median headroom $H(k \mid D, N)$ is 0.12, compared with 0.09 for high-D, 0.26 for mid-D, and 0.38 for low-D pairs. Small headroom means more random questions add little. Large headroom means additional evaluation may help. Together with the ceiling, this distinguishes **budget-limited** pairs (low current reliability, high ceiling, more questions help) from **rankability-limited** pairs (low current reliability, low ceiling, they do not).

Rankability depends on the target population. Restricting the target to progressively stronger, more closely matched models lowers the median rankability ceiling. Ranking reproducibility is thus population-relative. A benchmark can provide much less information for separating a narrow ability band than for ranking a broad population. As the target population narrows, both empirical and the predicted rankability ceiling decrease (Figure 3c). The predictor remains optimistic under population shift, with the gap increasing for narrower populations, indicating that D alone on one model population does not fully transfer to the different population.

5.4 Population Dependence: A Boundary Condition

Because D is population-relative, the usefulness of an efficient evaluation depends on which models estimate D and which are ranked. We summarize how this affects the estimate of D and the value of source-based selection. The full analysis is in Appendices E and F.

Estimating D requires a representative source pool. Random evaluation of the target is source-independent, but the estimate of D is not. Across source-selection strategies, pools that span the ability range recover values close to the full-population estimate ($D \approx 0.30$). Random, ability-spread, by-size, and cluster sampling give $D \approx 0.30$–0.38, whereas selecting only the extremes gives $D = 0.54$ and selecting only the top or bottom band gives $D = 0.11$ and 0.08 (Appendix E). Concentrating the source pool in a narrow region therefore distorts D.

Source-based selection can lose its advantage under population shift. Under matched populations ($M = 30$ source models, a disjoint random 25% of the remaining models as target, 10% budget), question selection gives modest gains over random. mRMR++ is the strongest, at $+0.033$ mean reliability over random, and agrees with random within 0.05 on 68% of instances (Appendix F). As the target is restricted to progressively stronger models while the source is held fixed, however, source-fitted selectors degrade. At a 10% budget on the top-25% target population, reliability falls to 0.29 for mRMR++ and 0.16 for metabench, versus 0.59 for random sampling, while the estimator and anchor methods (AIPW, PPI++, Anchor Points) remain comparatively stable (Figure 4). The effect varies across the tested population-shift spectrum rather than switching at a binary interpolation/extrapolation boundary. The full source-target grid is in Appendix F.

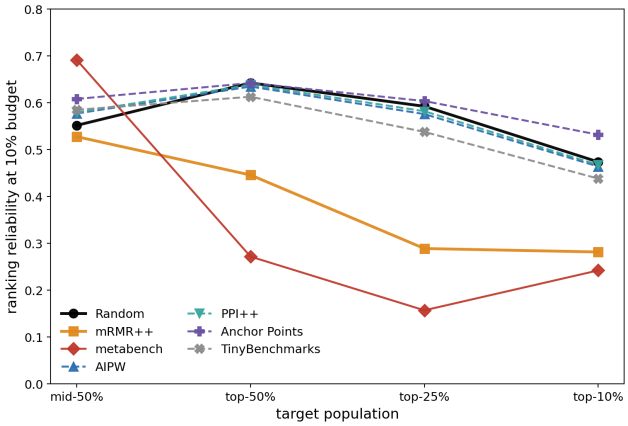


Figure 4: Population shift affects evaluation methods differently. At a 10% budget, the target population is restricted to higher-ability subsets while the source remains fixed. Random sampling and estimator/anchor methods remain stable, whereas source-fitted selection methods (mRMR++, metabench) degrade under increasing population mismatch.

D is a partial diagnostic. The strongest degradation coincides with narrow target populations whose low D_T indicates weak item-level signal for separating the target models. This gives a partial explanation for why source-fitted methods can fail under shift, but it is not complete. Methods with similar D_T can still behave differently, and D alone does not predict the magnitude of every failure. We therefore treat population-relative discrimination as an additional diagnostic for when source-based efficient evaluation is likely to transfer, not as a full account of selection behavior.

6 Discussion

Practical implications. Evaluation planning can be based on $\rho(k \mid D)$ and headroom $H(k \mid D, N)$. High predicted reliability suggests that a small random subset may suffice, low reliability with substantial headroom motivates a larger budget, while a low predicted ceiling indicates limited gains from additional random evaluation. Because D is population-relative, source-based question selection may also lose its advantage when the source and target populations differ.

Limitations. The predictor follows from a classical measurement model that assumes a dominant single ability, approximately independent item errors, and an effective per-item signal-to-noise level summarized by D. Real benchmarks can violate these through heterogeneous or correlated items, so the predictor is an empirical predictive approximation rather than an exact identity, with a small characterized positive bias. The workflow still requires a representative source population with full evaluations, so estimating D for a newly introduced benchmark with no such pool incurs an initial cost. Our comparison methods are not exhaustive, and the results do not establish random sampling as universally optimal; selection methods can behave differently under population shift. Results are limited to the benchmark families and populations studied. The predictor should eventually be tested on newer benchmarks, larger models, and non-binary scoring.

7 Conclusion

Subset size alone does not determine how well random evaluation recovers an LLM model ranking. Across 114 benchmark instances, population-relative item discrimination D explains substantial cross-benchmark variation and provides a simple predictor for subset-to-full ranking agreement, required evaluation size, and the full-benchmark rankability ceiling. Because D depends on the model population, its estimate may not transfer under population shift. These results provide a practical way to assess when random evaluation is sufficient and when benchmark properties limit the achievable ranking agreement.

AI USE STATEMENT

We used generative AI tools during the preparation of this paper for literature discovery and analysis and feedback on experimental methodology and polishing of the manuscript. All AI-assisted suggestions, code-related suggestions, and manuscript text were reviewed and verified by the authors. The authors take responsibility for the final content, claims, analyses, and artifacts in this work.

REPRODUCIBILITY STATEMENT

We provide an anonymous supplementary package containing the analysis code and derived statistics required to reproduce the reported numerical results and figures. The package documents the software environment, random seeds, and commands for reproducing the main analyses. Full benchmark and population construction details, the reliability-predictor derivation, additional reliability analyses, source-population estimation, and population-shift experiments are provided in the appendices. The raw model-question correctness matrices are obtained from the external benchmark sources described in the paper and are not redistributed in the supplementary package.

REFERENCES

Gabrielle Berrada, Jannik Kossen, Freddie Bickford Smith, Muhammed Razzak, Yarin Gal, and Tom Rainforth. Scaling up active testing to large language models. In *Advances in Neural Information Processing Systems*, 2025. URL https://arxiv.org/abs/2508.09093.

David Bodoff and Pu Li. Test theory for assessing IR test collections. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 367–374, 2007. doi: 10.1145/1277741.1277805.

Sam Bowyer, Acyr Locatelli, and Kris Cao. Efficient benchmarking is just feature selection and multiple regression. *arXiv preprint arXiv:2605.25773*, 2026. URL https://arxiv.org/abs/2605.25773.

William Brown. Some experimental results in the correlation of mental abilities. *British Journal of Psychology*, 3(3):296–322, 1910. doi: 10.1111/j.2044-8295.1910.tb00207.x.

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, 2018. URL https://arxiv.org/abs/1803.05457. arXiv:1803.05457.

Paula Cordero Encinar, Taylan Cemgil, Arnaud Doucet, Virginia Aglietti, and Silvia Chiappa. BayesAME: Bayesian active model evaluation, 2026. URL https://arxiv.org/abs/2607.27023.

Adam Fisch, Daniel Deutsch, Joshua Maynez, Alekh Agarwal, Jonathan Berant, William Cohen, Amir Globerson, and Jacob Eisenstein. CollabEval: Statistically efficient collaborative model evaluation via matrix completion, 2026. URL https://arxiv.org/abs/2607.05046.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations (ICLR)*, 2021a. URL https://arxiv.org/abs/2009.03300.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. In *Advances in Neural Information Processing Systems*, 2021b. URL https://arxiv.org/abs/2103.03874.

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. URL https://openreview.net/forum?id=mxcCg9YRqj.

Chia-Yu Hsu and Shubhanshu Shekhar. Efficient sequential evaluation of large language models, 2026. URL https://arxiv.org/abs/2607.17409.

Yizheng Huang, Wenjun Zeng, Aditi Kumaresan, and Zi Wang. ProEval: Proactive failure discovery and efficient performance estimation for generative AI evaluation, 2026. URL https://arxiv.org/abs/2604.23099.

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.

Diana Kornbrot. *Point Biserial Correlation*. John Wiley & Sons, 2005. doi: 10.1002/0470013192.bsa485.

Jannik Kossen, Sebastian Farquhar, Yarin Gal, and Tom Rainforth. Active testing: Sample-efficient model evaluation. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5753–5763. PMLR, 2021.

Lele Liao, Qile Zhang, Ruofan Wu, and Guanhua Fang. Toward a unified framework for data-efficient evaluation of large language models, 2025. URL https://arxiv.org/abs/2510.04051.

Fan Lin, Shuyi Xie, Yong Dai, Wenlin Yao, Tianjiao Lang, and Yu Zhang. Idgen: Item discrimination induced prompt generation for llm evaluation. In *Advances in Neural Information Processing Systems*, 2024.

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*, 2022. URL https://aclanthology.org/2022.acl-long.229/.

Zeli Liu, Jiancheng Zhang, Cong Liu, and Yinglun Zhu. Active testing of large language models via approximate Neyman allocation, 2026. URL https://arxiv.org/abs/2605.10075.

Frederic M. Lord and Melvin R. Novick. *Statistical Theories of Mental Test Scores*. Addison-Wesley, Reading, MA, 1968.

Felipe Maia Polo, Lucas Weber, Leshem Choshen, Yuekai Sun, Gongjun Xu, and Mikhail Yurochkin. tinyBenchmarks: evaluating LLMs with fewer examples. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 34303–34326. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/maia-polo24a.html.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*, 2024. URL https://arxiv.org/abs/2311.12022.

Alexander Rubinstein, Benjamin Raible, Martin Gubri, and Seong Joon Oh. DISCO: Diversifying sample condensation for efficient model evaluation. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=SoOgBHa3dZ.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. WinoGrande: An adversarial winograd schema challenge at scale. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020. URL https://arxiv.org/abs/1907.10641.

Gayathri Saranathan, Cong Xu, Mahammad Parwez Alam, Tarun Kumar, Martin Foltin, Soon Yee Wong, and Suparna Bhattacharya. SubLIME: Subset selection via rank correlation prediction for data-efficient LLM evaluation. 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. 30572–30593, Vienna, Austria, July 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.acl-long.1477. URL https://aclanthology.org/2025.acl-long.1477/.

Charles Spearman. "General Intelligence," Objectively Determined and Measured. *The American Journal of Psychology*, 15(2):201–293, 1904. doi: 10.2307/1412107.

Charles Spearman. Correlation calculated from faulty data. *British Journal of Psychology*, 3(3): 271–295, 1910. doi: 10.1111/j.2044-8295.1910.tb00206.x.

Zayne Sprague, Xi Ye, Kaj Bostrom, Swarat Chaudhuri, and Greg Durrett. MuSR: Testing the limits of chain-of-thought with multistep soft reasoning. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2310.16049.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V. Le, Ed H. Chi, Denny Zhou, and Jason Wei. Challenging BIG-Bench tasks and whether chain-of-thought can solve them. In *Findings of the Association for Computational Linguistics: ACL 2023*, 2023. URL https://arxiv.org/abs/2210.09261.

Julián Urbano, Mónica Marrero, and Diego Martín. On the measurement of test collection reliability. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '13, pp. 393–402, New York, NY, USA, 2013. Association for Computing Machinery. doi: 10.1145/2484028.2484038.

Rajan Vivek, Kawin Ethayarajh, Diyi Yang, and Douwe Kiela. Anchor points: Benchmarking models with much fewer examples. In Yvette Graham and Matthew Purver (eds.), *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1576–1601, St. Julian's, Malta, March 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.eacl-long.95. URL https://aclanthology.org/2024.eacl-long.95/.

Ganghua Wang, Zhaorun Chen, Bo Li, , and Haifeng Xu. Cer-Eval: Certifiable and cost-efficient evaluation framework for LLMs, 2025. URL https://arxiv.org/abs/2505.03814.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. MMLU-Pro: A more robust and challenging multi-task language understanding benchmark. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2024. URL https://arxiv.org/abs/2406.01574.

Emine Yilmaz, Peter Hayes, Raza Habib, Jordan Burgess, and David Barber. Sample efficient model evaluation, 2021. URL https://arxiv.org/abs/2109.12043.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019. URL https://aclanthology.org/P19-1472/.

Guanhua Zhang, Florian E. Dorner, and Moritz Hardt. How benchmark prediction from fewer data misses the mark. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://openreview.net/forum?id=o3bftqj17e.

Hongli Zhou, Hui Huang, Ziqing Zhao, Lvyuan Han, Huicheng Wang, Kehai Chen, Muyun Yang, Wei Bao, Jian Dong, Bing Xu, Conghui Zhu, Hailong Cao, and Tiejun Zhao. Lost in benchmarks? Rethinking large language model benchmarking with item response theory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 35085–35093, 2026. doi: 10.1609/aaai.v40i41.40814.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models, 2023. URL https://arxiv.org/abs/2311.07911.

APPENDIX

A DERIVATION OF THE RELIABILITY PREDICTOR

This section provides the theoretical motivation for the reliability predictor (Section 4), using the definitions of Y, a_i, and D from Section 3.2. The derivation follows the classical reliability framework and makes the assumptions explicit.

A.1 CLASSICAL RELIABILITY

Our predictor is based on the classical Spearman-Brown formula (Spearman, 1910; Brown, 1910), which relates the reliability of a measurement to the number of comparable items used to construct it. If a single item has reliability r, the reliability of a composite of k items is

$$R_k = \frac{kr}{1 + (k-1)r}. \tag{10}$$

The corresponding index of reliability, defined as the correlation between the observed and true scores, is $\sqrt{R_k}$ (Lord & Novick, 1968). Our application identifies the squared population-relative item discrimination D^2 with an effective single-item reliability. Substituting $r \approx D^2$ into Equation 10 gives the predictor evaluated in this work. This identification is an approximation, not a new derivation of the Spearman-Brown relationship.

A.2 RELATING DISCRIMINATION TO RELIABILITY

The item discrimination D_j measures the association between correctness on item j and full-benchmark accuracy a. We use D_j to approximate the item's association with the underlying model ability, and hence its effective reliability:

$$D_j = \mathrm{Corr}(Y_j, a), \qquad r_j \approx D_j^2. \tag{11}$$

This approximation is motivated by classical reliability theory (Lord & Novick, 1968): when full-benchmark accuracy is a sufficiently precise proxy for latent ability, the squared item-accuracy correlation approximates the item's squared association with that ability. It is not an exact identity, particularly for heterogeneous items or a noisy full-benchmark score.

We use the benchmark-level statistic $D = \frac{1}{N}\sum_j D_j$ to define the effective reliability $r \approx D^2$. In general, $\frac{1}{N}\sum_j D_j^2 \neq (\frac{1}{N}\sum_j D_j)^2$ thus, this scalar approximation does not follow exactly from averaging item reliabilities. Whether it adequately predicts ranking agreement is evaluated empirically in Section 5.2.

Remark on item-total and item-rest correlation. Our main definition correlates Y_j with the full accuracy a, which includes item j. A leave-one-out (item-rest) correlation removes this self-inclusion term. For large N the two coincide, and we use the item-total form throughout the main text. Appendix E uses the item-rest form for $\widehat{D}$ estimation, and we verify the two give consistent conclusions.

A.3 RELIABILITY AS A FUNCTION OF EVALUATION BUDGET

Under classical test theory, the reliability of a composite of k comparable items follows the Spearman-Brown relationship in Equation 10. Substituting the effective single-item reliability $r \approx D^2$ gives

$$\rho(k \mid D) = \sqrt{\frac{kD^2}{kD^2 + 1 - D^2}}\,. \tag{12}$$

This expression is a classical reliability index applied using our population-relative discrimination statistic, the functional form itself is not a new theoretical result. Its theoretical interpretation is the

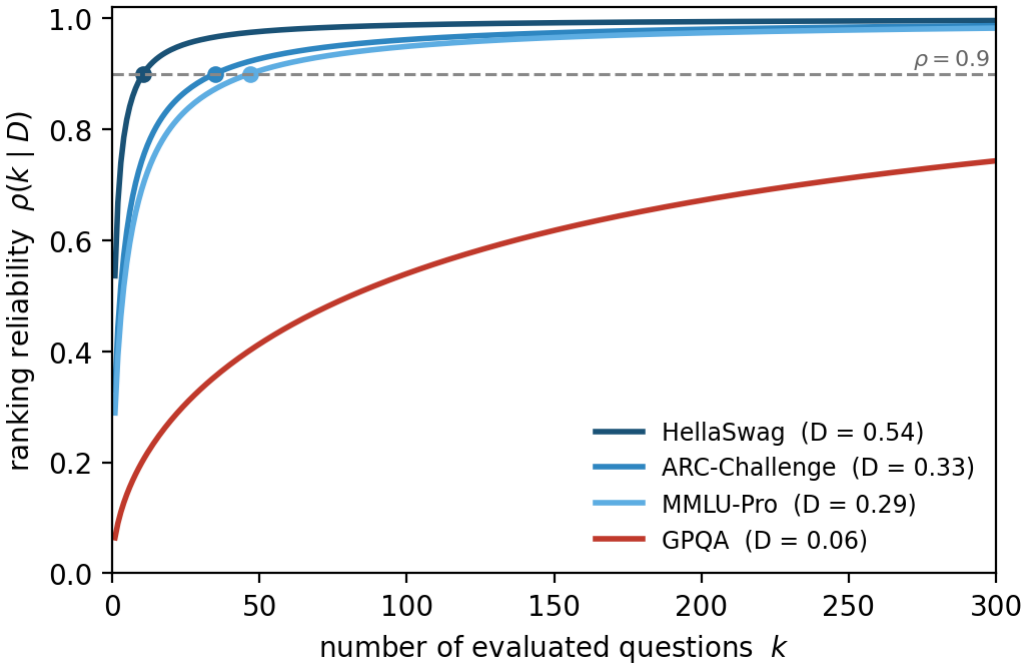


Figure 5: Predicted ranking reliability $\rho(k \mid D)$ as a function of the number of evaluated questions k, for several discrimination values. Higher discrimination reaches a given reliability with fewer evaluated questions. For example, to reach $\rho = 0.9$, HellaSwag ($D = 0.54$) requires about 11 questions, MMLU-Pro ($D = 0.29$) about 47, and GPQA ($D = 0.06$) on the order of a thousand.

Pearson correlation between a reduced measurement and latent ability. Our primary empirical target, however, is Spearman agreement between the ranking from a random subset and the full-benchmark ranking. These quantities are not identical in general. We therefore evaluate the relationship empirically in Section 5.2.

For the secondary held-out analysis, the two rankings are obtained from disjoint subsets of sizes k and $N - k$. Under the classical measurement model, their expected agreement is approximated by $\rho(k \mid D)\rho(N - k \mid D)$, motivating the two-form analysis in Appendix D.

A.4 Required Evaluation Budget

Inverting Equation 12 for a target $\rho^\star$,

$$k^\star(\rho^\star \mid D) = \frac{\rho^{\star 2}}{1 - \rho^{\star 2}} \frac{1 - D^2}{D^2} \,. \tag{13}$$

The operational budget is $\lceil k^\star \rceil$. If $k^\star > N$ the target exceeds the predicted full-budget value and is not attainable under the model, which is the basis for the feasibility analysis in Section 5.3.

A.5 Scope of the Derivation

The predictor combines the classical Spearman-Brown relationship with the approximation $r \approx D^2$. Its use for LLM ranking reliability is not an exact theoretical consequence of classical test theory: items may be heterogeneous or correlated, correctness is binary, and the empirical target is Spearman agreement with a finite full-benchmark ranking rather than correlation with latent ability. The empirical validation in Section 5.2 and the analyses in Appendix D assess the accuracy and limitations of this approximation, including under population shift.

Algorithm 1 Using the reliability predictor for efficient benchmark evaluation.

Require: Benchmark with N questions, source models $\mathcal{P}_S$ with full evaluations, fixed budget k or target reliability $\rho^\star$

Estimate item discriminations D_j over $\mathcal{P}_S$
$D \leftarrow \frac{1}{N}\sum_{j=1}^{N} D_j$
if a fixed budget k is given **then**
    $\hat{\rho} \leftarrow \rho(k \mid D)$
else
    $k^\star \leftarrow \dfrac{\rho^{\star 2}}{1-\rho^{\star 2}} \cdot \dfrac{1-D^2}{D^2}$
end if
$\rho_{\max} \leftarrow \rho(N \mid D)$
$H \leftarrow \rho_{\max} - \rho(k \mid D)$
if H is small **then**
    Increasing k is expected to provide limited additional reliability
else
    Consider increasing k or using a question-selection strategy
end if
Evaluate each target model on the resulting question set and rank by score
return $\hat{\rho}$, $k^\star$, $\rho_{\max}$, H, and target-model ranking

Table 3: Agreement between the predicted reliability $\rho(k \mid D)$ and observed held-out reliability $R_{\text{held}}(k)$ across 114 instances and budgets from 1% to 20%.

D range	n	Pearson r	MAE	bias (pred − obs)
high ($D > 0.27$)	360	0.91	0.054	+0.044
mid (0.14–0.27)	140	0.94	0.067	+0.054
low ($D < 0.14$)	70	0.91	0.072	+0.054
overall	570	0.96	0.060	+0.048

B Benchmark Collection and Statistical Units

This section documents the benchmark collection and the unit of analysis. We apply Y, a_i, and D as defined in Section 3.

B.1 Benchmark Suites and Benchmark Instances

Our analyses use 114 benchmark instances drawn from 11 benchmark suites. A *benchmark suite* is a named evaluation such as MMLU, BBH, or GPQA. A *benchmark instance* is a separately scored question set with its own model-question correctness matrix, for which we can compute discrimination and ranking reliability. Several suites are natively organized into subject or task-level splits, which we retain as separate instances because they are independently scored and can exhibit substantially different discrimination and reliability. For example, MMLU contributes one aggregate evaluation and 57 subject-level instances, giving 58 instances. BBH, MMLU-Pro, MATH, and MuSR are treated analogously. The 114 instances are the observational units of our analysis, not 114 unrelated benchmark families. Because aggregates share items with their constituent splits, we do not treat instances from the same suite as statistically independent (Appendix D).

B.2 Data Scale

Across the 114 instances, questions per instance range from 100 to 14,042 (median 250) and model populations from 218 to 6,558 (median 425), giving approximately 9.5×10^7 model-question outcomes. Each instance is analyzed using its native scored question set and available model population. Source and target construction is described in Appendix C.

Table 4: Composition of the evaluation collection. "Instances" includes aggregate evaluations and their native subject or task splits. Reported D is the median instance-level discrimination within each suite. References. MMLU (Hendrycks et al., 2021a), BBH (Suzgun et al., 2023), MMLU-Pro (Wang et al., 2024), MATH (Hendrycks et al., 2021b), MuSR (Sprague et al., 2024), GPQA (Rein et al., 2024), ARC-Challenge (Clark et al., 2018), HellaSwag (Zellers et al., 2019), IFEval (Zhou et al., 2023), TruthfulQA (Lin et al., 2022), WinoGrande (Sakaguchi et al., 2020).

Benchmark suite	Instances	Questions/instance	Models	Median D
MMLU	58	100–14,042	425	0.316
BBH	25	146–5,761	446	0.321
MMLU-Pro	15	381–12,032	448	0.289
MATH	6	132–894	447	0.294
MuSR	4	250–756	448	0.101
GPQA	1	1,192	447	0.064
ARC-Challenge	1	1,172	218	0.330
HellaSwag	1	10,042	5,227	0.536
IFEval	1	541	448	0.475
TruthfulQA	1	817	5,228	0.367
WinoGrande	1	1,267	6,558	0.304
Total	**114**	100–14,042	218–6,558	0.304

B.3 Empirical Discrimination Regimes

Several analyses group instances into low-, mid-, and high-discrimination regimes. These are empirical descriptive bins, not theoretical categories. The boundaries $D = 0.14$ and $D = 0.27$ organize the observed distribution and are not universal thresholds.

Table 5: Empirical discrimination regimes used in the main analyses. Boundaries are descriptive.

Regime	D range	Instances
Low	$D < 0.14$	14
Mid	$0.14 \leq D \leq 0.27$	28
High	$D > 0.27$	72
Total	—	**114**

B.4 Within-Suite Heterogeneity: MMLU

Figure 6 shows discrimination across the 57 MMLU subject instances, spanning $D = 0.05$ to $D = 0.48$ while the aggregate MMLU evaluation has $D = 0.289$. The aggregate does not capture this variation, which is directly relevant to the predictor. Different subject-level question sets provide very different amounts of information about the same population, so treating them as separate instances preserves the variation the predictor explains.

C Source and Target Population Construction

For each benchmark instance, the available models form a pool $\mathcal{P}$ of size m. We order models by full-benchmark accuracy a_i and index them $0, \ldots, m-1$ from weakest to strongest. Source models receive full evaluations and are used only to estimate benchmark properties, chiefly D and, for selection methods, the selected subset. Target models are ranked from a reduced evaluation. Source and target sets are always disjoint.

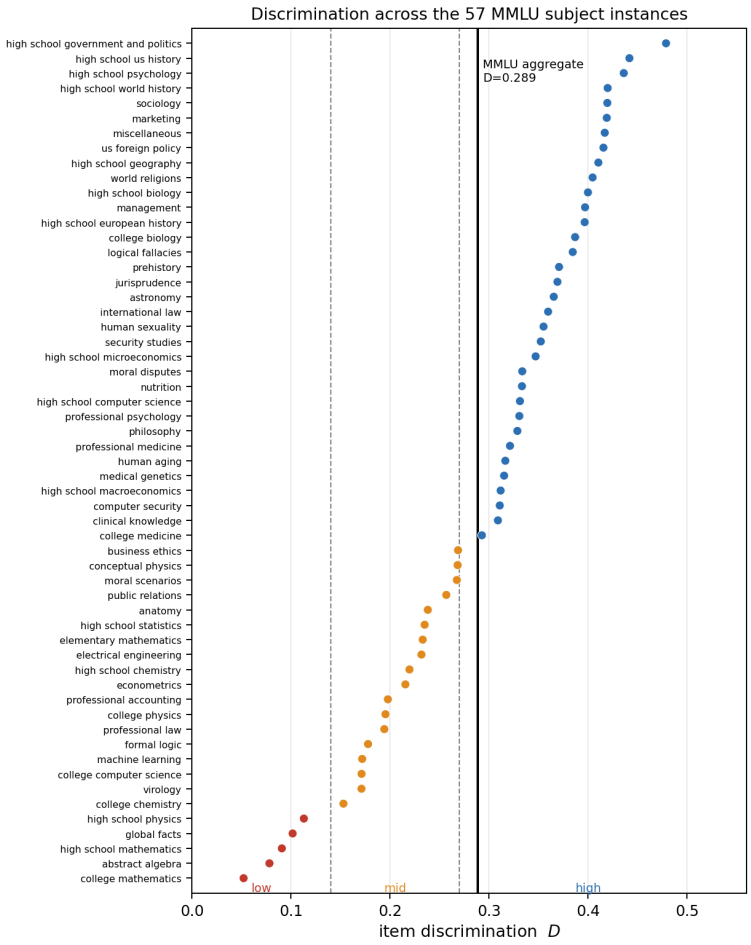


Figure 6: **Within-suite variation in discrimination for MMLU.** Discrimination D across the 57 MMLU subject instances. The solid line marks the aggregate value ($D = 0.289$), and dashed lines mark the empirical regime boundaries. Subject-level D ranges from 0.05 to 0.48.

C.1 Matched (Interpolation) Regime

Source and target are sampled without replacement from the same pool. For a permutation π of the m models, the source is $\pi[0{:}M]$ and the target a random subset of the remainder. The canonical setting is $M = 30$, $n_{\text{tgt}} = \max(12, \lceil 0.25(m - M) \rceil)$, and $k = 0.10N$, with 6 deterministic seeds ($\texttt{RandomState}(2000 + s)$). Secondary budgets $k \in \{0.05N, 0.20N\}$ are also recorded.

C.2 Source-Pool Size and Cost Variants

To study how the source pool affects end-to-end evaluation we sweep the source size M and the fraction f of questions the source models are evaluated on, with source $= \pi[0{:}M]$, target $= \pi[M{:}M + n_{\text{tgt}}]$, $n_{\text{tgt}} = \max(5, \lceil 0.25m \rceil)$, over $M \in \{5, 10, 15, 20, 30, 50\}$ and $f \in \{0.10, 0.25, 0.50, 1.00\}$ at $k = 0.10N$ (5 seeds, $\texttt{RandomState}(1000s + 7)$). Building the source pool costs $M \cdot f$ benchmark-equivalent passes. This is distinct from the discrimination-estimation analysis, which redraws $M \in \{2, 3, 5, 8, 10, 15, 20, 30, 50\}$ models and recomputes $\widehat{D}$ (300 seeds per M) with no method fitting.

C.3 Population-Shift Bands

For the population-shift analysis the target is an ability band and the source is its complement. With models ordered weakest-to-strongest, mid-half $= [m/4, 3m/4)$, top-half $= [m/2, m)$, top-quartile $= [3m/4, m)$, and top-decile $= [9m/10, m)$. Each method is fit on the complement and evaluated on the band over $k \in \{0.02, 0.05, 0.10, 0.20\}N$ (3 seeds, minimum source size 20). Because the source is the complement of the target, its size changes with the band, so we read this analysis as a controlled, monotone axis of population shift rather than a fixed-source experiment. We use "extrapolation" narrowly for this constructed ability shift, not for arbitrary distribution shift.

Rankability populations. The population-restricted rankability analysis (Section 5.3) uses the same accuracy ordering, evaluating the full population and the top $50/25/10\%$, middle 50%, and bottom $50/25\%$ bands to compute population-specific D and predicted ceilings. These are not source-target splits.

C.4 Ability-Disaggregated Distribution Study

A finer study, used only for the heatmaps of Appendix F, crosses target ability quartiles ($\text{Q1}_{\text{weak}}, \ldots, \text{Q4}_{\text{strong}}$) with source ability bands (percentile windows all_weak $= [0, 40]\%$, lower_mid $= [20, 60]\%$, mid $= [30, 70]\%$, upper_mid $= [40, 80]\%$, all_strong $= [60, 100]\%$, plus spread and random), at $k = 0.10N$ with 8 seeds. For each target quartile a random half is held out as the target.

C.5 What Is Held Fixed

Within each comparison the following are fixed so the population relationship is isolated. The benchmark and its full-benchmark reference ranking, the budget k, the sampling and fitting protocol, the number of source models M where applicable, and the target scoring procedure (Spearman against the target's own full-benchmark ranking).

D Additional Budget and Reliability Analyses

This section reports the finer budget sweep behind the predictor validation. Held-out reliability scores a random k-question subset against a disjoint held-out set, so the two rankings share no items. Statistics are computed for the full population, averaged over seeds to one observation per benchmark-budget cell, giving $n = 114 \times 5 = 570$ observations. We test the predictor as a predictor with no fitted coefficient.

D.1 Predictor Validation Across Budgets

Sweeping $k \in \{1, 2, 5, 10, 20\}\%$ of N, the prediction $\rho(k \mid D)$ agrees with observed held-out reliability at $r = 0.963$, MAE $= 0.060$, bias $= +0.048$ (Table 6, Figure 8a).

Table 6: Predictor validation by discrimination band, pooled over the five budgets.

Band	D range	n	Pearson r	MAE	Bias
High	$D > 0.27$	360	0.912	0.054	$+0.044$
Mid	$0.14 \le D \le 0.27$	140	0.939	0.067	$+0.054$
Low	$D < 0.14$	70	0.912	0.072	$+0.054$
Overall	—	570	0.963	0.060	$+0.048$

D.2 Fixed-Budget Analysis and Dependence

Holding k constant, the cross-benchmark fit remains strong at every budget ($r = 0.930$–0.969, Table 7) with bias close to $+0.05$, so the predictor predicts the relative reliability of benchmarks at

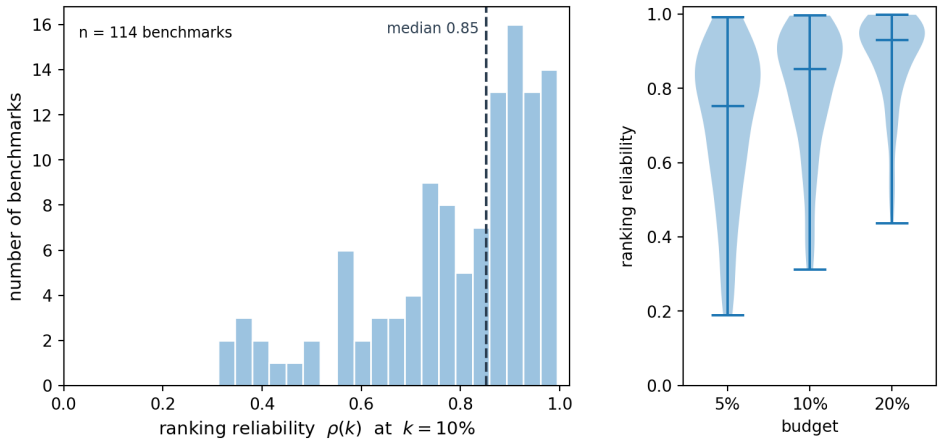


Figure 7: Ranking reliability of random subsampling across evaluation budgets. The median full-benchmark ranking agreement increases from 0.75 at 5% to 0.85 at 10% and 0.93 at 20%, while substantial variation across benchmark-population pairs remains at each budget.

a fixed budget, not merely their common increase with k. Because instances are not independent, since aggregates share items with their splits, we also recompute the pooled fit excluding aggregate instances and find it materially unchanged, and we report suite-level medians alongside the pooled value.

Table 7: Cross-benchmark predictor validation at each fixed budget ($n = 114$ per row).

Budget k	n	Pearson r	MAE	Bias
1%	114	0.930	0.070	+0.050
2%	114	0.941	0.069	+0.049
5%	114	0.966	0.057	+0.046
10%	114	0.969	0.052	+0.046
20%	114	0.969	0.051	+0.050

D.3 Partial Correlation Controlling for Budget

Regressing predicted and observed reliability on the budget and correlating the residuals gives $r_{\text{partial}} = 0.956$, close to the pooled $r = 0.963$. The agreement therefore reflects benchmark-level structure captured by D, not the shared increase of both quantities with k.

D.4 The Positive Bias and a Two-Form Diagnostic

The predictor over-predicts held-out reliability by about $+0.05$. The offset is essentially independent of budget (residual versus k, $r = 0.00$) and of discrimination (residual versus D, $r = 0.05$, Figure 8b), but is moderately larger for smaller benchmarks (residual versus $\log N$, $r = -0.36$). This is the direction expected if the offset arises from the held-out reference itself being a finite, noisy ranking. Under the measurement model the expected agreement between two disjoint forms of sizes k and $N - k$ is $\rho(k \mid D)\,\rho(N - k \mid D)$. Using this two-form prediction reduces the mean bias from $+0.048$ to $+0.025$ and the MAE from 0.060 to 0.050, while leaving the correlation unchanged ($r = 0.964$). Roughly half the offset is thus explained by the finite held-out reference. We report the single-form predictor as the primary predictor because it requires only D and k, and treat the two-form product as a diagnostic that identifies the mechanism of the residual bias rather than a fitted correction. Residual bias also varies by family, slightly negative on MMLU-Pro and GPQA and largest on the MATH split, indicating further structure not captured by a single effective D.

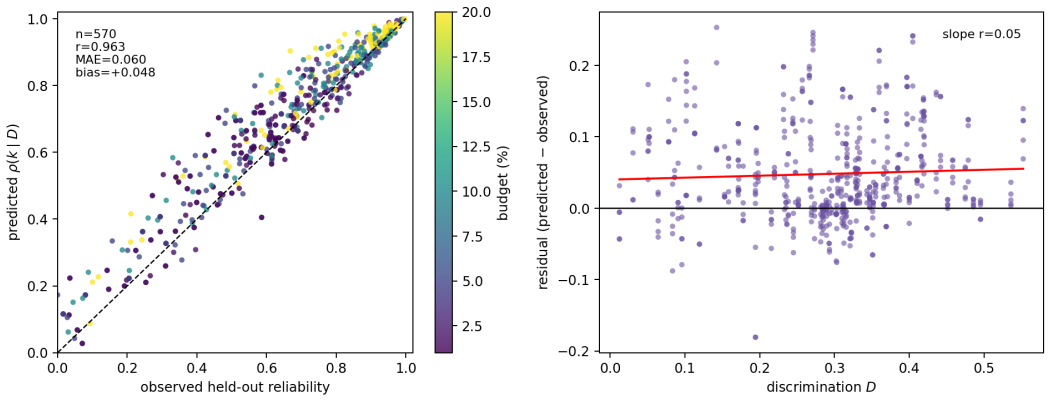


Figure 8: **Budget-sweep validation** ($n = 570$, budgets 1–20%). (a) Predicted $\rho(k \mid D)$ versus observed held-out reliability, colored by budget ($r = 0.963$, MAE $= 0.060$, bias $+0.048$). (b) Residual versus D. The offset is weakly related to D ($r = 0.05$).

Table 8: Accuracy, bias, and stability of $\widehat{D}$ versus source-pool size M (300 draws per M). SD/draw is the mean within-benchmark standard deviation across draws, and downstream r is the correlation between $\rho(k \mid \widehat{D})$ and observed held-out reliability at $k = 10\%$.

M	$r(\widehat{D}, D_{\text{full}})$	MAE	Bias	SD/draw	Downstream r
2	0.885	0.065	+0.052	0.362	0.944
3	0.887	0.072	+0.069	0.182	0.946
5	0.908	0.055	+0.054	0.113	0.949
8	0.933	0.039	+0.038	0.080	0.949
10	0.948	0.030	+0.028	0.071	0.952
15	0.962	0.024	+0.023	0.056	0.953
20	0.971	0.018	+0.016	0.047	0.955
30	0.984	0.011	+0.010	0.036	0.958
50	0.991	0.007	+0.005	0.027	0.961

E Estimating Discrimination from Source Models

We quantify how well D can be estimated from a small source population before evaluating the target models. Estimation uses the item-rest discrimination on the sampled source models, and the full-population D is the reference. All results are on the 114 instances.

E.1 Effect of Source-Pool Size

Drawing M models uniformly at random (300 draws per M), Table 8 reports the cross-benchmark correlation between mean $\widehat{D}$ and D_{full}, the MAE and bias of $\widehat{D}$, its per-draw standard deviation, and the downstream correlation between $\rho(k \mid \widehat{D})$ and observed held-out reliability at $k = 10\%$. Accuracy improves monotonically ($r = 0.908$ at $M = 5$, 0.962 at $M = 15$, 0.971 at $M = 20$), and the downstream prediction saturates earlier, reaching $r = 0.952$ at $M = 10$ and 0.953 at $M = 15$, versus 0.969 using the full-population D.

Uncertainty in $\widehat{D}$ propagates to $k^\star$. Because $k^\star \propto (1 - D^2)/D^2 \approx D^{-2}$ for small D, relative error in $\widehat{D}$ is amplified in $k^\star$. A 20% relative error in $\widehat{D}$ produces roughly a 40% error in $k^\star$, most severely for low-D benchmarks. The strong downstream correlation of $\rho(k \mid \widehat{D})$ therefore does not by itself guarantee accurate budget selection, and the practical recommendation of $M \approx 10$–20 source models is a tested regime rather than a universal threshold.

Table 9: Source-selection strategy and estimated pool discrimination, pooled over $M \in \{10, 20, 50\}$. Median reference $D = 0.304$.

Strategy	Selection	Median D	Δ vs. ref.
extremes	weakest + strongest only	0.544	$+0.238$
spread	even coverage of ability range	0.384	$+0.079$
by_provider	stratified by model provider	0.334	$+0.035$
by_size	stratified by model size	0.331	$+0.034$
random	uniform	0.331	$+0.033$
cluster	clustered in ability	0.307	$+0.013$
diverse	diversity-maximizing	0.289	-0.013
top	strongest models only	0.114	-0.182
bottom	weakest models only	0.075	-0.222

E.2 Source-Model Selection Strategy

Because D is population-relative, the source-selection strategy changes the estimate even at fixed M. Across nine strategies (Table 9, pooled over $M \in \{10, 20, 50\}$), pools that span the ability range stay near the reference (0.30–0.38), while selecting only the extremes inflates D to 0.544 and single-tail pools deflate it to 0.114 (top) and 0.075 (bottom). A source pool should therefore cover the ability range relevant to the target.

E.3 Source D Predicts Target Reliability

For each instance we draw $M = 30$ source models and a disjoint random 25% of the remaining models as target (median 99 models). Target models are evaluated on a random 10% subset and ranked against their full-benchmark ranking (6 seeds). Source-estimated D predicts observed target reliability with $r = 0.93$ (MAE $= 0.058$), and the target population's own D gives $r = 0.92$ (MAE $= 0.058$). The two estimates correlate at $r(\widehat{D}_S, \widehat{D}_T) = 0.985$. In a complementary same-population test, estimating $\widehat{D}$ from M randomly drawn models and predicting the full-population held-out reliability at 10% gives $r = 0.953$ at $M = 15$ (0.952 at $M = 10$, 0.958 at $M = 30$, 0.969 using the full reference). These results concern matched populations. Population shift is examined in Appendix F.

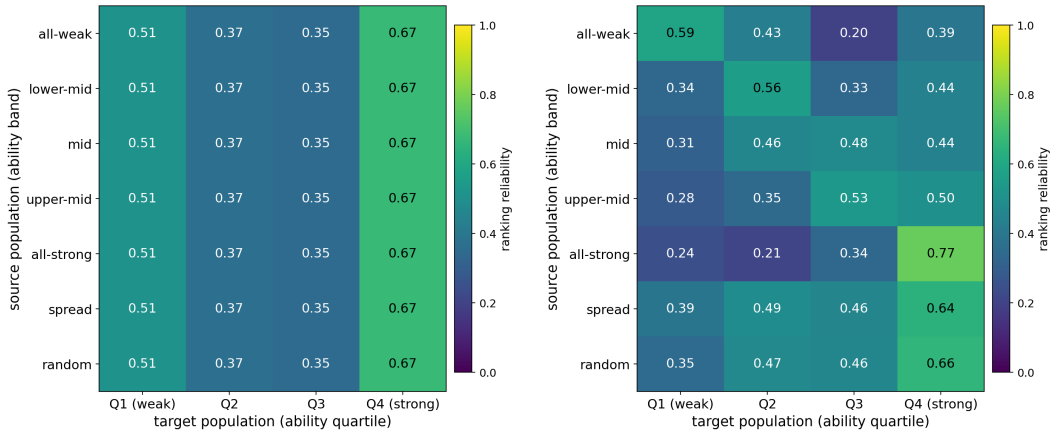
F Population Shift Across the Spectrum

The main-paper band comparison (Figure 4) samples an axis of source-target mismatch. Here we vary the source and target populations independently across the ability range, at $k = 10\%$, median over the 114 instances and seeds, with $M = 30$ source models.

Grid and mismatch metric. The target is one of four ability quartiles ($\text{Q1}_{\text{weak}} \rightarrow \text{Q4}_{\text{strong}}$) and the source is one of seven ability bands (Appendix C.4). Each cell has a scalar mismatch $\Delta a = \bar{a}_{\text{tgt}} - \bar{a}_{\text{src}}$, ranging from -0.29 (strong source, weak target) to $+0.29$ (weak source, strong target).

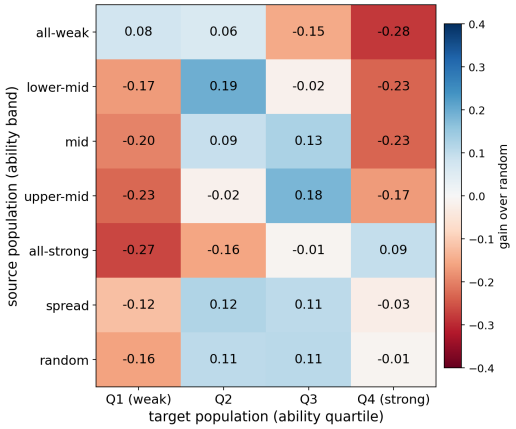
Matched-population selection gain. Under matched populations ($M = 30$, disjoint 25% target, 10% budget), mRMR++ gives the highest mean reliability, $+0.033$ over random, agreeing with random within 0.05 on 68% of instances. The other methods lie between random and mRMR++. Selection thus helps modestly under matched populations but does not establish a uniform advantage.

Cumulative bands. Table 10 reports selection gain over random as the target narrows. mRMR++ and metabench degrade sharply, while AIPW, PPI++, and Anchor Points stay within about ± 0.06 of random. The behavior changes across the population relationship rather than switching at a binary interpolation/extrapolation boundary.



(a) Random reliability

(b) mRMR++ reliability



(c) mRMR++ gain over random

Figure 9: **Reliability and selection gain across the source-target grid.** Rows are source ability bands, columns target ability quartiles. (a) Random reliability varies across target quartiles but is constant across source bands. (b) mRMR++ reliability varies with both. (c) mRMR++ gain over random is positive when source and target are aligned and negative under larger mismatch, down to -0.28.

Table 10: **Selection gain over random across the mismatch spectrum** (method minus random reliability at $k = 10\%$, medians). Target bands use the complement source (Appendix C.3).

Method	Mid-half	Top-half	Top-quartile	Top-decile
mRMR++	-0.024	-0.197	-0.301	-0.192
metabench	$+0.152$	-0.368	-0.424	-0.223
TinyBenchmarks	$+0.045$	-0.027	-0.045	-0.027
Anchor Points	$+0.056$	-0.000	$+0.014$	$+0.058$
AIPW	$+0.025$	-0.008	-0.014	-0.010
PPI++	$+0.028$	-0.004	-0.008	-0.006

Interpretation. These results separate ranking reliability (target-dependent and source-independent under random sampling), selection gain (the change from a source-based method), and source-target mismatch (Δa). Source-based selection inherits information about the population on which its item statistics are estimated, so its gains need not transfer under shift. We do not claim a causal mechanism beyond the experimental comparison. The interpolation and extrapolation regimes of prior work are points along this broader spectrum rather than a binary switch.