Lost in Benchmarks? Rethinking Large Language Model Benchmarking with Item Response Theory

Hongli Zhou[1*], Hui Huang[1*], Ziqing Zhao[1], Lvyuan Han[1], Huicheng Wang[1], Kehai Chen[2], Muyun Yang[1†], Wei Bao[3†], Jian Dong[3†], Bing Xu[1], Conghui Zhu[1], Hailong Cao[1], Tiejun Zhao[1]

[1]Faculty of Computing, Harbin Institute of Technology, Harbin, China
[2]School of Computer Science and Technology, Harbin Institute of Technology, Shenzhen, China
[3]China Electronics Standardization Institute, Beijing, China
hongli.joe@stu.hit.edu.cn, yangmuyun@hit.edu.cn, {baowei, dongjian}@cesi.cn

Abstract

The evaluation of large language models (LLMs) via benchmarks is widespread, yet inconsistencies between different leaderboards and poor separability among top models raise concerns about their ability to accurately reflect authentic model capabilities. This paper provides a critical analysis of benchmark effectiveness, examining mainstream prominent LLM benchmarks using results from diverse models. We first propose Pseudo-Siamese Network for Item Response Theory (PSN-IRT), an enhanced Item Response Theory framework that incorporates a rich set of item parameters within an IRT-grounded architecture. PSN-IRT can be utilized for accurate and reliable estimations of item characteristics and model abilities. Based on PSN-IRT, we conduct extensive analysis on 11 LLM benchmarks comprising 41,871 items, revealing significant and varied shortcomings in their measurement quality. Furthermore, we demonstrate that leveraging PSN-IRT is able to construct smaller benchmarks while maintaining stronger alignment with human preference.

Code — https://github.com/Joe-Hall-Lee/PSN-IRT
Extended version — https://arxiv.org/abs/2505.15055

1 Introduction

As the scale and performance of large language models (LLMs) continue to grow, accurately measuring their capabilities has become increasingly important (Chang et al. 2024; Zhou et al. 2024; Huang et al. 2025). Currently, the performance of LLMs is primarily evaluated through various benchmarks (Wang et al. 2024), which are comprehensive test suites consisting of carefully designed questions to assess model behavior across different tasks. However, in practice, existing benchmarks often exhibit significant limitations, prompting consideration of their effectiveness (McIntosh et al. 2024).

As illustrated in Figure 1, on one hand, even benchmarks that are designed to measure similar underlying capabilities often produce inconsistent leaderboards, leading to substantial ranking variations for the same model (Perlitz et al. 2024). On the other hand, many benchmarks show weak separability among top models, limiting the ability to understand performance differences and further model refinement. Given these limitations, there is a growing need for a more systematic analysis of LLM benchmarks.

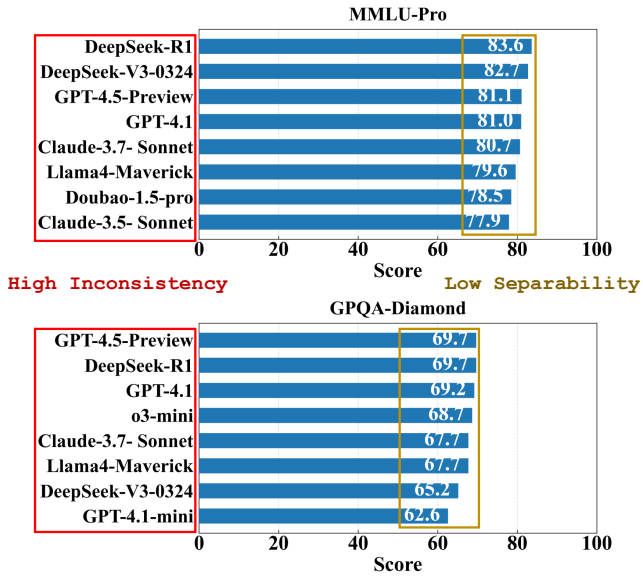

Figure 1: Illustration of weak separability and ranking inconsistencies in LLM benchmarks[1].

In this paper, we conduct a comprehensive analysis for various popular LLM benchmark datasets. Our experiment is based on Item Response Theory (IRT) (Lord, Novick, and Birnbaum 1968; Baker and Kim 2004), a psychometric framework widely used in educational assessment to analyze item properties such as difficulty by modeling item response against examinee ability[2]. However, traditional IRT struggles with the complexity and scale of modern datasets, and its assumption of the normal distribution of abilities

*These authors contributed equally.
†Corresponding authors.

[1]https://opencompass.org.cn
[2]In the paper, an item refers to a benchmark sample.

does not always hold.

To advance our objective, we propose the Pseudo-Siamese Network for IRT (PSN-IRT), a framework for comprehensive and interpretable analysis of benchmark test sets. PSN-IRT processes model identifiers and item identifiers through independent neural network pathways. These pathways respectively estimate latent model ability and a rich set of item parameters. The estimated properties are then integrated using an IRT-based formulation to predict outcome probabilities. This architectural design ensures both powerful modeling capabilities and strong theoretical interpretability.

To validate the effectiveness of PSN-IRT, we conduct extensive experiments on evaluation results from 12 LLMs across 11 benchmarks. Our results show that PSN-IRT outperforms previous approaches in both parameter estimation accuracy and reliability. Based on PSN-IRT, we perform in-depth analyses of these benchmarks using key metrics, with main findings as follows:

1. LLM benchmarks fail to achieve simultaneous excellence across multiple measurements.
2. LLM benchmarks suffer from widespread saturation and insufficient difficulty ceilings, limiting their ability to challenge and accurately evaluate advanced models.
3. LLM benchmarks exhibit data contamination in numerous items, compromising their reliability.

Furthermore, we find that model rankings derived from high-quality datasets selected by PSN-IRT are more consistent with human preference and offer stronger discriminability among top models.

Our contributions are summarized as follows:

1. We propose PSN-IRT, a benchmark analysis framework with superior estimation accuracy and reliability.
2. We use PSN-IRT for an in-depth analysis of mainstream benchmarks and find that current LLM benchmarks present deficiencies in many aspects.
3. We show that selecting items with PSN-IRT leads to model rankings that better align with human preference.

2 Background

2.1 Related Work

Researchers have explored various methods to evaluate and analyze the quality of LLM benchmarks. Recent efforts have introduced quantitative metrics that operate at a holistic, dataset level. For example, Benchmark Agreement Testing (BAT) assesses benchmark reliability by comparing the consistency of model rankings across different leaderboards (Perlitz et al. 2024; White et al. 2025). Other novel metrics have also been proposed for more nuanced dataset-level analysis (Li et al. 2024). Delving deeper than macro-level comparisons, another significant line of work focuses on content-based analysis to ground evaluation in concrete capabilities. For instance, ADeLe (Zhou et al. 2025) leverages scalable cognitive rubrics and item demand features to build highly interpretable capability profiles for AI systems.

Item Response Theory (IRT) (Lord, Novick, and Birnbaum 1968; Baker and Kim 2004) offers a psychometric framework for data-driven, item-level analysis. Within natural language processing (NLP), IRT has been primarily utilized to diagnose fundamental item properties such as difficulty and discriminability (Vania et al. 2021; Byrd and Srivastava 2022). More recently, its principles have also been used to inspire methods for improving evaluation efficiency, for instance, through optimized item selection (Maia Polo et al. 2024) or adaptive testing paradigms (Zhuang et al. 2025; Truong et al. 2025). However, due to the complexity and scale of modern datasets and its assumption of normally distributed abilities, the application of IRT on LLM benchmarks is still underexplored (Ye et al. 2025).

2.2 Preliminary: Item Response Theory

Item Response Theory (IRT) (Lord, Novick, and Birnbaum 1968) is a psychometric framework widely used in educational and cognitive assessments to model the relationship between benchmark items and examinee abilities. Unlike Classical Test Theory (CTT) (DeVellis 2006), which relies on aggregate test scores, IRT analyzes individual item responses as a function of a latent ability θ, enabling precise measurement of both item and examinee properties.

Central to IRT is the Item Characteristic Curve (ICC), a mathematical function that describes the probability of a correct response, $P(X = 1 \mid \theta)$, as a function of ability θ. Typically logistic, the ICC visualizes how item properties influence performance, serving as the foundation for IRT.

The simplest IRT model, the One-Parameter Logistic (1PL) model, defines the ICC as:

$$P(X = 1 \mid \theta) = \frac{1}{1 + e^{-(\theta - b)}} \tag{1}$$

where b denotes item difficulty. When $\theta = b$, the probability of the examinee generating a correct response is 0.5. The 1PL assumes all items have equal discriminability, a simplification that may not hold in complex testing scenarios.

More advanced IRT models introduce additional parameters to capture diverse item properties. However, traditional IRT often suffer from inaccurate parameter estimation and rely on assumptions like normal ability distributions, which may not align with real-world data (Tsutsumi, Kinoshita, and Ueno 2021).

3 Pseudo-Siamese Network for IRT

In this section, we propose the Pseudo-Siamese Network for IRT (PSN-IRT), designed to diagnose benchmark item quality by analyzing LLM responses. Specifically, one property can be inferred for each model: *model-ability*, and four properties can be inferred for each benchmark item: *discriminability*, *difficulty*, *guessing-rate* and *feasibility*.

Model Architecture. The PSN-IRT architecture comprises two independent networks: a model network and an item network. The model network processes one-hot encoded LLM identifiers to estimate *model-ability* θ, while the item network handles one-hot encoded item identifiers to produce the four IRT parameters: *discriminability* a, *difficulty* b, *guessing-rate* c, and *feasibility* d.

Each network consists of three fully connected layers with ReLU activation functions, enabling efficient and stable parameter estimation. After that, these estimated properties are then processed by a Logistic Calculation Layer. This layer operates using the Four-Parameter Logistic (4PL) model, an advanced IRT formulation that extends simpler models like the 1PL to capture nuanced behaviors:

$$P(X = 1 \mid \theta) = c + (d - c) \cdot \frac{1}{1 + e^{-a(\theta - b)}} \tag{2}$$

where each parameter controls a different aspect of benchmark item behavior:

- The difficulty b determines the ability level where the probability changes most significantly.
- The discriminability a reflects an item's power to differentiate between models of varying abilities.

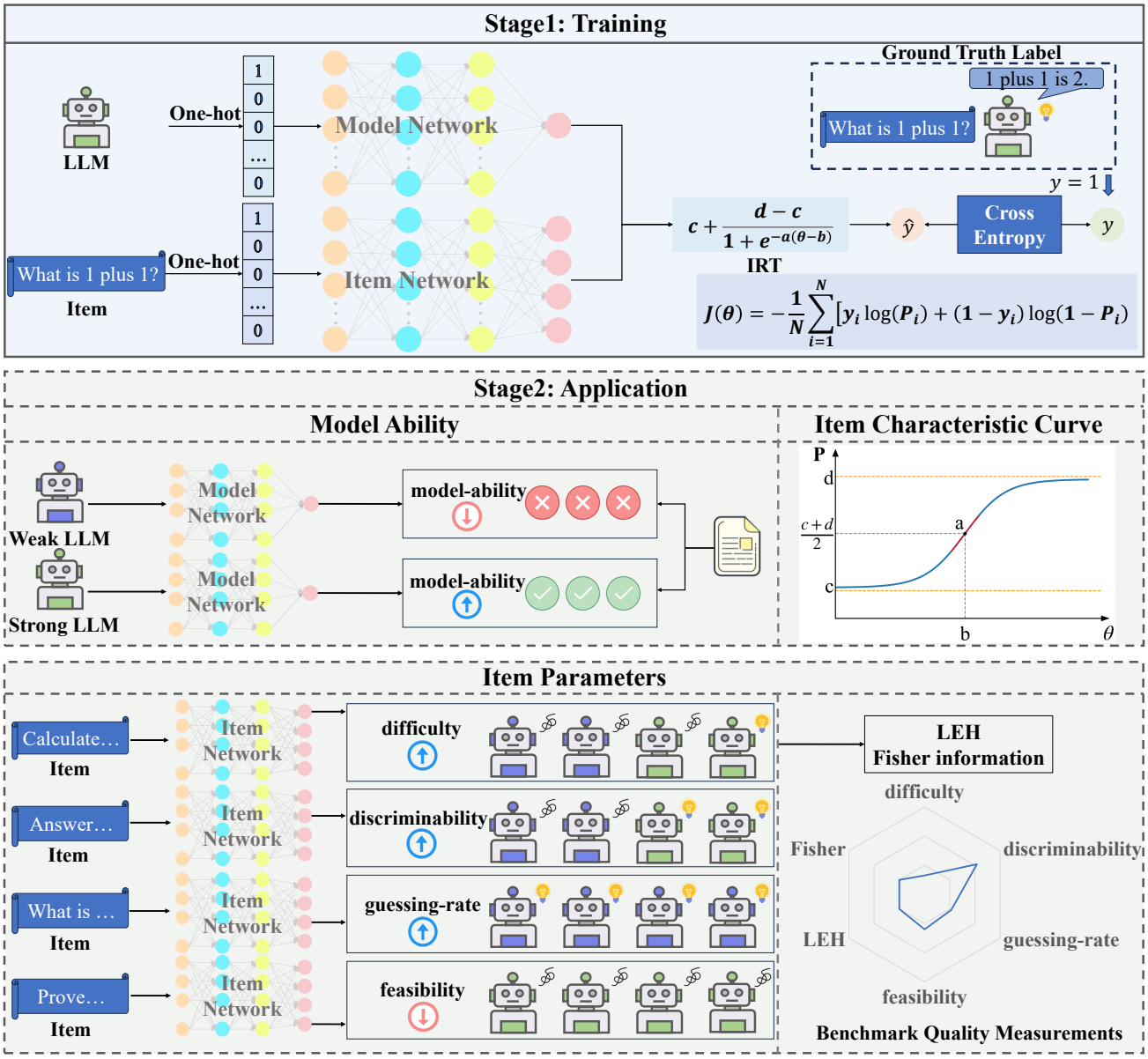

Figure 2: The illustration of our proposed PSN-IRT. Separate neural networks estimate model-ability (θ) and item parameters (a, b, c, d), which are then combined via the IRT formula to predict the probability of a correct response. After that, the networks can be leveraged for estimating properties for models or items, respectively.

- The guessing-rate c captures how likely models are to succeed without full understanding.
- The feasibility d represents the maximum probability that even highly proficient models will correctly answer the item.

Notably, PSN-IRT follows standard IRT assumptions such as unidimensionality and monotonicity, which generally hold for LLM benchmarking (Kipnis et al. 2025; Truong et al. 2025).

Training and Applications. The PSN-IRT is trained end-to-end using the observed binary response data in the form of (Model, Item, Response, Outcome), where the binary outcome indicates the correctness of each LLM on each benchmark item. During training, PSN-IRT processes encoded pairs of LLM and item identifiers through their respective network pathways. For each pair, the PSN-IRT framework first internally estimates the specific model and item properties. These estimated properties are then passed to a Logistic Calculation Layer, with the training objective as follows:

$$J(\theta) = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(P_i) + (1-y_i)\log(1-P_i)] \quad (3)$$

where N is the batch size. The training objective is to minimize the discrepancy between the model's prediction and the observed binary outcomes. Consequently, all learnable weights within both the model and item networks are updated simultaneously, concurrently optimizing the estimated properties for both models and benchmark items.

Upon completion of training, the two networks of PSN-IRT can be used for evaluating model performance and diagnosing benchmark characteristics, respectively. Specifically, the model network, by processing a model's encoding, outputs an estimate of the model-ability θ. Meanwhile, the item network, given a benchmark item's encoding, outputs its four estimated psychometric parameters a, b, c, d, which can be used for an in-depth diagnosis of individual item characteristics and the overall benchmark quality.

4 Efficacy of PSN-IRT

4.1 Experimental Setup

Datasets. To evaluate existing LLM benchmarks, we select 11 datasets that vary in domain coverage. They serve as a basis for comparing different IRT-based methods and analyzing test set properties. An overview of the datasets is provided in Table 1.

To construct training data, we conduct evaluations using OpenCompass (Contributors 2023). The evaluation results are collected in the form of binary outcomes, indicating whether each model answered each item correctly. Then, we construct a unified outcome matrix by aggregating item-level results across all models and benchmarks. We split the interactions into training, validation, and test sets.

We mainly focus on currently high-performing models, including 360GPT2-Pro[3], DeepSeek-V3[4], Doubao-pro[5], Gemini-1.5[6], Hunyuan-Turbo[7], Moonshot-v1[8], Qwen-Plus[9], and Yi-Lightning[10]. However, testing only strong models might obscure differences in the discriminability of the test sets, as most items could be solved easily, thus limiting a comprehensive evaluation of test set quality (Martínez-Plumed et al. 2019). To address this, we intentionally included several relatively weaker models, including Gemma-2B-it, Mistral-7B-Instruct-v0.1, Qwen2.5-3B-Instruct, and Vicuna-7B-v1.3 (Chiang et al. 2023).

Baselines. We mainly compare PSN-IRT with traditional IRT methods, namely IRT estimated based on traditional parameter estimation techniques, including Maximum Likelihood Estimation (MLE), Markov Chain Monte Carlo (MCMC) (Hastings 1970), Variational Inference (VI) (Jordan et al. 1999), and VIBO (Wu et al. 2020). We also compare with Deep-IRT (Tsutsumi, Kinoshita, and Ueno 2021), which is an extension of 1PL IRT that learns item-trait interactions via deep networks.

Metrics. To assess whether the learned parameters meaningfully reflect corresponding properties, we use two quantitative metrics:

- **Prediction accuracy**: Following standard practice in educational testing (Eignor 2013), we assess how well a method predicts whether a model answers an item correctly, reporting accuracy, F1 score, and ROC AUC.
- **Rank stability**: To evaluate the reliability of model ranking induced by each method, we split the test set into two subsets, estimate model abilities separately, and compute Kendall's τ between the resulting rankings.

[3]https://ai.360.cn/

[4]https://platform.deepseek.com/

[5]https://www.volcengine.com/product/doubao

[6]https://ai.google.dev/

[7]https://hunyuan.tencent.com/

[8]https://platform.moonshot.cn/

[9]https://qwen.ai/apiplatform

[10]https://platform.lingyiwanwu.com/

4.2 Main Results

Table 2 shows the estimation accuracy and reliability of PSN-IRT compared to traditional IRT and Deep-IRT. Both Deep-IRT and PSN-IRT significantly outperform traditional IRT in prediction accuracy. Despite its simpler Logistic output layer rooted in the IRT formula, PSN-IRT attains prediction accuracy on par with the more complex Deep-IRT, and surpasses it in rank reliability.

Given our goal to comprehensively capture diverse item characteristics, we adopt the more expressive 4PL structure for all methods except Deep-IRT. Experiments also confirm that PSN-IRT performs best with 4PL.

4.3 Ablation Study

To assess whether more complex architectures offer benefits over our straightforward design, we conduct an ablation study. Specifically, we compare PSN-IRT with two variants: one replaces one-hot inputs with semantic embeddings from Instructor (Su et al. 2023), and the other uses a GNN (Su et al. 2022) instead of the MLP backbone.

As shown in Table 3, both alternatives underperform our original design. We argue that semantic similarity between item texts does not necessarily reflect similarity in measurement characteristics. For the GNN variant, its neighborhood aggregation mechanism risks diluting the unique statistical signals of individual models and items, thereby hindering precise parameter estimation. These results suggest that in this estimation scenario, preserving the individuality of models and items is more effective than introducing architectural complexity.

Furthermore, since PSN-IRT estimates parameters that are specific to the training data, requiring retraining to analyze new items, we also investigate the stability of these estimates when the item pool changes. To simulate this, we conduct an experiment by independently training separate PSN-IRT models on random subsets of the items. We then compute the Pearson correlation between the item parameters estimated from these subsets and those estimated from the full dataset.

As shown in Table 4, the results reveal a high degree of correlation across all parameters. This indicates that PSN-IRT learns highly consistent intrinsic properties for the items regardless of the specific data sample, demonstrating the stability of parameter estimation.

5 Benchmark Analysis with PSN-IRT

5.1 Benchmark Quality Measurements

In this section, we delve into a detailed analysis of the benchmarks based on different measurements. Specifically, we leverage the 4 item parameters (difficulty, discriminability, guessing-rate and feasibility) estimated by PSN-IRT, along with two additional metrics for evaluating item quality: Local Efficiency Headroom (LEH) (Vania et al. 2021) and Fisher information (Lord 1980).

LEH score assesses the potential of a test example to evaluate future progress in LLMs. It is calculated as the derivative of the item's Item Characteristic Curve (ICC) with respect to the highest observed latent ability. A high LEH score indicates that even the top model remains far from the saturation point of the ICC.

Fisher information measures the amount of information an item provides about a model's ability level. Items with higher Fisher information are more informative for estimating model abilities. For 4PL IRT, it is defined as:

$$I(\theta) = \frac{a^2(P(\theta) - c)^2(d - P(\theta))^2}{(d - c)^2 P(\theta)(1 - P(\theta))}, \tag{4}$$

where $P(\theta)$ is defined as Equation 2.

Benchmark	Data Size	Domain	Metric	Format
ARC-C (Clark et al. 2018)	295	General	EM	Multiple Choice
BBH (Suzgun et al. 2023)	6,511	General	EM	Mixed
Chinese SimpleQA (He et al. 2024)	3,000	Knowledge	LLM-as-a-Judge	QA
GPQA Diamond (Rein et al. 2024)	198	Science	EM	Multiple Choice
GSM8K (Cobbe et al. 2021)	1,319	Math	EM	QA
HellaSwag (Zellers et al. 2019)	10,042	General	EM	Multiple Choice
HumanEval (Chen et al. 2021)	164	Code	Pass@1	QA
MATH (Hendrycks et al. 2021b)	5,000	Math	EM	QA
MBPP (Austin et al. 2021)	500	Code	Pass@1	QA
MMLU (Hendrycks et al. 2021a)	14,042	General	EM	Multiple Choice
TheoremQA (Chen et al. 2023)	800	Science	EM	QA

Table 1: Benchmarks used in this paper.

Model	Parameter	Method	ACC	F1	AUC	Kendall's τ
IRT	4PL	MLE	0.7211	0.8034	0.7012	0.9697
		MCMC	0.7070	0.7811	0.7278	0.9697
		VI	0.7201	0.8015	0.6940	0.9091
		VIBO	0.7188	0.8007	0.7055	0.9697
Deep-IRT	1PL	Deep Learning	0.7974	0.8516	**0.8519**	0.9697
PSN-IRT	4PL	Deep Learning	**0.7998**	**0.8538**	0.8485	**1.0000**

Table 2: Comparison of prediction accuracy and rank reliability across different methods.

Method	ACC	F1	AUC	Kendall
PSN-IRT	**0.7998**	**0.8538**	**0.8485**	**1.0000**
+ Embedding	0.7808	0.8413	0.8310	0.9394
+ GNN	0.7928	0.8490	0.8197	0.7273

Table 3: Ablation study on PSN-IRT's input representation and network architecture.

Data	Difficulty	Discrim.	Guessing	Feasibility
30%	0.9009	0.8186	0.8274	0.9025
50%	0.7437	0.7835	0.8776	0.9330
70%	0.9519	0.7414	0.8320	0.9442

Table 4: Pearson correlation coefficients of item parameters estimated on random subsets of the data versus those estimated on the full dataset. All correlations are statistically significant ($p < 0.0001$).

5.2 Benchmark Analysis

Based on the measurements, we provide an aggregate rank for each benchmark, calculated from the sum of individual ranks in Table 5, where a lower total rank value indicates better overall quality. Additionally, Figure 3 visualizes the distribution of item-level properties across the 11 LLM benchmarks. Building on these experimental results, our detailed analysis is presented as follows.

Finding 1: LLM benchmarks fail to achieve simultaneous excellence across multiple measurement properties. As shown in Table 5, the results clearly demonstrate this lack of simultaneous excellence. While Chinese SimpleQA achieves the best overall score, no benchmark performs universally well across the individual metrics. For instance, Chinese SimpleQA itself suffers from poor feasibility with a rank of 9. This pattern of strengths being offset by significant weaknesses is common across benchmarks, which underscores the inherent challenges and trade-offs in benchmark design. Moreover, the property distributions visualized in Figure 3 further highlight this variability and the difficulty for any single benchmark to achieve balanced, high quality across all desired measurement properties.

Finding 2: Current LLM benchmarks exhibit an insufficient difficulty ceiling to challenge the most advanced models. As shown in Table 5, difficulty is varied among different benchmarks. Datasets such as TheoremQA, Chinese SimpleQA, and GPQA Diamond lead in difficulty and continue to challenge many LLMs. In contrast, other benchmarks including ARC-C, HellaSwag, and MMLU have become comparatively easy, where many of the items are readily solved by high-performing models, diminishing their utility for differentiating top-tier systems.

Crucially, as shown in Figure 3, even the most challenging items in existing benchmarks exhibit relatively low difficulty levels. For instance, on the IRT scale, the highest item difficulty values rarely exceed 1.0, while top models such as DeepSeek-V3 have estimated abilities well above 3.0. This indicates that even the hardest questions are significantly below the capability level of elite LLMs, highlighting a lack of sufficient challenge for frontier capabilities.

Finding 3: Generally low LEH scores across most benchmarks reveal widespread item saturation. While Table 5 indicates that datasets like GPQA Diamond and TheoremQA achieve the highest relative ranks for average LEH, suggesting they offer more headroom than other benchmarks, the absolute LEH values in Figure 3 are often smaller than ideal. Conversely, benchmarks such as GSM8K and ARC-C exhibit lower average LEH scores. This signifies that current high-performing models are already approaching or have reached the performance ceiling for a substantial portion of items within these test sets. This necessitates the development of new benchmark items and datasets explicitly designed with greater headroom as models continue to evolve.

Dataset	Difficulty	Discriminability	Guessing	Feasibility	LEH	Fisher	Total
Chinese SimpleQA	2	3	1	9	3	5	23
TheoremQA	1	10	3	11	2	2	29
MBPP	5	6	4	8	4	4	31
MATH	4	1	2	7	7	10	31
GPQA Diamond	3	11	6	10	1	1	32
BBH	6	5	8	6	6	7	38
MMLU	9	9	9	5	5	3	40
GSM8K	8	2	5	3	11	11	40
HumanEval	7	4	7	4	9	9	40
HellaSwag	10	7	10	2	8	6	43
ARC-C	11	8	11	1	10	8	49

Table 5: Ranks of datasets based on average item-level properties estimated by PSN-IRT. Each rank is obtained by averaging the property over all items in a dataset and then sorting these averages; smaller rank values correspond to better performance.

Finding 4: Outlier-high guessing-rates in many benchmark items flag risks of data contamination. The guessing-rate reflects the chance that a model can answer a question correctly without actual understanding, typically by exploiting format-based cues or shortcuts. As shown in Figure 3, datasets like ARC-C, HellaSwag, and MMLU generally present higher guessing-rates, possibly due to their multiple-choice formats. Notably, further inspection of the item guessing-rate parameter distributions, as depicted in Figure 3, reveals that a considerable number of items across several benchmarks exhibit unusually high guessing-rates. This observation regards the potential for such items to indicate data contamination, namely the content or answers for these items might have been present in the models' pretraining data (Zhuang et al. 2025).

Finding 5: Widespread low-feasibility items in scientific benchmarks reveal flawed question design. Feasibility estimates how well a question can be answered based on the information provided. Low feasibility often arises from underspecified or overly broad questions, where even a strong model cannot determine a clear answer. As shown in Figure 3, datasets like TheoremQA and GPQA Diamond exhibit lower feasibility, likely due to complex scientific contexts or vague problem descriptions. In contrast, ARC-C and HellaSwag have high feasibility, suggesting their questions are well-formed and specific. Low feasibility more often reflects weaknesses in annotation quality or task design, and such items can distort evaluation by penalizing models for failing to resolve ambiguities (Rodriguez et al. 2021).

6 Efficient Benchmarking with PSN-IRT

While the pursuit of comprehensive LLM evaluation has led to increasingly large benchmarks, sheer item volume neither guarantees quality nor comes without significant computational cost. This section, therefore, explores how PSN-IRT can facilitate more efficient benchmarking. We investigate whether strategically curating smaller, highly informative item sets can yield model comparisons that maintain or even surpass the accuracy and reliability of larger, undifferentiated collections.

Experimental Setup. To evaluate the effectiveness of different item selection strategies, we design an experiment to evaluate the effectiveness of different item selection strategies. Our methodology involves Benchmark Agreement Testing (BAT) (Perlitz et al. 2024). We first establish a reference model ranking by aggregating results from two public leaderboards: Chatbot Arena (Chiang et al. 2024) and OpenCompass Arena (Contributors 2023)[11].

Method	w/ weak models		w/o weak models	
	Variance	Kendall	Variance	Kendall
All	0.0434	0.6444	0.0010	0.2381
Top 400				
Random	0.0421	0.6444	0.0007	0.2381
Clustering	0.0308	0.6000	**0.0058**	0.1429
Discrim.	**0.1841**	0.5556	0.0000	0.2381
Fisher	0.0049	**0.6889**	0.0033	**0.7143**
Top 1000				
Random	0.0406	0.6444	0.0009	0.2381
Clustering	0.0166	0.6444	**0.0041**	0.2381
Discrim.	**0.1805**	0.6000	0.0000	0.3813
Fisher	0.0039	**0.6889**	0.0030	**0.9048**

Table 6: Variance and Kendall's τ for various data selection strategies with and without weak models.

Subsequently, various subsets of benchmark items are chosen based on different criteria. Model rankings generated using these subsets are then compared against the reference arena ranking using Kendall's τ to measure agreement. Alongside agreement, we also consider the variance of model scores on these subsets; higher variance generally indicates the subset's capacity to differentiate more clearly between models. The evaluations are conducted using two distinct model groupings: one set comprising a mix of stronger and weaker models ("w/ weak models"), and another set from which weaker models are excluded ("w/o weak models"), to specifically assess separability among stronger contenders, with weak models introduced in Section 4.1.

Results. As shown in Table 6, selecting items based on Fisher information consistently produces model rankings with superior alignment to the reference arena ranking. For instance, a carefully selected subset composed of just 1,000 items chosen via the Fisher information criterion achieves a Kendall's τ of up to 0.9048. This level of agreement markedly surpasses that achieved by rankings derived from using all available items or from similarly sized randomly selected subsets.

In contrast, other item selection approaches are less effective for distinguishing strong models. For instance, selecting items based

[11]These arenas are based on human evaluation and thus reflect human preference regarding overall model capabilities.

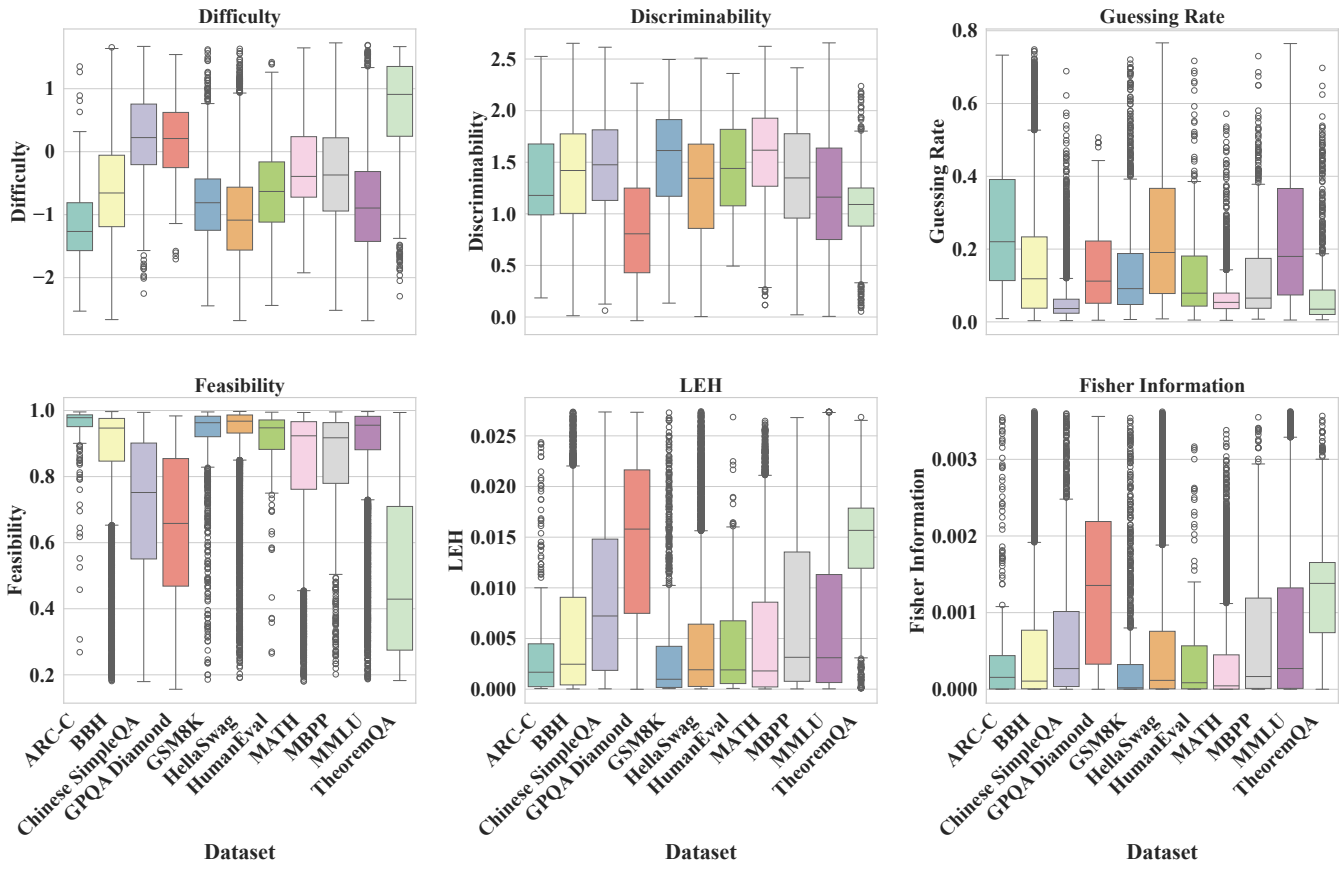

Figure 3: Distribution of item-level properties across 11 LLM benchmarks.

on discriminability yields zero score variance when weak models are excluded, indicating it merely separates broad capability tiers rather than providing granularity among top-performers. Separately, clustering items based on their success/failure vectors (Pacchiardi, Cheke, and Hernández-Orallo 2024) also fails to improve ranking correlation, even when it increases score separability.

7 Conclusion

In this work, we introduce PSN-IRT, a framework that combines psychometrics with neural networks to provide comprehensive and reliable analyses of LLM benchmarks. Through extensive experiments on 12 LLMs across 11 diverse datasets, we demonstrate that current benchmarks often suffer from uneven measurement properties, insufficient difficulty ceilings, item saturation, and data contamination. Moreover, we show that strategically selecting smaller, high-quality item sets using PSN-IRT enhances alignment with human preference and separability among top models.

Acknowledgments

This work is supported by National Natural Science Foundation of China (62276077, 62376075, 62376076), Department of Science and Technology of Heilongjiang (Grant No. ZY04JD04), and the Key Laboratory of Computing Power Network and Information Security, Ministry of Education (Grant No. 2023ZD027).

References

Austin, J.; Odena, A.; Nye, M.; Bosma, M.; Michalewski, H.; Dohan, D.; Jiang, E.; Cai, C.; Terry, M.; Le, Q.; and Sutton, C. 2021. Program Synthesis with Large Language Models. arXiv:2108.07732.

Baker, F. B.; and Kim, S.-H. 2004. *Item response theory: Parameter estimation techniques*. CRC press.

Byrd, M.; and Srivastava, S. 2022. Predicting Difficulty and Discrimination of Natural Language Questions. In Muresan, S.; Nakov, P.; and Villavicencio, A., eds., *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 119–130. Dublin, Ireland: Association for Computational Linguistics.

Chang, Y.; Wang, X.; Wang, J.; Wu, Y.; Yang, L.; Zhu, K.; Chen, H.; Yi, X.; Wang, C.; Wang, Y.; Ye, W.; Zhang, Y.; Chang, Y.; Yu, P. S.; Yang, Q.; and Xie, X. 2024. A Survey on Evaluation of Large Language Models. *ACM Trans. Intell. Syst. Technol.*, 15(3).

Chen, M.; Tworek, J.; Jun, H.; Yuan, Q.; de Oliveira Pinto, H. P.; Kaplan, J.; Edwards, H.; Burda, Y.; Joseph, N.; Brockman, G.; Ray, A.; Puri, R.; Krueger, G.; Petrov, M.; Khlaaf, H.; Sastry, G.; Mishkin, P.; Chan, B.; Gray, S.; Ryder, N.; Pavlov, M.; Power, A.; Kaiser, L.; Bavarian, M.; Winter, C.; Tillet, P.; Such, F. P.; Cummings, D.; Plappert, M.; Chantzis, F.; Barnes, E.; Herbert-Voss, A.; Guss, W. H.; Nichol, A.; Paino, A.; Tezak, N.; Tang, J.; Babuschkin, I.; Balaji, S.; Jain, S.; Saunders, W.; Hesse, C.; Carr, A. N.; Leike, J.; Achiam, J.; Misra, V.; Morikawa, E.; Radford, A.; Knight, M.; Brundage, M.; Murati, M.; Mayer, K.; Welinder,

P.; McGrew, B.; Amodei, D.; McCandlish, S.; Sutskever, I.; and Zaremba, W. 2021. Evaluating Large Language Models Trained on Code.

Chen, W.; Yin, M.; Ku, M.; Lu, P.; Wan, Y.; Ma, X.; Xu, J.; Wang, X.; and Xia, T. 2023. TheoremQA: A Theorem-driven Question Answering Dataset. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 7889–7901. Singapore: Association for Computational Linguistics.

Chiang, W.-L.; Li, Z.; Lin, Z.; Sheng, Y.; Wu, Z.; Zhang, H.; Zheng, L.; Zhuang, S.; Zhuang, Y.; Gonzalez, J. E.; Stoica, I.; and Xing, E. P. 2023. Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality.

Chiang, W.-L.; Zheng, L.; Sheng, Y.; Angelopoulos, A. N.; Li, T.; Li, D.; Zhu, B.; Zhang, H.; Jordan, M.; Gonzalez, J. E.; and Stoica, I. 2024. Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference. In Salakhutdinov, R.; Kolter, Z.; Heller, K.; Weller, A.; Oliver, N.; Scarlett, J.; and Berkenkamp, F., eds., *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, 8359–8388. PMLR.

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. *arXiv:1803.05457v1*.

Cobbe, K.; Kosaraju, V.; Bavarian, M.; Chen, M.; Jun, H.; Kaiser, L.; Plappert, M.; Tworek, J.; Hilton, J.; Nakano, R.; Hesse, C.; and Schulman, J. 2021. Training Verifiers to Solve Math Word Problems. *arXiv preprint arXiv:2110.14168*.

Contributors, O. 2023. OpenCompass: A Universal Evaluation Platform for Foundation Models. https://github.com/open-compass/opencompass.

DeVellis, R. F. 2006. Classical test theory. *Medical care*, 44(11): S50–S59.

Eignor, D. R. 2013. The standards for educational and psychological testing.

Hastings, W. K. 1970. Monte Carlo sampling methods using Markov chains and their applications.

He, Y.; Li, S.; Liu, J.; Tan, Y.; Wang, W.; Huang, H.; Bu, X.; Guo, H.; Hu, C.; Zheng, B.; Lin, Z.; Liu, X.; Sun, D.; Lin, S.; Zheng, Z.; Zhu, X.; Su, W.; and Zheng, B. 2024. Chinese SimpleQA: A Chinese Factuality Evaluation for Large Language Models. arXiv:2411.07140.

Hendrycks, D.; Burns, C.; Basart, S.; Zou, A.; Mazeika, M.; Song, D.; and Steinhardt, J. 2021a. Measuring Massive Multitask Language Understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Hendrycks, D.; Burns, C.; Kadavath, S.; Arora, A.; Basart, S.; Tang, E.; Song, D.; and Steinhardt, J. 2021b. Measuring Mathematical Problem Solving With the MATH Dataset. *NeurIPS*.

Huang, H.; Bu, X.; Zhou, H.; Qu, Y.; Liu, J.; Yang, M.; Xu, B.; and Zhao, T. 2025. An Empirical Study of LLM-as-a-Judge for LLM Evaluation: Fine-tuned Judge Model is not a General Substitute for GPT-4. In Che, W.; Nabende, J.; Shutova, E.; and Pilehvar, M. T., eds., *Findings of the Association for Computational Linguistics: ACL 2025*, 5880–5895. Vienna, Austria: Association for Computational Linguistics. ISBN 979-8-89176-256-5.

Jordan, M. I.; Ghahramani, Z.; Jaakkola, T. S.; and Saul, L. K. 1999. An introduction to variational methods for graphical models. *Machine learning*, 37: 183–233.

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 *The Thirteenth International Conference on Learning Representations*.

Li, T.; Chiang, W.-L.; Frick, E.; Dunlap, L.; Wu, T.; Zhu, B.; Gonzalez, J. E.; and Stoica, I. 2024. From Crowdsourced Data to High-Quality Benchmarks: Arena-Hard and BenchBuilder Pipeline. *arXiv preprint arXiv:2406.11939*.

Lord, F.; Novick, M.; and Birnbaum, A. 1968. Statistical theories of mental test scores.

Lord, F. M. 1980. Applications of Item Response Theory to Practical Testing Problems.

Maia Polo, F.; Weber, L.; Choshen, L.; Sun, Y.; Xu, G.; and Yurochkin, M. 2024. tinyBenchmarks: evaluating LLMs with fewer examples. In Salakhutdinov, R.; Kolter, Z.; Heller, K.; Weller, A.; Oliver, N.; Scarlett, J.; and Berkenkamp, F., eds., *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, 34303–34326. PMLR.

Martínez-Plumed, F.; Prudêncio, R. B.; Martínez-Usó, A.; and Hernández-Orallo, J. 2019. Item response theory in AI: Analysing machine learning classifiers at the instance level. *Artificial intelligence*, 271: 18–42.

McIntosh, T. R.; Susnjak, T.; Arachchilage, N.; Liu, T.; Watters, P.; and Halgamuge, M. N. 2024. Inadequacies of large language model benchmarks in the era of generative artificial intelligence. *arXiv preprint arXiv:2402.09880*.

Pacchiardi, L.; Cheke, L. G.; and Hernández-Orallo, J. 2024. 100 instances is all you need: predicting the success of a new LLM on unseen data by testing on a few instances. arXiv:2409.03563.

Perlitz, Y.; Gera, A.; Arviv, O.; Yehudai, A.; Bandel, E.; Shnarch, E.; Shmueli-Scheuer, M.; and Choshen, L. 2024. Do These LLM Benchmarks Agree? Fixing Benchmark Evaluation with BenchBench. arXiv:2407.13696.

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*.

Rodriguez, P.; Barrow, J.; Hoyle, A. M.; Lalor, J. P.; Jia, R.; and Boyd-Graber, J. 2021. Evaluation Examples are not Equally Informative: How should that change NLP Leaderboards? In Zong, C.; Xia, F.; Li, W.; and Navigli, R., eds., *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 4486–4503. Online: Association for Computational Linguistics.

Su, H.; Shi, W.; Kasai, J.; Wang, Y.; Hu, Y.; Ostendorf, M.; Yih, W.-t.; Smith, N. A.; Zettlemoyer, L.; and Yu, T. 2023. One Embedder, Any Task: Instruction-Finetuned Text Embeddings. In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., *Findings of the Association for Computational Linguistics: ACL 2023*, 1102–1121. Toronto, Canada: Association for Computational Linguistics.

Su, Y.; Cheng, Z.; Wu, J.; Dong, Y.; Huang, Z.; Wu, L.; Chen, E.; Wang, S.; and Xie, F. 2022. Graph-based cognitive diagnosis for intelligent tutoring systems. *Knowledge-Based Systems*, 253: 109547.

Suzgun, M.; Scales, N.; Schärli, N.; Gehrmann, S.; Tay, Y.; Chung, H. W.; Chowdhery, A.; Le, Q.; Chi, E.; Zhou, D.; and Wei, J. 2023. Challenging BIG-Bench Tasks and Whether Chain-of-Thought Can Solve Them. In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., *Findings of the Association for Computational*

Linguistics: ACL 2023, 13003–13051. Toronto, Canada: Association for Computational Linguistics.

Truong, S. T.; Tu, Y.; Liang, P.; Li, B.; and Koyejo, S. 2025. Reliable and Efficient Amortized Model-based Evaluation. In Singh, A.; Fazel, M.; Hsu, D.; Lacoste-Julien, S.; Berkenkamp, F.; Maharaj, T.; Wagstaff, K.; and Zhu, J., eds., *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, 60238–60265. PMLR.

Tsutsumi, E.; Kinoshita, R.; and Ueno, M. 2021. Deep item response theory as a novel test theory based on deep learning. *Electronics*, 10(9): 1020.

Vania, C.; Htut, P. M.; Huang, W.; Mungra, D.; Pang, R. Y.; Phang, J.; Liu, H.; Cho, K.; and Bowman, S. R. 2021. Comparing Test Sets with Item Response Theory. In Zong, C.; Xia, F.; Li, W.; and Navigli, R., eds., *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 1141–1158. Online: Association for Computational Linguistics.

Wang, Y.; Ma, X.; Zhang, G.; Ni, Y.; Chandra, A.; Guo, S.; Ren, W.; Arulraj, A.; He, X.; Jiang, Z.; Li, T.; Ku, M.; Wang, K.; Zhuang, A.; Fan, R.; Yue, X.; and Chen, W. 2024. MMLU-Pro: A More Robust and Challenging Multi-Task Language Understanding Benchmark. In Globerson, A.; Mackey, L.; Belgrave, D.; Fan, A.; Paquet, U.; Tomczak, J.; and Zhang, C., eds., *Advances in Neural Information Processing Systems*, volume 37, 95266–95290. Curran Associates, Inc.

White, C.; Dooley, S.; Roberts, M.; Pal, A.; Feuer, B.; Jain, S.; Shwartz-Ziv, R.; Jain, N.; Saifullah, K.; Dey, S.; Shubh-Agrawal; Sandha, S. S.; Naidu, S. V.; Hegde, C.; LeCun, Y.; Goldstein, T.; Neiswanger, W.; and Goldblum, M. 2025. LiveBench: A Challenging, Contamination-Limited LLM Benchmark. In *The Thirteenth International Conference on Learning Representations*.

Wu, M.; Davis, R. L.; Domingue, B. W.; Piech, C.; and Goodman, N. D. 2020. Variational Item Response Theory: Fast, Accurate, and Expressive. In Rafferty, A. N.; Whitehill, J.; Romero, C.; and Cavalli-Sforza, V., eds., *Proceedings of the 13th International Conference on Educational Data Mining, EDM 2020, Fully virtual conference, July 10-13, 2020*. International Educational Data Mining Society. ISBN 978-1-7336736-1-7.

Ye, H.; Jin, J.; Xie, Y.; Zhang, X.; and Song, G. 2025. Large language model psychometrics: A systematic review of evaluation, validation, and enhancement. *arXiv preprint arXiv:2505.08245*.

Zellers, R.; Holtzman, A.; Bisk, Y.; Farhadi, A.; and Choi, Y. 2019. HellaSwag: Can a Machine Really Finish Your Sentence? In Korhonen, A.; Traum, D.; and Màrquez, L., eds., *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 4791–4800. Florence, Italy: Association for Computational Linguistics.

Zhou, H.; Huang, H.; Long, Y.; Xu, B.; Zhu, C.; Cao, H.; Yang, M.; and Zhao, T. 2024. Mitigating the Bias of Large Language Model Evaluation. In Maosong, S.; Jiye, L.; Xianpei, H.; Zhiyuan, L.; and Yulan, H., eds., *Proceedings of the 23rd Chinese National Conference on Computational Linguistics (Volume 1: Main Conference)*, 1310–1319. Taiyuan, China: Chinese Information Processing Society of China.

Zhou, L.; Pacchiardi, L.; Martínez-Plumed, F.; Collins, K. M.; Moros-Daval, Y.; Zhang, S.; Zhao, Q.; Huang, Y.; Sun, L.; Prunty, J. E.; et al. 2025. General scales unlock ai evaluation with explanatory and predictive power. *arXiv preprint arXiv:2503.06378*.

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 Singh, A.; Fazel, M.; Hsu, D.; Lacoste-Julien, S.; Berkenkamp, F.; Maharaj, T.; Wagstaff, K.; and Zhu, J., eds., *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, 82483–82508. PMLR.