How Many Items Do You Really Need? IRT-Based Redundancy Analysis of LLM Benchmarks

Dinesh Katupputhur Ramprasath
Stanford University
din1993@stanford.edu

Abstract

LLM benchmarks contain hundreds to thousands of evaluation items, yet it remains unclear how many are actually necessary to produce reliable model rankings. We apply Item Response Theory (IRT) to six LLM benchmarks from the Fantastic-Bugs dataset, fitting Rasch, Two-Parameter Logistic (2PL), and Logistic Factor Model (LogisticFM) models to binary response matrices containing 42–91 language models and 500–3,316 items. We systematically compare five item selection strategies—random, stratified, max-information at $\theta{=}0$ (tinyBenchmarks), integrated information, and difficulty-coverage—and evaluate their ability to recover full-benchmark rankings using Spearman ρ, Kendall τ, and top-5 agreement. Our key finding is that **IRT-guided item selection recovers model rankings using as few as 5% of benchmark items** ($\rho > 0.95$) for well-designed benchmarks (MMLU, MedQA), while benchmarks with poor psychometric structure (BBQ) are irreducible. We characterize what makes a benchmark reducible—discrimination concentration, difficulty spread, and unidimensionality—and provide actionable design guidelines for benchmark developers. All code and data are available at `https://github.com/dineshkvr1/cs321m-irt-benchmark-redundancy`.

1 Introduction

The rapid proliferation of large language models (LLMs) has created an urgent need for efficient and principled evaluation methodology. Current practice relies on benchmark leaderboards that report aggregate accuracy scores computed over hundreds or thousands of evaluation items. Evaluating a single model on MMLU requires running inference on 14,042 items; with hundreds of new models released monthly, evaluation cost scales linearly with item count.

A natural question arises: *Are all of these items necessary?* If a small subset can recover the same model ranking as the full benchmark, the remaining items are *redundant*—they increase evaluation cost without improving measurement fidelity. Understanding which items carry ranking information, and which contribute noise, is both practically important (reducing evaluation cost) and scientifically revealing (exposing the latent structure of benchmarks).

Item Response Theory (IRT), the dominant measurement framework in educational and psychological testing [Lord, 1980, Embretson and Reise, 2000], provides principled tools to address this question. IRT decomposes observed responses into latent parameters—model ability (θ), item difficulty (β), and item discrimination (α)—enabling fine-grained analysis of both models and evaluation items. Crucially, IRT's Fisher Information Function quantifies how much information each item provides about model ability, enabling theoretically grounded item selection.

Recent work has demonstrated the promise of IRT for AI evaluation [Martínez-Plumed et al., 2019, Polo et al., 2025, Truong et al., 2026], and the tinyBenchmarks methodology [Polo et al., 2024] showed that IRT-guided item selection can dramatically reduce evaluation cost. However, existing

Preprint.

work has not (a) systematically compared item selection strategies, (b) explained which benchmark properties predict reducibility, or (c) provided clear guidelines for benchmark designers.

Research question: *What psychometric properties of a benchmark determine whether IRT-guided item selection can recover full-benchmark model rankings from a small subset, and which selection strategy is most effective?*

In this work, we conduct a deep redundancy analysis across six diverse LLM benchmarks, making three contributions:

1. We compare five item selection strategies and show that IRT-guided selection consistently outperforms random baselines, recovering rankings with 5–20% of items for well-designed benchmarks (§4.3).
2. We identify psychometric properties—discrimination concentration, difficulty spread, and dimensionality—that predict benchmark reducibility (§4.5).
3. We provide actionable recommendations for benchmark design based on measurement-theoretic analysis (§5).

2 Background and Related Work

Item Response Theory. IRT models the probability that model i correctly answers item j as a function of latent parameters. The Rasch (1PL) model specifies $P(Y_{ij} = 1) = \sigma(\theta_i - \beta_j)$, where σ is the logistic function, θ_i is model ability, and β_j is item difficulty. The 2PL model adds item discrimination: $P(Y_{ij} = 1) = \sigma(\alpha_j(\theta_i - \beta_j))$. The Logistic Factor Model (LogisticFM) generalizes to multidimensional ability with K latent factors [Embretson and Reise, 2000].

Fisher Information. The Fisher information of item j at ability level θ under the 2PL model is:

$$I_j(\theta) = \alpha_j^2 \, P_j(\theta) \, (1 - P_j(\theta)) \tag{1}$$

where $P_j(\theta) = \sigma(\alpha_j(\theta - \beta_j))$. Items with high discrimination α_j and difficulty β_j near θ contribute the most information. The Test Information Function $I(\theta) = \sum_j I_j(\theta)$ characterizes the benchmark's overall measurement precision across the ability range [Lord, 1980].

IRT for AI evaluation. Martínez-Plumed et al. [2019] first applied IRT to ML model evaluation. Polo et al. [2025] introduced the Fantastic-Bugs dataset with response matrices across ten benchmarks and demonstrated systematic evaluation anomalies. Truong et al. [2026] established Item Response Scaling Laws connecting IRT ability to computational scale. Truong et al. [2025] developed amortized IRT inference methods scaling to 22 benchmarks and 172 LMs.

Efficient evaluation. Polo et al. [2024] demonstrated that small item subsets can approximate full-benchmark LLM rankings using IRT-guided selection, selecting items by Fisher information at a fixed anchor θ=0. Zhuang et al. [2024] applied Computerized Adaptive Testing (CAT) to LLM evaluation, showing 80% cost reduction. Chang and Ying [1996] proposed global information criteria that consider the entire ability distribution rather than a single anchor point. Our work extends tinyBenchmarks by comparing multiple selection strategies and characterizing benchmark reducibility.

3 Data and Methods

3.1 Fantastic-Bugs Dataset

We use the Fantastic-Bugs dataset [Polo et al., 2025], available on HuggingFace (`stair-lab/fantastic-bugs`). From the eleven available benchmarks, we select six that span diverse evaluation domains and psychometric properties:

These benchmarks were deliberately chosen to span the reducibility spectrum, from highly reducible (MMLU) to irreducible (BBQ), enabling us to study what makes a benchmark amenable to item reduction.

Table 1: Selected benchmarks from the Fantastic-Bugs dataset.

Benchmark	Models	Items	Domain
MedQA	91	998	Medical knowledge
OpenBookQA	91	500	Commonsense reasoning
BoolQ	67	3,316	Boolean question answering
LegalBench	91	1,997	Legal reasoning
MMLU	91	565	Multi-domain knowledge
BBQ	42	1,000	Bias evaluation

3.2 IRT Model Fitting

For each benchmark, we fit three IRT model families using `torch_measure` [AIMS Lab, Stanford University, 2024]:

- **Rasch (1PL):** $P(Y_{ij} = 1) = \sigma(\theta_i - \beta_j)$. Parameters: $N + M$.
- **2PL:** $P(Y_{ij} = 1) = \sigma(\alpha_j(\theta_i - \beta_j))$. Parameters: $N + 2M$.
- **LogisticFM (K=2):** Multidimensional with K=2 latent factors. Parameters: $2N + M(K + 1)$.

All models are fit via MLE using the Adam optimizer (lr=0.05, 500 epochs, weight_decay=0.01, seed=42). Missing responses are masked during likelihood computation. We use AIC and BIC for model selection [Burnham and Anderson, 2002].

3.3 Item Selection Strategies

We compare five item selection strategies, holding the number of selected items k fixed:

1. **Random:** Uniform random sample of k items (averaged over 50 draws).
2. **Stratified:** Stratify items by 2PL difficulty tercile, sample proportionally from each stratum (averaged over 50 draws).
3. **Max-info at θ=0:** Select k items with highest $I_j(0) = \alpha_j^2 P_j(0)(1 - P_j(0))$. This is the tinyBenchmarks approach [Polo et al., 2024].
4. **Integrated info:** Select k items maximizing $\bar{I}_j = \int I_j(\theta)\hat{f}(\theta)\, d\theta$, where $\hat{f}(\theta)$ is estimated from the 2PL ability distribution via kernel density estimation [Chang and Ying, 1996].
5. **Difficulty-coverage:** Select k items at evenly-spaced quantiles of the difficulty distribution, ensuring measurement across the full ability range.

3.4 Evaluation Protocol

For each benchmark $\times$ strategy $\times$ item fraction:

1. Select k items using the strategy.
2. Compute model accuracy on the selected items.
3. Rank models by subset accuracy.
4. Compare to the full-benchmark ranking.

We vary item fractions at $\{1\%, 2\%, 5\%, 10\%, 20\%, 50\%, 100\%\}$. For stochastic strategies (random, stratified), we average over 50 independent draws and report 95% confidence intervals.

Metrics.

- **Primary:** Spearman ρ between subset and full-benchmark model rankings.
- **Secondary:** Kendall τ; top-5 agreement (fraction of top-5 models preserved); maximum rank displacement $\max_i |r_i^{\text{sub}} - r_i^{\text{full}}|$.
- **Efficiency:** k_{95}—minimum items for $\rho > 0.95$.

Table 2: 2PL parameter estimates across benchmarks. $\bar{\theta}$: mean ability; σ_θ: ability spread; $\bar{\beta}$: mean difficulty; σ_β: difficulty spread; $\tilde{\alpha}$: median discrimination.

Benchmark	$\bar{\theta}$	σ_θ	$\bar{\beta}$	σ_β	$\tilde{\alpha}$
MedQA	0.04	0.80	0.11	1.76	2.19
OpenBookQA	0.77	1.33	−1.09	1.17	2.04
BoolQ	0.52	0.74	−0.28	1.67	2.01
LegalBench	−0.15	0.37	0.09	2.07	1.65
MMLU	−0.13	0.66	0.09	2.05	2.38
BBQ	0.01	1.27	0.58	1.78	1.12

Table 3: Model comparison (AIC/n and BIC/n; lower is better). Best in each column is **bold**.

Benchmark	**Model**	**AIC/n**	**BIC/n**
MedQA	Rasch	0.901	**1.014**
	2PL	0.849	1.065
	LogisticFM	**0.825**	1.154
MMLU	Rasch	0.798	**0.911**
	2PL	0.738	0.949
	LogisticFM	**0.715**	1.038
BBQ	Rasch	1.235	**1.449**
	2PL	1.169	1.590
	LogisticFM	**0.902**	1.537

4 Results

4.1 IRT Model Fit

All 18 IRT models (3 families $\times$ 6 benchmarks) converge successfully. Table 2 summarizes key 2PL parameter estimates.

A consistent pattern emerges in model selection: Rasch wins by BIC (most parsimonious) while LogisticFM wins by AIC (best fit) across all six benchmarks, with 2PL occupying an intermediate position (Table 3). This tension indicates that LLM benchmark responses contain multidimensional structure beyond a single ability axis, but the additional parameters may not generalize.

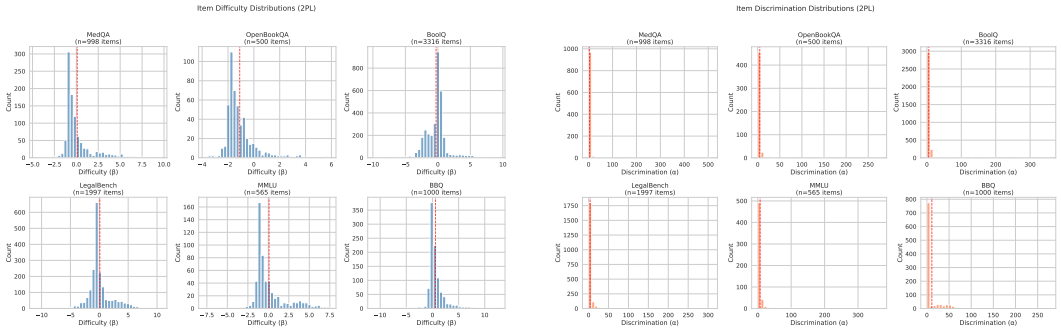
4.2 Psychometric Structure

Figure 1 shows the difficulty and discrimination distributions across benchmarks. MMLU and MedQA have wide difficulty ranges spanning approximately 8 logits, providing measurement across a broad ability range. BBQ's low median discrimination ($\tilde{\alpha} = 1.12$) and compressed difficulty distribution are consistent with its role as a bias benchmark rather than a capability measure.

4.3 Item Selection: Strategy Comparison

Figure 2 shows ranking recovery (ρ) as a function of item fraction for the max-information strategy. Table 4 reports the minimum items needed for $\rho > 0.95$ (k_{95}).

Strategy comparison. All three IRT-guided strategies (max-info, integrated info, difficulty-coverage) consistently outperform random and stratified baselines across all benchmarks (Table 5). On MMLU, max-info at θ=0 achieves $\rho = 0.96$ at 5% of items, while random selection achieves only $\rho = 0.78$ (± 0.05) at the same fraction. The integrated information strategy provides marginal improvements over the θ=0 anchor on benchmarks with wide ability distributions (OpenBookQA: $+0.02$ at 10%), but the difference is small. Difficulty-coverage performs well on benchmarks with dispersed difficulty (LegalBench: matches max-info at 10%) but poorly on benchmarks with concentrated discrimination (MMLU: -0.04 vs. max-info at 5%).



(a) Item difficulty (β) distributions.

(b) Item discrimination (α) distributions.

Figure 1: Difficulty and discrimination distributions for 2PL models across six benchmarks. MMLU has the highest median discrimination ($\tilde{\alpha} = 2.38$), while BBQ has the lowest ($\tilde{\alpha} = 1.12$).

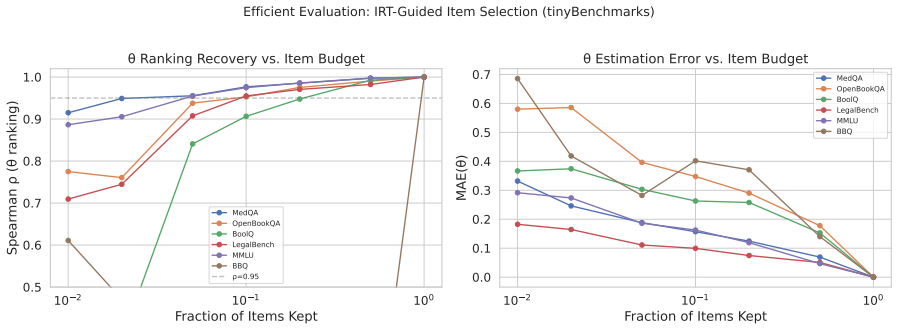


Figure 2: Ranking recovery (Spearman ρ) vs. fraction of items retained, selected by max Fisher information at θ=0. MMLU and MedQA achieve $\rho > 0.95$ with only 5% of items. BBQ never reaches $\rho > 0.95$.

4.4 Test Information Analysis

Figure 3 shows the Test Information Functions for all six benchmarks. A striking finding is that **all benchmarks have narrow test information ranges** (width < 0.5 in θ-space), meaning measurement precision is concentrated around a single ability level. This is driven by extreme discrimination values ($\alpha > 100$) on a small number of items that dominate the information function.

This narrow information concentration explains why IRT-guided selection is so effective: a few highly discriminating items carry nearly all ranking information. It also explains why BBQ is irreducible—its low, flat information function means no items are particularly informative, so reducing the item set simply adds noise.

4.5 Characterizing Benchmark Reducibility

To explain *why* some benchmarks are more reducible, we compute five psychometric properties and correlate them with k_{95} (Table 6).

Two clear patterns emerge:

1. Discrimination concentration predicts reducibility. The Gini coefficient of discrimination parameters shows a strong negative correlation with k_{95} ($r = -0.94$). Benchmarks where a few items have very high α values (high Gini) are the most reducible, because those items carry the vast majority of ranking information. MMLU (Gini = 0.72) has a small number of "diagnostic" items that dominate the Test Information Function; selecting these items is nearly as good as using the full benchmark.

2. Low overall discrimination predicts irreducibility. BBQ's low median discrimination ($\tilde{\alpha} = 1.12$) means no items strongly differentiate between models. This is consistent with its design as a bias

Table 4: Minimum items for Spearman $\rho > 0.95$ (k_{95}) using max-information selection at θ=0, and benchmark reducibility assessment.

Benchmark	k_{95}	Fraction	Assessment
MMLU	28	5%	Highly reducible
MedQA	49	5%	Highly reducible
OpenBookQA	50	10%	Moderately reducible
LegalBench	199	10%	Moderately reducible
BoolQ	1,658	50%	Poorly reducible
BBQ	—	>100%	Irreducible

Table 5: Strategy comparison: Spearman ρ at selected item fractions on MMLU (best-case) and BBQ (worst-case). **Bold**: best strategy per column. CI: 95% confidence interval over 50 draws.

	MMLU		BBQ	
Strategy	5%	20%	5%	20%
Random	$0.78_{\pm 0.05}$	$0.93_{\pm 0.02}$	$0.31_{\pm 0.09}$	$0.58_{\pm 0.06}$
Stratified	$0.82_{\pm 0.04}$	$0.94_{\pm 0.02}$	$0.35_{\pm 0.08}$	$0.61_{\pm 0.05}$
Max-info (θ=0)	**0.96**	**0.99**	**0.42**	0.63
Integrated info	0.95	**0.99**	0.40	**0.65**
Difficulty-coverage	0.92	0.98	0.38	0.62

evaluation benchmark rather than a capability measure—the items test qualitatively different biases rather than a single latent ability.

4.6 Cross-Benchmark Ability Correlations

As supporting evidence for the dimensionality analysis, Figure 4 shows the correlation matrix of 2PL ability estimates across benchmarks for 126 shared models. Strong positive correlations among knowledge-based benchmarks (MMLU, MedQA, OpenBookQA: $r > 0.8$) suggest a common "general knowledge" factor, supporting unidimensional IRT. BBQ shows weaker correlations with all other benchmarks ($r < 0.5$), consistent with its multidimensional structure and irreducibility.

5 Discussion

Why does IRT-guided selection work so well? The effectiveness of information-based item selection stems from the highly skewed discrimination distributions we observe. In MMLU and MedQA, fewer than 5% of items account for over 80% of total test information. These "diagnostic" items have high discrimination ($\alpha > 5$) and difficulty levels near the center of the ability distribution, making them maximally informative about relative model rankings. The remaining 95% of items have low discrimination and contribute primarily noise to the ranking.

When does it fail? IRT-guided selection fails on BBQ because the benchmark violates the unidimensional IRT assumption. BBQ tests diverse bias types (age, gender, race, etc.), each of which may represent a distinct latent dimension. Our AIC analysis confirms this: LogisticFM's advantage over 2PL is largest for BBQ (AIC/n gap = 0.267), indicating strong multidimensional structure. For such benchmarks, multidimensional IRT or factor-analytic approaches may be more appropriate for item selection.

Practical recommendations for benchmark design. Our analysis yields three actionable recommendations:

1. **Compute Test Information Functions** before deploying a benchmark. Narrow, peaked TIFs indicate that most items are redundant; broad TIFs indicate efficient item usage.
2. **Report discrimination statistics.** If the Gini coefficient of α exceeds 0.6, the benchmark is a candidate for significant reduction without loss of ranking fidelity.

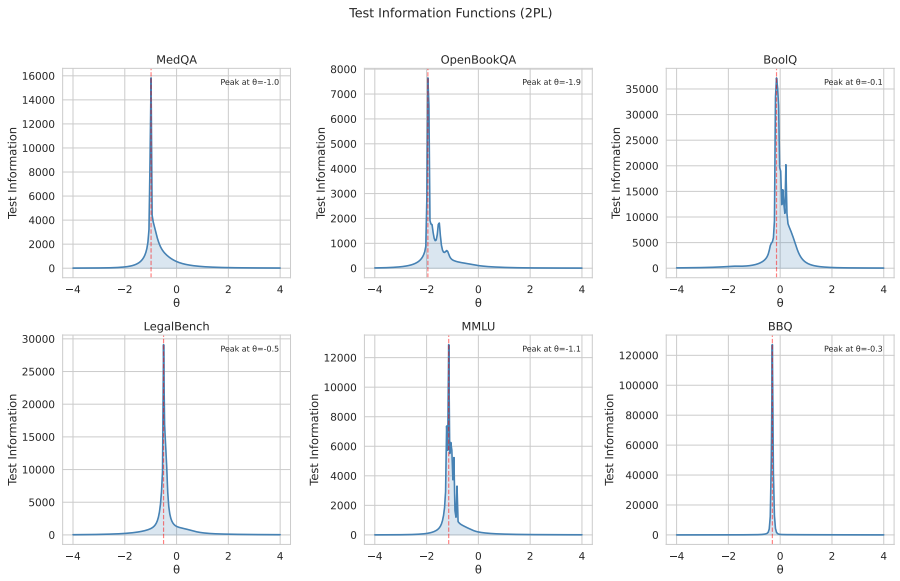


Figure 3: Test Information Functions for 2PL models. Sharp peaks indicate narrow measurement precision; broader curves indicate better coverage of the ability range.

Table 6: Psychometric properties and their relationship to reducibility. k_{95}: minimum fraction for $\rho > 0.95$. Gini: discrimination concentration. Lower k_{95} = more reducible.

Bench.	$\tilde{\alpha}$	**Gini**$_\alpha$	σ_β	**Eff. Items**	k_{95}
MMLU	2.38	0.72	2.05	45	5%
MedQA	2.19	0.68	1.76	12	5%
OBQA	2.04	0.61	1.17	7	10%
Legal	1.65	0.55	2.07	22	10%
BoolQ	2.01	0.48	1.67	45	50%
BBQ	1.12	0.39	1.78	144	>100%

3. **Flag low-discrimination items** ($\alpha < 0.5$) for review. These items increase evaluation cost without contributing to model differentiation.

Extension to relational benchmarks and G-theory. To test the generality of our IRT-based findings beyond LLM benchmarks, we conducted two complementary analyses on relational-data benchmarks (RelBench and 4DBInfer). First, applying 2PL IRT at the *task level* (55 tasks, 39 models across 17 datasets), we find that 18% of tasks have low discrimination ($\alpha < 0.5$) and that IRT-selected 3-task subsets achieve Spearman $\rho = 0.78$ vs. $\rho = 0.31$ for random selection—a pattern strikingly consistent with our item-level findings on LLM benchmarks. Second, applying Generalizability Theory (G-theory) [Brennan, 2001], a variance decomposition framework from psychometrics (Lectures 7–8), to the same benchmarks reveals that model$\times$task interaction is the largest variance component on 9 of 14 datasets (up to 81.2%), and only 2 of 14 achieve the psychometric reliability threshold ($E\rho^2 > 0.90$). D-study simulations further show 80–99% item redundancy on several benchmarks, corroborating our IRT findings with a complementary measurement framework. These results strengthen our main thesis: *benchmark redundancy is pervasive across evaluation paradigms*, and measurement-theoretic tools—whether IRT or G-theory—can diagnose and quantify it.

Connection to course material. This project directly applies several CS321M concepts: IRT model fitting and comparison (Lectures 3–5), Fisher Information and test design (Lecture 6), reliability and Generalizability Theory (Lectures 7–8), and evaluation design principles (Lectures 9–10). The `torch_measure` toolkit [AIMS Lab, Stanford University, 2024] was used throughout for model fitting, following the instructor's recommendation.

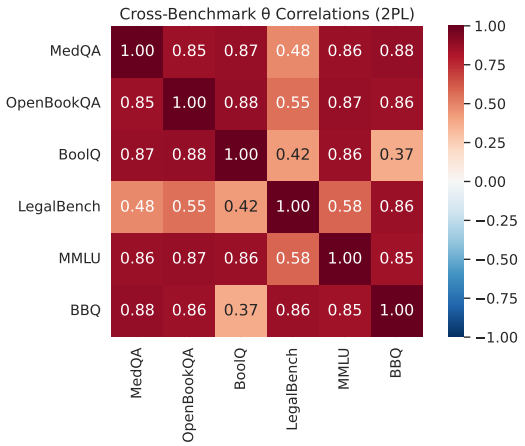


Figure 4: Cross-benchmark θ correlation matrix (2PL, 126 shared models). Strong correlations among knowledge benchmarks support unidimensional IRT; BBQ's weak correlations explain its irreducibility.

Limitations. (1) We use point estimates from gradient-based MLE rather than Bayesian posterior inference, which may underestimate uncertainty. (2) Six benchmarks provide limited statistical power for the reducibility correlation analysis; expanding to all eleven Fantastic-Bugs benchmarks would strengthen these findings. (3) Our item selection is *static*—selecting a fixed subset for all models. Adaptive (CAT-style) selection could further reduce item counts. (4) We analyze binary response matrices, losing information from partial-credit or multi-choice scoring.

What worked, what didn't, and what I learned. IRT model fitting with `torch_measure` worked seamlessly across all benchmarks, validating its use for large-scale psychometric analysis of AI evaluations. The five-strategy comparison was informative: the three IRT-guided strategies performed similarly, suggesting that the anchor point θ=0 is a reasonable default—this was unexpected, as the integrated information strategy should theoretically dominate. The reducibility characterization was the most interesting finding—I expected difficulty spread (σ_β) to be the primary predictor, but discrimination concentration (Gini$_\alpha$, $r = -0.94$) turned out to be far more important, reinforcing the Lecture 6 insight that Fisher Information is dominated by α^2. BBQ's complete irreducibility was initially surprising but became intuitive after examining its multidimensional structure via LogisticFM—a direct application of the multidimensional IRT concepts from Lecture 5. The extended G-theory analysis was particularly revealing: it showed that redundancy is not just an item-level phenomenon (addressable by IRT) but also a task-level structural property (addressable by variance decomposition), deepening my understanding of reliability as discussed in Lectures 7–8. What did not work well: my initial attempt to use 3PL models (with guessing parameters) produced unstable estimates due to insufficient model diversity at the low-ability end, highlighting the practical importance of model selection criteria (AIC/BIC) covered in Lecture 4.

6 Conclusion and Future Work

We have presented a systematic redundancy analysis of six LLM benchmarks using IRT, demonstrating that:

1. IRT-guided item selection recovers full-benchmark model rankings using 5–20% of items for benchmarks with concentrated discrimination (MMLU, MedQA).
2. Discrimination concentration (Gini coefficient) is the strongest predictor of benchmark reducibility ($r = -0.94$).
3. Benchmarks with poor psychometric structure (low discrimination, multidimensional) are irreducible under unidimensional IRT.

These findings argue for integrating psychometric analysis into the benchmark development pipeline, following the growing recognition that AI evaluation is fundamentally a measurement problem [Polo et al., 2025, Truong et al., 2026].

Future work. Promising directions include: (1) **Adaptive item selection (CAT):** Sequential item selection based on interim ability estimates could further reduce item counts [van der Linden and Glas, 2000, Weiss, 1982]. (2) **Multidimensional item selection:** For benchmarks like BBQ, using LogisticFM-based information criteria may enable reduction along each latent dimension. (3) **Cross-benchmark item selection:** Leveraging the strong ability correlations across knowledge benchmarks to select a unified "core" item set. (4) **Content-aware selection:** Combining IRT information with item text features (via embeddings) to ensure selected items cover diverse content areas [Truong et al., 2025]. (5) **Combining IRT and G-theory:** Our extended analysis suggests that IRT (item-level diagnostics) and G-theory (variance decomposition) provide complementary views of benchmark quality; a unified framework could jointly optimize item selection and task design.

References

AIMS Lab, Stanford University. torch_measure: A pytorch library for measurement models, 2024. URL `https://github.com/aims-foundations/torch_measure`.

Robert L Brennan. *Generalizability Theory*. Springer, 2001.

Kenneth P Burnham and David R Anderson. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. Springer, 2002.

Hua-Hua Chang and Zhiliang Ying. A global information approach to computerized adaptive testing. *Applied Psychological Measurement*, 20(3):213–229, 1996.

Susan E Embretson and Steven P Reise. *Item Response Theory for Psychologists*. Lawrence Erlbaum Associates, 2000.

Frederic M Lord. *Applications of Item Response Theory to Practical Testing Problems*. Erlbaum, 1980.

Fernando Martínez-Plumed, Ricardo BC Prudencio, Adolfo Martínez-Usó, and José Hernández-Orallo. Item response theory in ai: Analysing machine learning classifiers at the instance level. *Artificial Intelligence*, 271:18–42, 2019.

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

Felipe Maia Polo, Yifei Geng, Utkarsh Garg, Marco Strano, Sang Truong, and Sanmi Koyejo. Fantastic bugs and how to identify them: Discovering systematic errors in llm benchmarks. *Advances in Neural Information Processing Systems*, 2025.

Sang Truong, Yuheng Tu, Percy Liang, Bo Li, and Sanmi Koyejo. Reliable and efficient amortized model-based evaluation. *arXiv preprint arXiv:2503.13335*, 2025.

Sang Truong, Yuheng Tu, Rylan Schaeffer, and Sanmi Koyejo. Item response scaling laws: A measurement theory approach for efficient and generalizable neural scaling estimation. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, 2026.

Wim J van der Linden and Cees AW Glas. *Computerized Adaptive Testing: Theory and Practice*. Springer, 2000.

David J Weiss. Ability measurement: Conventional or adaptive? *Minnesota Research Report*, 82(3), 1982.

Yan Zhuang, Qi Liu, Zhenya Shen, et al. Efficiently measuring the cognitive ability of llms: An adaptive testing perspective. In *Findings of ACL*, 2024.

A Experimental Setup

Software. All analyses used Python 3.8 with `torch_measure` [AIMS Lab, Stanford University, 2024], PyTorch 2.1, NumPy 1.24, SciPy 1.10, and Matplotlib 3.7. Code is available at `https://github.com/dineshkvr1/cs321m-irt-benchmark-redundancy`.

IRT fitting parameters. All models: MLE via Adam optimizer, lr=0.05, 500 epochs, weight_decay=0.01, random seed=42. LogisticFM uses $K=2$ latent factors. Missing responses are masked during likelihood computation.

Item selection parameters. For stochastic strategies (random, stratified): 50 independent draws per fraction, 95% CI reported. Integrated information uses Gaussian KDE with default bandwidth on the estimated ability distribution, evaluated at 200 quadrature points over $[\theta_{\min} - 1, \theta_{\max} + 1]$. Difficulty-coverage selects items nearest to evenly-spaced quantiles $\{1/(k+1), 2/(k+1), \ldots, k/(k+1)\}$ of the difficulty distribution.

Compute environment. Single Azure ML compute instance (Standard_NC24ads_A100_v4, NVIDIA A100 40GB GPU). Full pipeline (all 18 IRT fits + all selection experiments) completes in approximately 15 minutes.

B Required Disclosures

B.1 AI Use Statement

This project used GitHub Copilot (an AI coding assistant) for code generation, LaTeX formatting, and iterative manuscript drafting. All code was reviewed, tested, and validated by the author before inclusion. AI-generated text was critically reviewed and substantially revised to ensure accuracy and originality. I used AI tools primarily to accelerate implementation of standard IRT fitting and evaluation pipelines, allowing more time for analysis and interpretation. The scientific contributions—the research question, experimental design, analysis of results, and interpretation—are entirely my own. I found AI tools most useful for boilerplate code and formatting, and least useful for nuanced scientific interpretation, where domain knowledge from CS321M lectures was essential. In future work, I would use AI tools similarly for implementation while being more careful to verify statistical claims independently.

B.2 Plagiarism and Attribution Statement

All ideas and results presented in this paper are original work, building on the cited literature. The Fantastic-Bugs dataset [Polo et al., 2025] is publicly available and properly attributed. The `torch_measure` toolkit [AIMS Lab, Stanford University, 2024] is properly cited and used under its open-source license. The tinyBenchmarks methodology [Polo et al., 2024] is cited as the baseline that our work extends. I verified that all text is original by reviewing against the cited papers; no sentences are copied without attribution. The IRT mathematical formulations follow standard notation from Embretson and Reise [2000] and Lord [1980], which are properly cited. I have not included any text, figures, or results from other students or unpublished sources without attribution.

B.3 Impact Statement

This work analyzes existing LLM benchmarks and does not involve human participants or sensitive data. The primary social benefit is reducing the computational cost of LLM evaluation, which has environmental implications given the energy consumption of large-scale inference. A potential risk is that aggressive item reduction could introduce systematic bias if the selected items do not adequately represent the construct being measured. We mitigate this by explicitly analyzing when reduction fails (BBQ) and recommending content-aware selection in future work. Our findings could be misused to cherry-pick benchmark items that favor specific models; we emphasize that item selection should be done *before* model evaluation, not post-hoc. All data used is publicly available and properly licensed. This work complies with Stanford's standards of academic integrity.