MetaEval: Measuring the Discrimination of Benchmarks for Efficient LLM Evaluation

Zhuo Wang[1, 2, 3], **Wen Wu**[3*], **Guoqing Wang**[3], **Guangze Ye**[1, 2, 3], **Zhenxiao Cheng**[3]

[1]Lab of Artificial Intelligence for Education, East China Normal University, Shanghai, China
[2]Shanghai Institute of Artificial Intelligence for Education, East China Normal University, Shanghai, China
[3]School of Computer Science and Technology, East China Normal University, Shanghai, China
{zhuowang, wgq, gzye, 51255901012}@stu.ecnu.edu.cn, wwu@cc.ecnu.edu.cn

Abstract

Benchmarks serve as standardized test systems to distinguish capabilities among large language models (LLMs). Discriminative items enable high-ability LLMs to favor correct answers, while causing low-ability models to assign lower plausibility to these answers and tend toward incorrect answers. Current methods for assessing benchmark quality primarily focus on coverage of difficulty levels and task diversity, yet lack direct quantification of discrimination—the core metric. Furthermore, large-scale benchmarks incur high evaluation costs. Although heuristic methods can reduce item counts to some extent, they cannot guarantee preservation of the benchmark's original discriminative properties. To address these limitations, we propose MetaEval, a meta-evaluation framework designed to precisely quantify per-item discrimination and enable efficient assessment. Central to MetaEval is our novel Signal Detection and Item Response (SD-IR) model, which simulates LLMs' detection of correct answers (signals) by representing each model's perception through two latent ability states: "known" and "unknown". For any item, discrimination is quantified as the difference in signal plausibility between these states. Leveraging these discrimination metrics, MetaEval introduces two strategies to replicate full-benchmark results using minimal subsets for efficient evaluation: (1) Distilling *metaBench*: a compact subset that retains discriminative power by removing redundant items; (2) Predicting performance on full-benchmark based on *metaBench*'s discrimination. Experiments across five benchmarks confirm that high-discrimination items capture greater performance variation among LLMs, align more closely with full-benchmark rankings, and exhibit superior predictive ability. Notably, in the best case, MetaEval achieves accurate full-benchmark estimation using only 2.5% of items, substantially reducing evaluation costs while preserving reliability.

Code — https://github.com/wangzhuo0092/MetaEval

1 Introduction

As large language models (LLMs) continue to advance rapidly, reliable and efficient evaluation is essential for guiding improvement (Wang et al. 2025c) and enabling comparison (Ye et al. 2024). Although recent work has proposed new benchmarks and leaderboard curation methods, such improvements typically focus on expanding difficulty coverage or diversifying tasks (Xuan et al. 2025; Liu et al. 2025; Ni et al. 2024). However, the core function of any benchmark remains underexamined: discriminative ability, or discrimination, i.e., the capacity to effectively distinguish LLMs of differing capabilities. Discriminative items guide high-ability LLMs to prefer correct answers, while low-ability ones tend to favor incorrect answers and assign lower plausibility to the correct answer, as illustrated in the upper part of Figure 1. By contrast, items with low discrimination fail to reflect true capability gaps. In such cases, strong and weak models may perform similarly, or worse, weaker models may outperform due to artifacts or shortcuts (Balepur, Ravichander, and Rudinger 2024; Li, Lan, and Yang 2025). As a result, rankings become highly sensitive to sampling and unstable: small changes in subset composition can cause significant shifts, particularly when model abilities appear close due to evaluation lacking sufficient discrimination (Alzahrani et al. 2024; Gao et al. 2025; Zhou et al. 2025). This undermines the credibility of evaluations and helps explain why benchmark results often fail to reflect real-world utility—they do not reliably reveal genuine LLM capability (Banerjee, Agarwal, and Singh 2024).

In addition, benchmarks often consist of hundreds or even

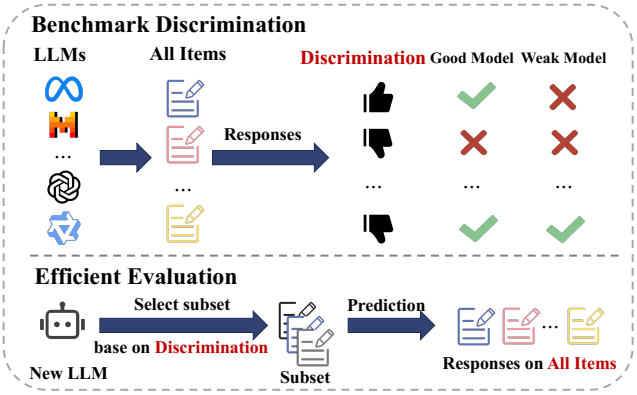

Figure 1: Illustration of Discrimination Quantification and Efficient Evaluation.

*The corresponding author.

thousands of items, making LLM evaluation both computationally and financially expensive, especially when repeated across training checkpoints (Li et al. 2025b; Wei et al. 2025). For instance, MMLU-ProX evaluation reportedly consumed 4k GPU hours, and HELM can exceed $10K per model via APIs (Xuan et al. 2025; Liang et al. 2023). To reduce cost, existing approaches commonly remove items with superficial task overlap or apply heuristic-based filters to discard seemingly uninformative examples (Perlitz et al. 2024). However, items that appear redundant in content may still elicit different responses from LLMs of varying capabilities, and thus offer unique discriminative value (Feng et al. 2024; Zhang et al. 2025). Pruning solely based on content risks discarding high-discrimination items while retaining less informative ones, thereby compromising the benchmark's original discriminative capacity and yielding evaluations that are cheaper but less reliable and potentially misleading.

To address these limitations, we propose MetaEval, a novel meta-evaluation framework that treats evaluation itself as the object of analysis (Murugadoss et al. 2025; Wang et al. 2025a). MetaEval focuses on quantifying the discrimination of benchmark items and leveraging it to enable reliable and efficient evaluation. Firstly, at the core of MetaEval is a novel SD-IR model for quantifying item discrimination, inspired by signal detection theory (SDT) (Green, Swets et al. 1966; Wang et al. 2025b) and item response theory (IRT) (Lord and Novick 2008). Specifically, the SD component models how LLMs assign plausibility to all candidate answers by simulating the detection of the correct answer (signal) among incorrect ones (noise). To capture how the plausibility distribution varies with LLM ability, the IR component simulates the interaction between each LLM's latent ability and item characteristics, thereby constraining each LLM's ability to two states: "known" and "unknown". The item's discrimination is then defined as the overall plausibility gap assigned to the signal between these two states.

Secondly, to enable efficient evaluation, we propose two strategies that preserve discriminative capacity while reducing item count. One strategy focuses on constructing discrimination vectors based on the SD-IR model, capturing each item's discriminative strength, distributional and interaction patterns. These vectors are then clustered, and representative items are selected from cluster centers to form a compact and diverse subset—*metaBench*. The other strategy further mitigates the loss of evaluation fidelity caused by removing items. We leverage *metaBench*'s discrimination to estimate LLM ability and predict full-benchmark performance, enabling reliable evaluation with only a small subset, as illustrated in the lower part of Figure 1.

To verify the effect of our method, we conduct experiments on five benchmarks. The results confirm that SD-IR effectively quantifies item discrimination by analyzing LLM performance, ranking consistency, and predictive ability across different levels of discrimination. Moreover, we demonstrate the practical value of MetaEval in enabling efficient evaluation. The two strategies require only a small subset of items and achieve average errors of 2.7% and 2.3% relative to full benchmark evaluations. In the best case, MetaEval estimates performance on MMLU-Pro using only 2.5% of the original data, with an average error just 1.2%.

To summarize, the contributions are listed as follows:

- We innovatively propose the SD-IR model, which quantifies the core discriminative ability of benchmarks by modeling plausibility distributions.
- We introduce MetaEval, a novel framework based on SD-IR that includes two efficient evaluation strategies: benchmark distillation and performance prediction.
- We empirically validate the effectiveness of MetaEval in discrimination quantification and efficient evaluation across five benchmarks.

2 Related work

2.1 Benchmark Evaluation

Scientific evaluation is essential for comparing LLMs, yet only 56.3% of benchmarks report quality measures (Zhao et al. 2024), raising concerns about their reliability. Recent studies reveal the sensitivity of LLM rankings to benchmark subset variations, suggesting that certain subsets are not robust in distinguishing model capabilities (Siska et al. 2024; Alzahrani et al. 2024). To examine benchmarks, Easy2Hard-Bench explores item difficulty from historical model performance (Ding et al. 2024), while others propose heuristic filtering using LLM-based annotation (Li et al. 2025a) or statistical properties like low variance and weak score correlation (Kipnis et al. 2025). However, existing efforts rely on heuristics or coarse criteria, overlooking both the benchmark's core discriminative capacity and how to quantify it.

2.2 Efficient Benchmarking

Recent work has sought to reduce evaluation costs through various strategies. Ye et al. (2023) explore task reduction in Big-Bench, while Zhuang et al. (2025) and Ding et al. (2025) propose adaptive testing that adjusts item difficulty to match LLM capabilities. Others compress benchmarks by sampling diverse subsets (Perlitz et al. 2024), clustering based on model confidence (Vivek et al. 2024), or selecting informative items to prioritize high-impact questions (Kipnis et al. 2025). Despite their efficiency, these methods rely on surface-level signals and risk weakening the benchmark's original discriminative capacity by discarding items that may be valuable for differentiating model performance.

In this work, we propose a unified framework to quantify item discrimination and leverage it for efficient evaluation.

3 Preliminaries

Let $\mathcal{L} = \{l_1, l_2, \ldots, l_N\}$, $\mathcal{J} = \{j_1, j_2, \ldots, j_M\}$ denote the sets of N LLMs and M benchmark items, respectively. Each item $j \in \mathcal{J}$ is associated with K_j candidate answers. We define the LLM response matrix $\mathbf{R} \in \mathbb{N}^{N \times M}$, where $R_{lj} \in \{1, \ldots, K_j\}$ denotes the answer selected by model l on item j. Based on this, we define the correctness matrix $\mathbf{Y} \in \{0, 1\}^{N \times M}$, where $Y_{lj} = 1$ if the selected answer R_{lj} is correct, and $Y_{lj} = 0$ otherwise. Our objective is to assess the discrimination of benchmarks and examine whether the discrimination of a small subset of the benchmark $\hat{\mathcal{J}} \subset \mathcal{J}$ can serve as a proxy for predicting LLMs' performance on

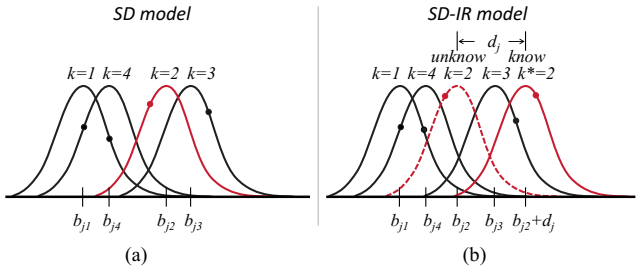

Figure 2: Illustration of (a) SD model and (b) SD-IR model.

the full benchmark $\mathcal{J}$. To this end, we design a task-adapted SD model and extend it into a mixed SD-IR model.

Definition 1. Signal Detection (SD) model. We first design an SD model, inspired by Signal Detection Theory (SDT), to model the plausibility distributions perceived by LLMs (Wickens 2001). As illustrated in Figure 2(a), the SD model assigns a Gumbel distribution to each candidate answer (e.g., $k = 1, 2, 3, 4$) for a given item j, where the correct answer (i.e., signal) is $k = 2$. It estimates plausibility scores $\boldsymbol{b}_j \in \mathbb{R}^{K_j}$, with each b_{jk} indicating the LLM's preference for answer k. The LLM will select the option with the highest b_{jk}, e.g., $k = 3$, even if it is incorrect.

Definition 2. SD-Item Response (SD-IR) model. To capture the impact of varying LLM abilities, we propose a novel extension to the SD model by incorporating an Item Response (IR) component, inspired by Item Response Theory (IRT) (Lord and Novick 2008; Brzezińska 2020). The improved IR component models the latent ability of LLM l, denoted by $\boldsymbol{\theta}_l$, and the skill requirement of item j, denoted by $\boldsymbol{\alpha}_j$. These jointly determine a binary variable $\delta_{lj} \in \{0, 1\}$, where $\delta_{lj} = 1$ indicates that LLM l possesses sufficient ability to solve item j. As shown in Figure 2(b), the signal plausibility distribution is split into "known" ($\delta_{lj} = 1$) and "unknown" ($\delta_{lj} = 0$) states. Their separation defines the item's discrimination $d_j \in \mathbb{R}$, reflecting how well the item distinguishes between LLMs of different abilities.

4 Proposed Framework

The overall framework of MetaEval is illustrated in Figure 3. It consists of two main steps: (1) Quantifying Item Discrimination using the proposed SD-IR model; and (2) Leveraging Discrimination for Efficient Evaluation, which includes two strategies: Strategy 1 — Benchmark Distillation to construct a compact *metaBench* that serves as a substitute for the full benchmark; and Strategy 2 — a combination of Ability Estimation and Performance Inference to approximate LLM performance on the full benchmark.

4.1 Step1: Item Discrimination Quantification

SD—Modeling the Plausibility Distribution over Candidates. For each benchmark item j, we model this decision process by associating each candidate answer $k \in \{1, \ldots, K_j\}$ with a continuous latent plausibility variable Ψ_{ljk} (DeCarlo 2020), representing the internal plausibility score that LLM l assigns to option k on item j:

$$R_{lj} = \arg\max_k \Psi_{ljk}, \quad \Psi_{ljk} \triangleq b_{jk} + \varepsilon_{ljk}, \tag{1}$$

where ε_{ljk} is a noise term with fixed zero-mean and variance to capture uncertainty in LLM preferences. Based on Equation (1), the probability of selecting candidate k is given by:

$$P(R_{lj} = k \mid \boldsymbol{b}_j) = \prod_{k' \neq k} p(\varepsilon_{ljk'} < b_{jk} - b_{jk'} + \varepsilon_{ljk}), \tag{2}$$

where $k' \neq k$ denotes all other candidate answers for item j. To simplify the inference, we marginalize over the noise term, assuming independence , derive the unconditional probability k:

$$P(R_{lj} = k \mid \boldsymbol{b}_j) = \mathbb{E}_{\varepsilon_{ljk}} \left[\prod_{k' \neq k} F(b_{jk} - b_{jk'} + \varepsilon_{ljk}) \right], \tag{3}$$

where $F(\cdot)$ denotes the cumulative distribution function (CDF) of the noise variable ε_{ljk}.

SD-IR — Modeling Plausibility Gaps for Discrimination Estimation. To account for varying LLM capabilities, we incorporate δ_{lj}, which indicates whether LLM l can solve item j. Firstly, the terms b_{jk} and ε_{ljk} follow the same definitions as in the SD model and are independent of correctness. Each candidate answer k is also associated with a binary signal variable X_{jk}, where $X_{jk} = 1$ represents the presence of a signal (i.e., the correct answer) and 0 represents a distractor. This signal is only utilized when the LLM is in a "known" state. To model how well an item distinguishes different LLMs, we incorporate its discrimination parameter d_j. A higher d_j means the item provides a stronger signal for capable LLMs. The plausibility function incorporates an interaction term $d_j \delta_{lj} X_{jk}$, resulting in the structural plausibility variable Ψ_{ljk}:

$$\Psi_{ljk} \triangleq b_{jk} + d_j \delta_{lj} X_{jk} + \varepsilon_{ljk}. \tag{4}$$

In both versions of the model, the decision rule is to select the answer with the highest plausibility. The probability marginalizes over δ_{lj}:

$$P(R_{lj} = k) = \sum_{\delta_{lj}=0}^{1} p(\delta_{lj})\, p(R_{lj} = k \mid \delta_{lj}). \tag{5}$$

The conditional probability $P(R_{lj} = k \mid \delta_{lj})$ follows the same form in the extended formulation Equation (2) and (3):

$$P(R_{lj} = k \mid \delta_{lj}) = \mathbb{E}_{\varepsilon_{ljk}} \left[\prod_{k' \neq k} F\left(\Delta\eta_{ljkk'} + \varepsilon_{ljk}\right) \right], \tag{6}$$

$$\Delta\eta_{ljkk'} = (b_{jk} + d_j \delta_{lj} X_{jk}) - (b_{jk'} + d_j \delta_{lj} X_{jk'}). \tag{7}$$

Under the assumption that $F(\cdot)$ follows the standard Gumbel distribution (Gumbel 1958), Equation (6) has a closed-form solution :

$$P(R_{lj} = k \mid \delta_{lj}) = \frac{e^{b_{jk} + \delta_{lj} d_j X_{jk}}}{\sum_{k'=1}^{K} e^{b_{jk'} + \delta_{lj} d_j X_{jk'}}}. \tag{8}$$

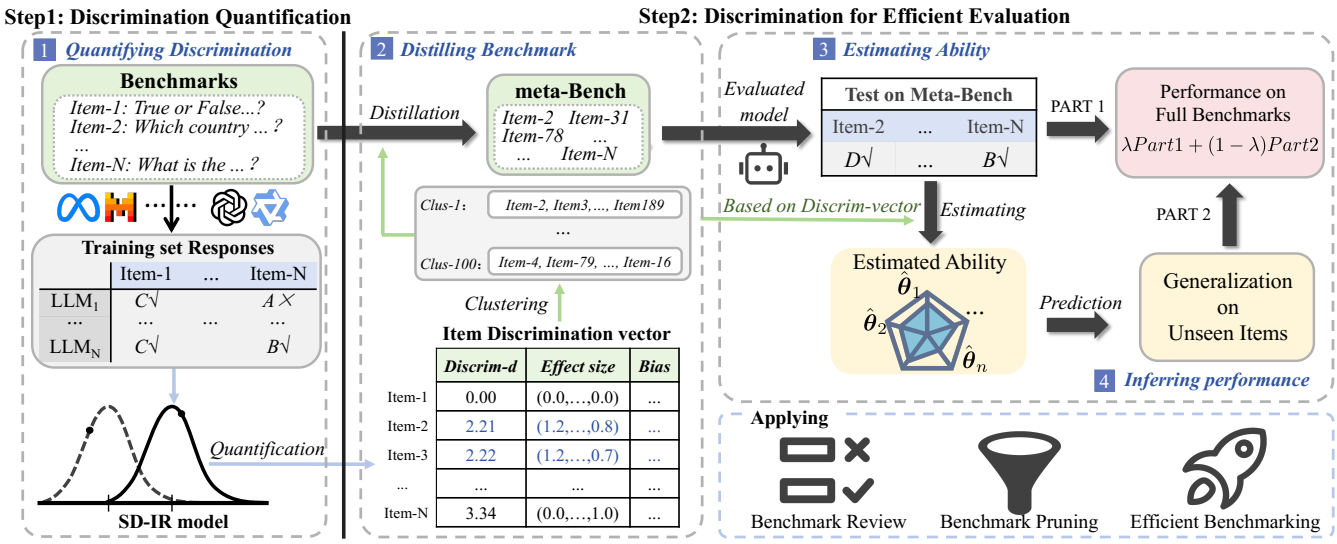

Figure 3: Overview of the proposed framework MetaEval for quantifying discrimination and efficient evaluation.

To capture the variation of knowledge across items and the ability differences across LLMs, we adopt a customized multidimensional IR component. The probability that LLM l "knows" item j is defined as:

$$p(\delta_{lj} = 1) \triangleq \frac{1}{1 + \exp(-\boldsymbol{\alpha}_j^\top \boldsymbol{\theta}_l)}. \tag{9}$$

Note that $\delta_{lj} = 0$ satisfies $p(\delta_{lj} = 0) = 1 - p(\delta_{lj} = 1)$, where $\boldsymbol{\theta}_l \in \mathbb{R}^{dim}$ represents the latent ability vector, and $\boldsymbol{\alpha}_j \in \mathbb{R}^{dim}$ the effect sizes of the interactions between item j and the corresponding ability dimensions. The inner product $\boldsymbol{\alpha}_j^\top \boldsymbol{\theta}_l$ reflects how well the abilities of LLM l align with the requirements of item j. The variable dim represents the number of distinct abilities, allowing the model to account for diverse ability requirements across items.

Fitting the SD-IR model. To estimate item-level parameters, we randomly select a training subset of LLMs $\mathcal{L}_{\text{tr}} \subset \mathcal{L}$, and choose the optimal dimension from $dim \in \{1, 3, 5, 7, 10\}$. Finally, we formalize the quantification process as the following mathematical optimization problem. Specifically, we maximize the log-likelihood of the observed responses, $\{R_{lj} = r_{lj} \mid l \in \mathcal{L}_{\text{tr}},\ j \in \mathcal{J}\}$:

$$\Theta^* = \arg\max_{\Theta} \sum_{l \in \mathcal{L}_{\text{tr}}} \sum_{j \in \mathcal{J}} \log P(R_{lj} = r_{lj}), \tag{10}$$

$$\Theta = \{\mathbf{b}_j, d_j, \boldsymbol{\alpha}_j\}_{j \in \mathcal{J}} \cup \{\boldsymbol{\theta}_l\}_{l \in \mathcal{L}_{\text{tr}}}. \tag{11}$$

The discrimination d_j quantifies the plausibility gap between "known" and "unknown" states, reflecting how well an item separates strong from weak models.

4.2 Step2: Discrimination for Efficient Evaluation

We aim to examine whether item discrimination can support efficient evaluation of LLMs $l \in \mathcal{L} \setminus \mathcal{L}_{\text{tr}}$ by selecting a small subset of items to reproduce their performance on the full benchmark. Formally, the target accuracy is defined as:

$$Z_l \triangleq \frac{1}{|\mathcal{J}|} \sum_{j \in \mathcal{J}} Y_{lj} \tag{12}$$

Benchmark Distillation. First, we select a subset of items $\hat{\mathcal{J}} \subset \mathcal{J}$, referred to as the *metaBench*—a compact and representative subset distilled from the full benchmark. To preserve the original discriminative capacity of the benchmark, we construct *metaBench* by leveraging item parameters from our SD-IR model. For each item $j \in \mathcal{J}$, we define a discrimination vector $\mathbf{D}_j = [d_j, \mathbf{b}_j, \boldsymbol{\alpha}_j]$, which comprehensively represents the item's discriminative characteristics. Specifically, d_j indicates the item's overall discriminative strength, $\mathbf{b}_j$ encodes the relative bias of candidate answers with respect to the discrimination signal, and $\boldsymbol{\alpha}_j$ captures the item's multidimensional skill requirements.

We then perform K-Means clustering over all $\mathbf{D}_j$, grouping items with similar discriminative characteristics. Finally, one representative item is selected from each cluster to form the *metaBench*. The resulting *metaBench* significantly reduces evaluation cost. Importantly, $\mathbf{D}_j$ is low-dimensional and interpretable, making the selection process principled.

Based on the LLM's responses over the *metaBench* $\hat{\mathcal{J}}$, we compute an observed score $Z_l^{\hat{\mathcal{J}}}$ as an estimate of overall performance, as **Strategy 1**:

$$Z_l^{\hat{\mathcal{J}}} \triangleq \frac{1}{|\hat{\mathcal{J}}|} \sum_{j \in \hat{\mathcal{J}}} Y_{lj}. \tag{13}$$

Ability Estimation. To estimate the accuracy on unobserved items, we approximate its conditional expectation given the LLM's responses on the *metaBench*:

$$\hat{Z}_l^{pred} \triangleq \hat{\mathbb{E}}\left[Z_l \,\middle|\, \{R_{lj}\}_{j \in \hat{\mathcal{J}}}\right], \tag{14}$$

where $\hat{Z}_l^{pred}$ denotes the predicted accuracy on the original benchmark. To obtain this estimate, we infer the LLM's latent ability $\hat{\boldsymbol{\theta}}_l$ by maximizing the log-likelihood of its responses on $\hat{\mathcal{J}}$, while keeping $\mathbf{D}_j$ fixed:

$$\hat{\boldsymbol{\theta}}_l = \arg\max_{\boldsymbol{\theta}_l} \sum_{j \in \hat{\mathcal{J}}} \log P(R_{lj} \mid \mathbf{D}_j), \tag{15}$$

Then, $\hat{\boldsymbol{\theta}}_l$ is used to estimate the accuracy $\hat{y}_{lj}$ on each unobserved item $j \in \mathcal{J} \setminus \hat{\mathcal{J}}$:

$$\hat{y}_{lj} \triangleq \mathbb{I}\left[\arg\max_k \mathbb{P}(R_{lj} = k \mid \hat{\boldsymbol{\theta}}_l, \mathbf{D}_j) = k^*\right] \tag{16}$$

$\hat{y}_{lj} = 1$ indicates that LLM l's most probable response is correct (i.e., k^*), and $\hat{y}_{lj} = 0$ otherwise. Subsequently, the predicted accuracy can be expressed as:

$$\hat{Z}_l^{pred} = \frac{1}{|\hat{\mathcal{J}}|} \sum_{j \in \hat{\mathcal{J}}} Y_{lj} + \frac{1}{|\mathcal{J} \setminus \hat{\mathcal{J}}|} \sum_{j \in \mathcal{J} \setminus \hat{\mathcal{J}}} \hat{y}_{lj} \tag{17}$$

LLM Performance Inference. The discrimination vector $\mathbf{D}_j$ enables the prediction of LLM performance on unobserved items. However, estimating $\hat{\boldsymbol{\theta}}_l$ using the discrimination capacity of only a small number of items may involve errors. These errors may affect the prediction of correctness on unobserved items $\hat{y}_{il}$.

To address these challenges, we combine the observed correctness on $\hat{\mathcal{J}}$ with predicted accuracy on unobserved items, balancing empirical evidence with discrimination-based inference (Polo et al. 2024), as **Strategy 2**:

$$\hat{Z}_l \triangleq \lambda Z_l^{\hat{\mathcal{J}}} + (1 - \lambda)\hat{Z}_l^{pred}, \tag{18}$$

where $\lambda \in [0, 1]$ controls the balance between the unbiased but high-variance empirical term and the low-variance but potentially biased prediction. Setting $\lambda = 1$ recovers **Strategy 1**, using only responses on *metaBench*. We compute λ based on a bias-variance heuristic (Song 1988):

$$\lambda = \frac{\hat{\beta}^2}{\hat{\beta}^2 + \hat{\sigma}^2/|\hat{\mathcal{J}}|}, \tag{19}$$

where $\hat{\sigma}^2$ is the empirical variance of Y_{lj} across training LLMs $l \in \mathcal{L}_{\text{tr}}$, and $\hat{\beta}^2$ quantifies the squared bias between the estimated probabilities and the observed correctness on *metaBench*. This formulation ensures that when the number of *metaBench* $|\hat{\mathcal{J}}|$ is small, more weight is placed on prediction, while as more observed responses become available, the estimator tends toward the empirical average.

5 Assessing MetaEval

We aim to answer the following research questions (RQs):

- **RQ1**: Can SD-IR effectively quantify item discrimination?
- **RQ2**: Can MetaEval precisely predict full-benchmark performance using only a small number of items?
- **RQ3**: What is the difference in predictive performance between high and low-discrimination items?
- **RQ4**: How does assigning higher weights to high-discrimination items affect LLM rankings?

5.1 Benchmarks

- ARC-Challenge (Clark et al. 2018) is a more difficult version of the ARC, containing 1,172 adversarial questions designed to defeat retrieval.
- MMLU-Pro (Wang et al. 2024) is an enhanced version of the MMLU, containing 12K items across 14 subjects, with 10 candidate answers per question instead of 4.
- Big-Bench-Hard (BBH) (Suzgun et al. 2023) is a subset of 23 reasoning tasks from the BIG-bench, containing approximately 5.7k items.
- GPQA (Rein et al. 2024) is a graduate-level multiple-choice QA benchmark, available in three nested sets: *Extended* (546 items), *Main* (448), and *Diamond* (198). We use the *Extended* version to ensure full coverage.
- MuSR (Sprague et al. 2024) is a multi-step reasoning benchmark consisting of 756 items.

5.2 Models

We collect the latest evaluation results of 213 LLMs from the Open LLM Leaderboard (Beeching et al. 2023), and split them into 80% training $\mathcal{L}_{\text{tr}}$ and 20% test sets.

5.3 Results and Analysis of RQ1

To validate the effectiveness of SD-IR in quantifying item discrimination, we analyze the consistency between discrimination levels and other metrics that reflect performance differences among models. Each benchmark is divided into five bins based on item discrimination, ranging from low to high. We consider Spearman Rank Correlation Coefficient (SRCC, ρ) (Ali Abd Al-Hameed 2022) to measure the correlation between LLM rankings within each discrimination bin and the entire benchmark ranking. A higher SRCC indicates that the items better preserve the overall performance

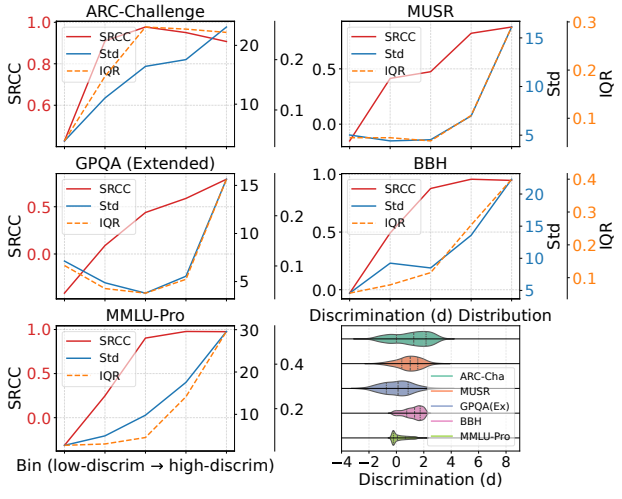

Figure 4: Consistency of Discrimination with Std, IQR, and SRCC (Ranking Correlation).

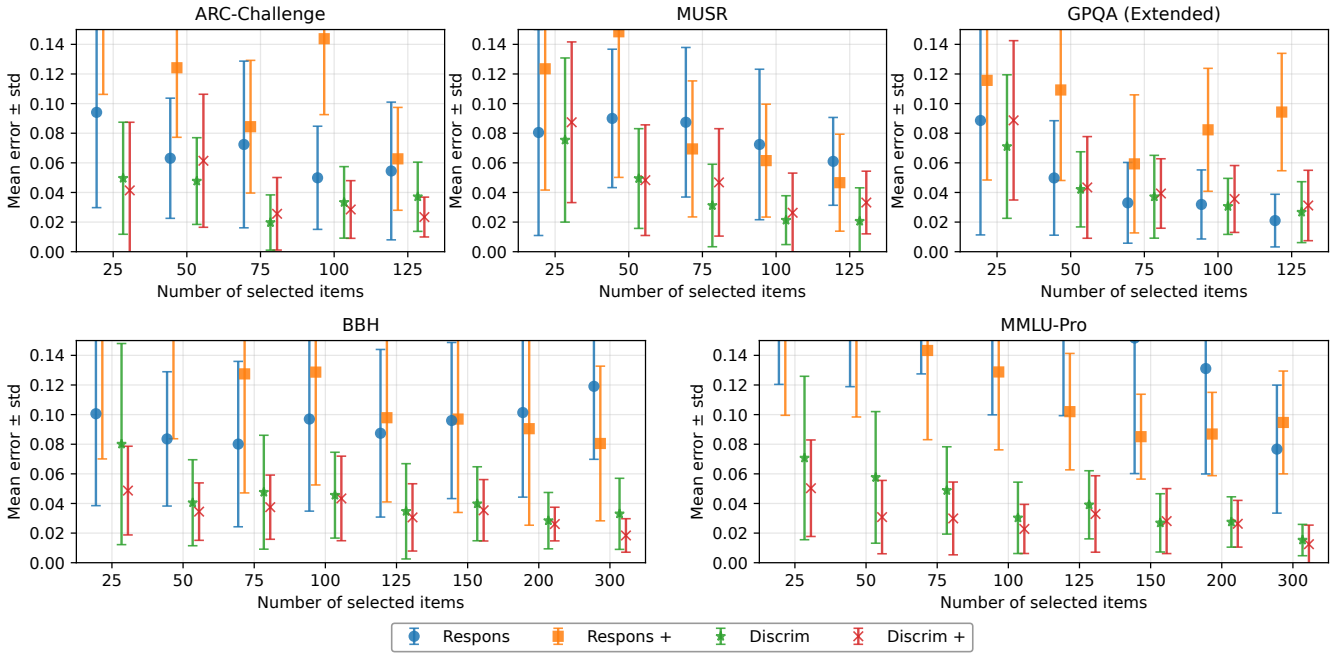

Figure 5: Performance prediction error (mean ± std) under different strategies across sampling sizes (rows) per benchmark.

differences. Within each bin, we consider Standard Deviation (Std) and Interquartile Range (IQR) (Arachchige, Prendergast, and Staudte 2022) to assess performance variation among models. Std measures the performance variation of all LLMs, while IQR measures the accuracy gap between the 75th and 25th percentile models, indicating the separation between stronger and weaker ones.

As shown in Figure 4, SRCC, Std, and IQR generally increase with higher levels of discrimination. The highest discrimination bin achieves an average Spearman's $\rho = 0.90$, with a 16.72% higher Std and a 29.51% higher IQR than the lowest bin . In addition, the violin plot in the bottom right shows the distribution of item discrimination across benchmarks. MUSR, GPQA, and MMLU-Pro show a high proportion of negatively discriminative items ($d < 0$), which aligns with their lower Std and IQR in the first three bins. Moreover, the lowest bin even shows a negative correlation with overall ranking ($\rho < 0$). In GPQA, however, the lowest bin exhibits unexpectedly high Std and IQR, likely due to randomness or artifacts where lower-ability models occasionally outperform stronger ones (Balepur, Ravichander, and Rudinger 2024). These findings confirm that SD-IR effectively quantifies discrimination, as high-discrimination items distinguish LLMs and align well with overall rankings, supporting its use as a core meta-evaluation metric for assessing benchmark's ability to differentiate LLMs.

5.4 Results and Analysis of RQ2

MetaEval aims to enable accurate yet low-cost evaluation via two strategies that use only a small subset of items. To validate their effectiveness, we compare the LLM performance estimated by these strategies with the actual performance on the full benchmark. *Discrim* and *Discrim+* refer to the benchmark distillation and performance inference strategies, respectively, as introduced in Section 4.2. As a baseline, we distill benchmark by clustering correctness patterns $\{Y_{lj}\}$ across training LLMs $l \in \mathcal{L}_{\text{tr}}$, denoted as *Respons*, and further inferring performance based on the subset, denoted as *Respons+*. The baseline approach follows a natural heuristic: selecting items where LLMs exhibit diverse correctness patterns, which may indicate informative items. The maximum sampling was set to 125 items for ARC-Challenge, MUSR, and GPQA, and increased to 300 for BBH and MMLU-Pro due to the substantially larger sizes.

Figure 5 reports prediction errors relative to the full benchmark. Each point marks the mean error across test LLMs, with vertical bars denoting standard deviation (mean ± std). Lower points imply higher accuracy; shorter bars, greater stability. The x-axis shows the number of selected items. The results demonstrate the effectiveness of the discrimination-based strategies *Discrim* and *Discrim+*. At the maximum sampling size across all benchmarks, *Discrim* and *Discrim+* achieve average errors of 2.7% (mean std: 2.1%) and 2.3% (mean std: 1.7%), respectively, with *Discrim+* demonstrating the highest robustness by consistently low error and variance. In contrast, *Respons* and *Respons+* exhibit higher errors and variability across nearly all cases.

Specifically, for ARC-Challenge, BBH, and MMLU-Pro, *Discrim+* achieves the lowest prediction errors at the largest sample size: 2.3% (std: 1.4%), 1.8% (std: 1.1%), and 1.2% (std: 1.3%), respectively. Additionally, for ARC-Challenge, *Discrim* also achieves a similarly low error of 2.0% (std: 1.9%) at 75 samples, but a relatively higher std. For MUSR, *Discrim* yields 2.1% at both 100 and 125 samples, with

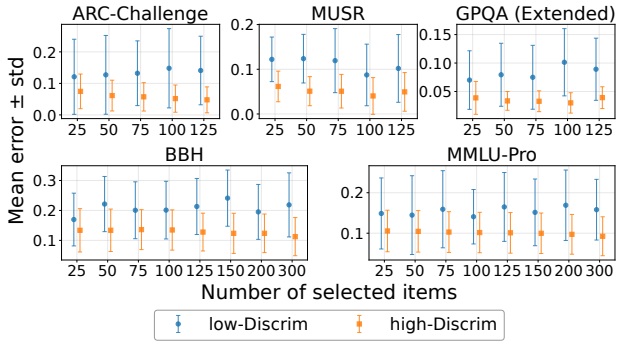

Figure 6: Prediction accuracy comparison between high and low discrimination items across sampling sizes (rows).

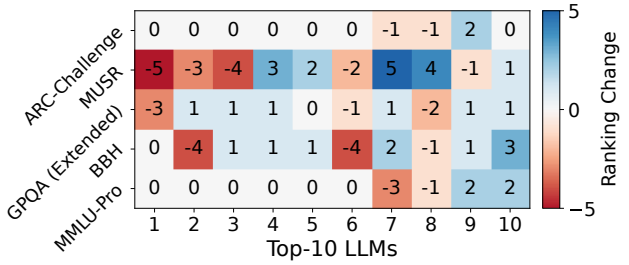

Figure 7: TOP10 LLM Ranking Change (Original - Weighted).

lower variability at 100 (std: 1.7% vs. 2.3%). GPQA is the only case where *Respons* slightly surpasses *Discrim* at 125 samples, yielding a lower error of 2.1% (std: 1.8%) compared to 2.7% (std: 2.1%). This may be because randomness in low-discrimination items weakens the effectiveness of discrimination-based strategies, while *Respons* occasionally benefits from capturing superficial patterns that correlate with model behavior. These results suggest that *Respons* and *Respons+*, which rely solely on surface-level correctness variation to identify informative items, fail to preserve the original discriminative structure. Clustering stability may further be undermined by the high-dimensional representations from numerous training LLMs. In contrast, *Discrim* and *Discrim+* maintain strong performance reconstruction while substantially reducing cost—for example, using only 2.5% of MMLU-Pro yields just 1.2% error. This demonstrates that our framework preserves discriminative ability using only a small subset, grounded in explicit modeling of item-level discrimination. Meanwhile, both strategies are based on discrimination, further highlighting its central role in effectively distinguishing LLM capabilities.

5.5 Results and Analysis of RQ3

Since predictive performance depends on how well items estimate LLM abilities, comparing predictive accuracy of items with different discrimination can also reveal the effectiveness of item discrimination in assessing LLM capabilities. Specifically, we follow the same sampling setup as RQ2 (along the x-axis) but focus solely on predictive ability by directly estimating performance on unseen items using two sets: one with the highest discrimination and another with near-zero discrimination. This yielding $\hat{Z}_l^{pred}$ (introduced in Section 4.2) without further inferring overall performance.

As shown in Figure 6, high-discrimination items yield significantly lower prediction errors and variances than low-discrimination ones across all cases ($p < 0.05$), with average error reductions of 7.53%, 6.02%, 4.81%, 7.94%, and 5.40% across benchmarks . Specifically, the performance differences between the two sets are relatively smaller in MUSR, GPQA, and MMLU-Pro, which corresponds to RQ1's finding that these benchmarks contain more negatively discriminative items. A possible reason is that responses to such items tend to be more random, making prediction more difficult. As the sampling size increases, prediction performance shows limited improvement—likely because, under the small sample setting, the 25 most discriminative items already provide sufficient signal for stable ability estimation. These results suggest that high-discrimination items provide stronger predictive capability. This capability arises from their alignment with the estimation of latent model abilities, indicating that high-discrimination items are more effective at distinguishing varying abilities of LLM.

5.6 Results and Analysis of RQ4

To better understand how item discrimination affects LLM rankings, we weight each item by its discrimination value d, assigning higher weights to more discriminative items . We then rank LLMs by their weighted accuracy and compare the resulting rankings with the original.

Figure 7 shows the impact of the weighting strategy on the top-10 model rankings. Each cell shows the change in the weighted ranking relative to the original, where 0 indicates no change. Ranking shifts appear across all benchmarks, especially in MUSR, GPQA, and BBH, where most models are affected; in MUSR, two models shift by up to 5 positions, highlighting the strong impact of incorporating item discrimination. Consistent with prior work, LLM rankings are sensitive to item distribution when benchmarks fail to reliably differentiate models (Siska et al. 2024; Alzahrani et al. 2024). This highlights the need to scrutinize items and suggests new directions for leaderboard curation that reflect LLM strengths on highly discriminative items.

6 Conclusion

In this paper, we propose an innovative meta-evaluation framework, MetaEval, which addresses the overlooked issue of benchmark discrimination and the high cost of evaluation. Through experimental analyses of LLM performance and item predictive capabilities across varying discrimination levels, we demonstrate MetaEval's effectiveness in quantifying discrimination. Moreover, leveraging discrimination enables benchmark distillation and performance inference for efficient evaluation, in the best case, just 1.2% error using only 2.5% of the original data. Thus, MetaEval not only provides a reliable mechanism for benchmark assessment but also supports rapid, cost-efficient evaluation.

Acknowledgments

This work is supported by the National Key Research and Development Program of China (under Grant No. 2024YFC3308500), the National Natural Science Foundation of China (under project No. 62377013), and the Fundamental Research Funds for the Central Universities. It is also supported by STI 2030-Major Projects 2021ZD0200500, the Research Project of Changning District Science and Technology Committee (under project No. CNKW2022Y37), and the Medical Master's and Doctoral Innovation Talent Base Project of Changning District (under project No. RCJD2022S07).

References

Ali Abd Al-Hameed, K. 2022. Spearman's correlation coefficient in statistical analysis. *International Journal of Nonlinear Analysis and Applications*, 13(1): 3249–3255.

Alzahrani, N.; Alyahya, H.; Alnumay, Y.; Alrashed, S.; Alsubaie, S.; Almushayqih, Y.; Mirza, F.; Alotaibi, N.; Al-Twairesh, N.; Alowisheq, A.; et al. 2024. When Benchmarks Are Targets: Revealing the Sensitivity of Large Language Model Leaderboards. In *Proceedings of the Association for Computational Linguistics (ACL)*, 13787–13805.

Arachchige, C. N.; Prendergast, L. A.; and Staudte, R. G. 2022. Robust analogs to the coefficient of variation. *Journal of Applied Statistics*, 49(2): 268–290.

Balepur, N.; Ravichander, A.; and Rudinger, R. 2024. Artifacts or Abduction: How Do LLMs Answer Multiple-Choice Questions Without the Question? In *Proceedings of the Association for Computational Linguistics (ACL)*, 10308–10330.

Banerjee, S.; Agarwal, A.; and Singh, E. 2024. The Vulnerability of Language Model Benchmarks: Do They Accurately Reflect True LLM Performance? *arXiv preprint arXiv:2412.03597*.

Beeching, E.; Fourrier, C.; Habib, N.; Han, S.; Lambert, N.; Rajani, N.; Sanseviero, O.; Tunstall, L.; and Wolf, T. 2023. Open LLM Leaderboard. https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard.

Brzezińska, J. 2020. Item response theory models in the measurement theory. *Communications in Statistics-Simulation and Computation*, 49(12): 3299–3313.

Clark, P.; Cowhey, I.; Etzioni, O.; Khot, T.; Sabharwal, A.; Schoenick, C.; and Tafjord, O. 2018. Think you have solved question answering? Try ARC, the AI2 Reasoning Challenge. In *Proceedings on Empirical Methods in Natural Language Processing (EMNLP)*, 4578–4587. Association for Computational Linguistics.

DeCarlo, L. T. 2020. An item response model for true–false exams based on signal detection theory. *Applied Psychological Measurement*, 44(3): 234–248.

Ding, M.; Deng, C.; Choo, J.; Wu, Z.; Agrawal, A.; Schwarzschild, A.; Zhou, T.; Goldstein, T.; Langford, J.; Anandkumar, A.; et al. 2024. Easy2Hard-Bench: Standardized Difficulty Labels for Profiling LLM Performance and Generalization. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, 44323–44365.

Ding, X.; Pan, C.; Li, Z.; Zhang, J.; Wang, S.; and Wei, Z. 2025. AutoJudger: An Agent-Driven Framework for Efficient Benchmarking of MLLMs. *arXiv preprint arXiv:2505.21389*.

Feng, K.; Ding, K.; Tan, H.; Ma, K.; Wang, Z.; Guo, S.; Cheng, Y.; Sun, G.; Zheng, G.; Zhang, Q.; et al. 2024. Sample-efficient human evaluation of large language models via maximum discrepancy competition. In *Proceedings of the Association for Computational Linguistics (ACL)*.

Gao, M.; Liu, Y.; Hu, X.; Wan, X.; Bragg, J.; and Cohan, A. 2025. Re-evaluating Automatic LLM System Ranking for Alignment with Human Preference. In *Findings of the Association for Computational Linguistics: NAACL*, 4605–4629.

Green, D. M.; Swets, J. A.; et al. 1966. *Signal detection theory and psychophysics*, volume 1. Wiley New York.

Gumbel, E. J. 1958. *Statistics of extremes*. Columbia university press.

Kipnis, A.; Voudouris, K.; Buschoff, L. M. S.; and Schulz, E. 2025. metabench-A Sparse Benchmark of Reasoning and Knowledge in Large Language Models. In *International Conference on Learning Representations (ICLR)*.

Li, T.; Chiang, W.; Frick, E.; Dunlap, L.; Wu, T.; Zhu, B.; Gonzalez, J. E.; and Stoica, I. 2025a. From Crowdsourced Data to High-Quality Benchmarks: Arena-Hard and Bench-Builder Pipeline. In *International Conference on Machine Learning (ICML)*.

Li, X.; Lan, Y.; and Yang, C. 2025. Treeeval: Benchmark-free evaluation of large language models through tree planning. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 39, 24485–24493.

Li, Y.; Ma, J.; Ballesteros, M.; Benajiba, Y.; and Horwood, G. 2025b. Active Evaluation Acquisition for Efficient LLM Benchmarking. In *International Conference on Machine Learning (ICML)*.

Liang, P.; Bommasani, R.; Lee, T.; Tsipras, D.; Soylu, D.; Yasunaga, M.; Zhang, Y.; Narayanan, D.; Wu, Y.; Kumar, A.; et al. 2023. Holistic Evaluation of Language Models. *Transactions on Machine Learning Research*.

Liu, C.; Jin, R.; Yao, Z.; Li, T.; Cheng, L.; Steedman, M.; and Xiong, D. 2025. Empirical Study on Data Attributes Insufficiency of Evaluation Benchmarks for LLMs. In *International Conference on Computational Linguistics (COLING)*, 6024–6038.

Lord, F. M.; and Novick, M. R. 2008. *Statistical theories of mental test scores*. IAP.

Murugadoss, B.; Poelitz, C.; Drosos, I.; Le, V.; McKenna, N.; Negreanu, C. S.; Parnin, C.; and Sarkar, A. 2025. Evaluating the evaluator: Measuring llms' adherence to task evaluation instructions. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 39, 19589–19597.

Ni, J.; Xue, F.; Yue, X.; Deng, Y.; Shah, M.; Jain, K.; Neubig, G.; and You, Y. 2024. MixEval: Deriving Wisdom of the Crowd from LLM Benchmark Mixtures. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Perlitz, Y.; Bandel, E.; Gera, A.; Arviv, O.; Dor, L. E.; Shnarch, E.; Slonim, N.; Shmueli-Scheuer, M.; and Choshen, L. 2024. Efficient Benchmarking of Language Models. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2519–2536.

Polo, F. M.; Weber, L.; Choshen, L.; Sun, Y.; Xu, G.; and Yurochkin, M. 2024. tinyBenchmarks: evaluating LLMs with fewer examples. In *International Conference on Machine Learning (ICML)*.

Rein, D.; Hou, B. L.; Stickland, A. C.; Petty, J.; Pang, R. Y.; Dirani, J.; Michael, J.; and Bowman, S. R. 2024. Gpqa: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*.

Siska, C.; Marazopoulou, K.; Ailem, M.; and Bono, J. 2024. Examining the robustness of LLM evaluation to the distributional assumptions of benchmarks. In *Proceedings of the Association for Computational Linguistics (ACL)*, 10406–10421.

Song, W. T. 1988. Minimal-MSE linear combinations of variance estimators of the sample mean. In *1988 Winter Simulation Conference Proceedings*, 414–421. IEEE.

Sprague, Z.; Ye, X.; Bostrom, K.; Chaudhuri, S.; and Durrett, G. 2024. MuSR: Testing the Limits of Chain-of-thought with Multistep Soft Reasoning. In *International Conference on Learning Representations (ICLR)*.

Suzgun, M.; Scales, N.; Schärli, N.; Gehrmann, S.; Tay, Y.; Chung, H. W.; Chowdhery, A.; Le, Q.; Chi, E.; Zhou, D.; et al. 2023. Challenging BIG-Bench Tasks and Whether Chain-of-Thought Can Solve Them. In *Findings of the Association for Computational Linguistics: ACL*, 13003–13051.

Vivek, R.; Ethayarajh, K.; Yang, D.; and Kiela, D. 2024. Anchor Points: Benchmarking Models with Much Fewer Examples. In *Proceedings of the European Chapter of the Association for Computational Linguistics*, 1576–1601.

Wang, G.; Dai, S.; Ye, G.; Gan, Z.; Yao, W.; Deng, Y.; Wu, X.; and Ying, Z. 2025a. Information Gain-based Policy Optimization: A Simple and Effective Approach for Multi-Turn LLM Agents. *arXiv preprint arXiv:2510.14967*.

Wang, G.; Wu, W.; Ye, G.; Cheng, Z.; Chen, X.; and Zheng, H. 2025b. Decoupling Metacognition from Cognition: A Framework for Quantifying Metacognitive Ability in LLMs. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 39, 25353–25361.

Wang, S.; Long, Z.; Fan, Z.; Huang, X.; and Wei, Z. 2025c. Benchmark Self-Evolving: A Multi-Agent Framework for Dynamic LLM Evaluation. In *International Conference on Computational Linguistics (COLING)*, 3310–3328.

Wang, Y.; Ma, X.; Zhang, G.; Ni, Y.; Chandra, A.; Guo, S.; Ren, W.; Arulraj, A.; He, X.; Jiang, Z.; et al. 2024. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*.

Wei, T.; Wen, W.; Qiao, R.; Sun, X.; and Ma, J. 2025. RocketEval: Efficient automated LLM evaluation via grading checklist. In *International Conference on Learning Representations (ICLR)*.

Wickens, T. D. 2001. *Elementary signal detection theory*. Oxford university press.

Xuan, W.; Yang, R.; Qi, H.; Zeng, Q.; Xiao, Y.; Xing, Y.; Wang, J.; Li, H.; Li, X.; Yu, K.; et al. 2025. Mmlu-prox: A multilingual benchmark for advanced large language model evaluation. *arXiv preprint arXiv:2503.10497*.

Ye, F.; Yang, M.; Pang, J.; Wang, L.; Wong, D.; Yilmaz, E.; Shi, S.; and Tu, Z. 2024. Benchmarking LLMs via Uncertainty Quantification. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, 15356–15385.

Ye, Q.; Fu, H.; Ren, X.; and Jia, R. 2023. How Predictable Are Large Language Model Capabilities? A Case Study on BIG-bench. In *Findings of the Association for Computational Linguistics: EMNLP*, 7493–7517.

Zhang, Z.; Zhao, X.; Fang, X.; Li, C.; Liu, X.; Min, X.; Duan, H.; Chen, K.; and Zhai, G. 2025. Redundancy Principles for MLLMs Benchmarks. In *Proceedings of the Association for Computational Linguistics (ACL)*.

Zhao, D.; Andrews, J. T.; Papakyriakopoulos, O.; and Xiang, A. 2024. Position: measure dataset diversity, don't just claim it. In *International Conference on Machine Learning (ICML)*, 60644–60673.

Zhou, H.; Huang, H.; Zhao, Z.; Han, L.; Wang, H.; Chen, K.; Yang, M.; Bao, W.; Dong, J.; Xu, B.; et al. 2025. Lost in Benchmarks? Rethinking Large Language Model Benchmarking with Item Response Theory. *arXiv preprint arXiv:2505.15055*.

Zhuang, Y.; Liu, Q.; Pardos, Z.; Kyllonen, P. C.; Zu, J.; Huang, Z.; Wang, S.; and Chen, E. 2025. Position: AI Evaluation Should Learn from How We Test Humans. In *International Conference on Machine Learning (ICML)*.