ADAPTIVE TESTING FOR LLM EVALUATION: A PSYCHOMETRIC ALTERNATIVE TO STATIC BENCHMARKS

Peiyu Li*, **Xiuxiu Tang***,
Si Chen, **Ying Cheng**, **Ronald Metoyer**, **Ting Hua**, **Nitesh V. Chawla**
University of Notre Dame
Notre Dame, IN 46556 USA
{pli9, xtang8, schen34, ycheng4, rmetoyer, thua, nchawla}@nd.edu

November 10, 2025

ABSTRACT

Large language model evaluation requires thousands of benchmark items, making evaluations expensive and slow. Existing methods compute average accuracy across fixed item sets, treating all items equally despite varying quality and informativeness. We present ATLAS an adaptive testing framework using Item Response Theory (IRT) to estimate model ability through Fisher information-guided item selection. Our analysis of five major benchmarks reveals that 3-6% of items exhibit negative discrimination, indicating annotation errors that corrupt static evaluation. ATLAS achieves 90% item reduction while maintaining measurement precision: on HellaSwag (5,608 items), we match full-benchmark estimates using only 42 items with 0.154 MAE. Our framework maintains item exposure rates below 10% and test overlap at 16-27%, compared to static benchmarks where every model sees all items (100% exposure). Among 4,000+ tested models, IRT ranks differ from accuracy ranks: models with the same accuracy get different IRT scores, and 23-31% of all models shift by more than 10 rank positions. Code and calibrated item banks are available at https://github.com/Peiyu-Georgia-Li/ATLAS.git.

1 Introduction

Large language model (LLM) evaluation relies on benchmarks with tens of thousands of test items. These benchmarks impose high computational costs and evaluation cycles spanning days or weeks. Despite benchmark growth—some proposals exceed 100,000 items—evaluation practice remains static: models receive average accuracy scores across fixed item sets. This approach ignores statistical information in large datasets and raises questions about efficiency and validity.

Current evaluation faces three fundamental limitations. First, average scores hide meaningful differences between models with distinct error patterns, particularly for lower-performing models where small ability differences are obscured by measurement noise. Second, static benchmarks create vulnerability to data contamination as items leak into pretraining corpora, enabling high scores through memorization rather than genuine capability. Third, evaluating complete benchmarks is inefficient, treating poorly discriminative items as equally informative as high-quality questions.

To address these limitations, we propose ATLAS (Adaptive Testing for LLM Ability Scoring), an adaptive evaluation framework based on computerized adaptive testing (CAT; [1, 2, 3]). ATLAS first calibrates benchmark items using three-parameter logistic (3PL) IRT models to estimate item difficulty, discrimination, and guessing parameters [4, 5]. Then, rather than administering fixed item sets, ATLAS dynamically selects items with maximum Fisher information for each model's current estimated ability, terminating when precision thresholds are reached. This approach directly addresses all three limitations: Fisher information-guided selection provides precise ability estimates that distinguish

*These authors contributed equally.

models with identical accuracy, dynamic item selection reduces contamination through diverse test forms, and adaptive termination achieves reliable evaluation with significantly fewer items.

We validate our framework across five major benchmarks (WinoGrande, TruthfulQA, HellaSwag, GSM8K, ARC). ATLAS achieves competitive or superior performance to static baselines: on TruthfulQA, it attains the lowest MAE (0.067) using only 51 items; on HellaSwag, it matches MetaBench accuracy (0.154 MAE) while requiring 3× fewer items (39 vs 93). Across benchmarks, ATLAS requires 30-78 items on average compared to thousands in full evaluation, while IRT-based estimates reveal systematic rank reordering that exposes limitations of accuracy-based evaluation.

Our contributions are: (1) We demonstrate fundamental limitations of average-score evaluation and establish the benefits of psychometric ability estimates for LLM evaluation. (2) We develop ATLAS, the first large-scale adaptive testing framework for LLMs, achieving 90% item reduction while maintaining measurement precision and natural contamination resistance through $< 10\%$ item exposure rates. (3) We provide comprehensive psychometric analysis of five major benchmarks, revealing that IRT-based estimates cause systematic rank reordering ($23 - 31\%$ of models shift more than 10 positions) and identifying that $3 - 6\%$ of items exhibit negative discrimination due to annotation errors, with calibrated item banks released for reproducible evaluation.

2 Related Work

2.1 Background

Large-scale benchmarks such as WinoGrande, TruthfulQA, HellaSwag, GSM8K and ARC have become standard for LLM evaluation but impose significant computational costs. These static approaches suffer from benchmark saturation, data contamination, and item redundancy, raising concerns about their ability to reflect genuine model capabilities.

Item Response Theory (IRT) has recently been applied to LLM evaluation [6, 7]. It provides item parameters such as difficulty, discrimination, and guessing, as well as latent ability estimates θ for models. However, existing IRT applications remain largely static: tinyBenchmarks [8] uses clustering for item selection but doesn't guarantee informativeness for θ estimation, while MetaBench [9] requires computationally expensive iterations to identify stable subsets. These approaches often lack proper psychometric validation and emphasize predictive accuracy over model fit diagnostics, which makes it difficult to ensure that ability estimates are valid, interpretable, and comparable across models.

Beyond these limitations of existing IRT applications, many evaluations continue to rely on average scores. Average scores tend to mask meaningful model differences and are often affected by form-dependence, nonlinear scaling, equal weighting of uninformative items, and contamination sensitivity (see Appendix A for detailed analysis). In contrast, IRT-based ability estimates (θ) provide form-invariant, uncertainty-aware alternatives that adjust for item difficulty and discrimination.

2.2 Adaptive Testing

Computerized adaptive testing (CAT) is a psychometric framework for efficiently measuring ability by selecting items that maximize Fisher information. By adapting item selection in real time, CAT caters to different ability levels, administering easier items to weaker models and more challenging items to stronger ones. CAT can operate with a fixed test length or with a variable length, terminating once a predefined precision threshold is reached [3].

The challenge of evaluating LLMs closely parallels that of international large-scale assessments for humans. Our dataset includes more than 4,000 models, ranging from simple pattern-matching systems with very low ability estimates ($\theta < -2$) to advanced models such as GPT-4 with high ability estimates ($\theta > 2$) . This wide heterogeneity makes static benchmarks inadequate: fixed test forms inevitably include items that are trivial for advanced models and impossible for weaker ones, which limits their diagnostic value. Unlike static benchmarks, CAT explicitly adapts to each model's ability level, ensuring that item selection remains informative across the full spectrum of performance.

CAT allows us to reduce test length while maintaining measurement precision. Modern extensions such as Multi-Stage Testing and the incorporation of process data [10, 11, 12] provide further gains in efficiency and validity. Prior attempts to explore adaptive evaluation for LLMs have been either limited in scope [13] or primarily conceptual [14]. Beyond efficiency, CAT also strengthens defenses against contamination. Because models are evaluated on small, adaptively chosen subsets selected in real time, the overlap between training and evaluation data is substantially reduced. Furthermore, CAT protocols can be refreshed with new items while preserving comparability of ability estimates θ. A detailed comparison of IRT-based approaches is provided in Appendix B.

3 Methodology

We introduce a novel adaptive testing framework that transforms LLM evaluation from static benchmarking to dynamic ability estimation. Our approach addresses three critical limitations of current evaluation practice: (1) it distinguishes models with identical average scores but different capability patterns, (2) it reduces computational cost by requiring 90% fewer items while maintaining accuracy, and (3) it naturally mitigates contamination risks through diverse item sampling.

This section presents our framework in four stages: problem formulation (Section 3.1), data construction with psychometric filtering (Section 3.2), item bank calibration using IRT models (Section 3.3), and adaptive testing with randomesque selection (Section 3.4).

3.1 Problem Formulation and Setup

We formulate LLM evaluation as a psychometric measurement problem. Let $\mathcal{I}$ denote the set of benchmark items and $\mathcal{L}$ the set of language models. For each model $\ell \in \mathcal{L}$ and item $i \in \mathcal{I}$, we observe a binary response $Y_{i,\ell} \in \{0, 1\}$, where 1 indicates correct and 0 incorrect. These responses form the item-response matrix $\{Y_{i,\ell}\}_{i\in\mathcal{I},\ell\in\mathcal{L}}$.

Unlike traditional approaches that rely solely on accuracy scores, our objective is to estimate the latent ability θ_ℓ of each model based on its response pattern $\{Y_{i,\ell}\}_{i\in\mathcal{I}}$, while simultaneously calibrating item-level parameters: discrimination a_i, difficulty b_i, and guessing c_i. This approach enables fine-grained model comparison even when models achieve identical accuracy, as θ_ℓ accounts for the varying informativeness of different items.

3.2 Discriminative Item Filtering

We construct the item-response matrix using data from the HuggingFace Open LLM Leaderboard. The item pool $\mathcal{I}$ spans five benchmarks: ARC, GSM8K, HellaSwag, TruthfulQA, and WinoGrande. To ensure data quality for IRT calibration, we apply two levels of filtering: removing unsuitable models and eliminating non-informative items.

Model Selection. We retain only models $\mathcal{L}$ with complete responses across all items to ensure unbiased ability estimates. Models with extreme scores (below 0.1st percentile) are excluded to prevent parameter estimation instability, as IRT's sigmoidal functions become under-constrained at the boundaries.

Item Filtering. We apply two complementary filters to retain only discriminative items:

- **Low-variance removal:** Items with response standard deviation $< 1\%$ or mean accuracy $> 95\%$ are discarded, as they provide litter information for differentiating between models.
- **Discrimination filtering:** We compute the point-biserial correlation $r_{pb}(i)$ between each item's response vector $\{Y_{i,\ell}\}_{\ell\in\mathcal{L}}$ and the models' total scores $T_\ell = \sum_{j\in\mathcal{I}} Y_{j,\ell}$ (see Appendix E for details). Items with $r_{pb}(i) < 0.1$ are removed as non-diagnostic.

This filtering process yields a refined response matrix that supports stable and reliable IRT calibration (see Table 5 in Appendix E for detailed results).

3.3 Scalable IRT Calibration

The calibration stage estimates item parameters (a_i, b_i, c_i) and computes reference ability estimates $\hat{\theta}_\ell^{\text{whole}}$ for each LLM ℓ for validation. To model the probability of a correct response, we adopt the three-parameter logistic (3PL) IRT model [4, 1]:

$$p_i(\theta_\ell) = c_i + \frac{1 - c_i}{1 + \exp(-a_i(\theta_\ell - b_i))}. \tag{1}$$

Here, a_i is the discrimination parameter, which determines how sharply item i differentiates between stronger and weaker models. b_i is the difficulty parameter, specifying the ability level at which a model has a 50% chance (beyond guessing) of answering item i correctly. c_i is the guessing parameter, setting the lower bound on the probability of success due to random guessing. These parameters enable Fisher information-based prioritization of items in our adaptive framework, distinguishing high-quality items from those with low discriminative power.

Common-Person Calibration at Scale. We develop a partition-based calibration strategy that takes advantage of the unique characteristics of LLM benchmarking. Unlike traditional testing where only subsets of items are administered

to each examinee, all LLM models respond to all items, enabling a novel application of common-person linking. We partition items into K non-overlapping subsets $\mathcal{I}_k$ (each with $|\mathcal{I}_k| \geq 100$ items), calibrate each subset independently, then apply mean-sigma transformations [15] using the model population $\mathcal{L}$ as linking anchors. This approach reduces computational complexity from $O(|\mathcal{I}|^3)$ to $O(K \cdot \max_k |\mathcal{I}_k|^3)$ while maintaining calibration accuracy through the redundancy of having all models as linkers [16].

Heterogeneity-Aware Ability Estimation. The extreme heterogeneity of LLM populations, spanning from random baselines ($\theta \approx -3$) to near-perfect models ($\theta \approx 3$), breaks standard estimation methods. Standard IRT practice relies on Expected A Posteriori (EAP) estimation [17] for ability estimation. However, EAP produces undefined values for all-correct or all-incorrect response patterns, affecting 12% of our model population spanning from random baselines to near-perfect systems. To address this LLM-specific challenge, we employ Weighted Likelihood Estimation (WLE) [18] with bias correction term $\frac{J(\theta)}{2I(\theta)}$ where $J(\theta) = \sum_i \frac{\partial I_i(\theta)}{\partial \theta}$ [18]. This choice enables finite estimates across the full ability spectrum while maintaining consistency properties [19], crucial for establishing reliable validation baselines in the extreme heterogeneity of LLM evaluation.

Multi-Subset Model Fit Validation. Unlike prior IRT applications to LLM evaluation [8, 9] that omit model validation, we conduct rigorous psychometric diagnostics to ensure calibration quality. We report the M_2 statistic [20] with RMSEA indices, establishing that our 3PL models achieve good fit (RMSEA < 0.05) across all benchmarks, as summarized in Appendix D. Our partition-based calibration strategy, which divides the full item bank into non-overlapping subsets (each containing ≥ 100 items for statistical stability), enables both computational feasibility and robust validation. Since the same set of models $\mathcal{L}$ acts as common persons across all partitions, diagnostic statistics reflect global calibration quality rather than partition-specific artifacts. This multi-subset linking design ensures that model fit metrics capture systematic patterns across the entire item bank, not just localized subsets. This validation is crucial for reliable adaptive testing, as misfitting items would compromise Fisher information calculations and degrade selection accuracy.

3.4 Adaptive Testing with Information Selection

Our proposed ATLAS dynamically selects the most informative items for each model, dramatically reducing the number of items needed while maintaining accuracy. Algorithm 1 presents the complete procedure. The algorithm includes several key design choices tailored to LLM evaluation:

Algorithm 1 Adaptive Testing for Model ℓ

Initialize: $\hat{\theta}_0 \leftarrow 0$, test record $R_\ell \leftarrow \emptyset$, $t \leftarrow 0$
while $t < \text{max\_items}$ and not converged **do**
  $t \leftarrow t + 1$
  if $t = 1$ **then**
    Select item i_t with $|b_{i_t} - \hat{\theta}_0|$ minimized
  else
    Compute Fisher information $I_i(\hat{\theta}_{t-1})$ for all unadministered items
    Select i_t randomly from top-5 most informative items
  end if
  Administer item i_t to model ℓ, observe response $Y_{i_t,\ell}$
  Update record: $R_\ell \leftarrow R_\ell \cup \{(i_t, Y_{i_t,\ell})\}$
  Update ability: $\hat{\theta}_t \leftarrow \text{EAP}(R_\ell)$
  Compute standard error: $\text{SE}(\hat{\theta}_t) \leftarrow 1/\sqrt{\sum_{j \in R_\ell} I_j(\hat{\theta}_t)}$
  if $t \geq \text{min\_items}$ and $\text{SE}(\hat{\theta}_t) \leq \tau$ **then**
    break ▷ Convergence achieved
  end if
end while
return $\hat{\theta}_\ell \leftarrow \hat{\theta}_t$, $\text{SE}(\hat{\theta}_\ell)$, R_ℓ

Initialization and Bounds. We initialize the ability estimate at $\hat{\theta}_0 = 0$. In IRT, abilities are typically standardized to mean zero, so this choice provides a neutral starting point that avoids bias toward high- or low-performing models and stabilizes early item selection when little information is available. We enforce minimum (30) and maximum (500) item limits. The minimum ensures stable estimation for models at performance extremes, while the maximum constrains computational cost and yields approximately 90% reduction in test length relative to full benchmarks.

Randomesque Item Selection. Rather than deterministically selecting the single most informative item, we randomly sample from the top-5 candidates ranked by Fisher information:

$$I_i(\theta) = a_i^2 \cdot p_i(\theta) \cdot [1 - p_i(\theta)]. \tag{2}$$

This randomesque strategy [21] prevents over-reliance on specific item types while still keeping high information, which is important for models with specialized capabilities.

Sequential Ability Updates. After each item administration, we update the ability estimate using Expected A Posteriori (EAP) estimation [17]:

$$\hat{\theta}_t = \mathbb{E}[\theta | R_\ell] = \int \theta \cdot p(\theta | R_\ell)\, d\theta. \tag{3}$$

EAP provides numerically stable updates with sparse early responses and incorporates prior knowledge about ability distributions. In contrast, WLE tends to become unstable when response patterns are extreme, a situation common in the early stages of adaptive testing.

Precision-Based Stopping. The procedure terminates when either the maximum number of items is reached or the standard error of the ability estimate, $\mathrm{SE}(\hat{\theta}_\ell)$, falls below a predefined threshold τ, provided that at least 30 items have been administered. This stopping rule achieves efficiency and ensures consistent precision across the ability spectrum.

Output and Validation. For each model ℓ, the algorithm produces: (1) the administered item sequence and responses R_ℓ, (2) the ability estimate trajectory $\{\hat{\theta}_t\}$ with associated standard errors, and (3) the final estimate $\hat{\theta}_\ell$. We validate these adaptive estimates against whole-bank references $\hat{\theta}_\ell^{\mathrm{whole}}$ to confirm that our dramatic reduction in items does not compromise measurement accuracy.

4 Experiments

We evaluate the proposed ATLAS framework across five benchmarks, comparing its efficiency and accuracy against static baselines while analyzing exposure rates, overlap patterns, and the advantages of IRT-based ability estimation over raw accuracy scores.

4.1 Experimental Setup and Metrics

We evaluate ATLAS across five diverse benchmarks covering different cognitive domains: WinoGrande (commonsense reasoning), TruthfulQA (factual consistency), HellaSwag (procedural inference), GSM8K (mathematical reasoning), and ARC (scientific question answering). All experiments use calibrated item banks from Section 3.3.

We compare against four static baselines that do not adapt to individual models: (1) **Random sampling** of 100 items uniformly from the full bank, (2) **TinyBenchmarks** [8] using predetermined subsets selected via clustering without Fisher information optimization, (3) **MetaBench-Primary** and (4) **MetaBench-Secondary** [9] using curated splits that require computationally expensive iterations to identify stable subsets. Unlike these static approaches, ATLAS uses three precision thresholds ($\mathrm{SE}(\hat{\theta}) \leq 0.1, 0.2, 0.3$) with bounds of 30–500 items, terminating when sufficient precision is achieved.

We evaluate using five metrics: (1) **Mean Absolute Error (MAE)** between ATLAS estimates $\hat{\theta}_\ell$ and full-bank references $\hat{\theta}_\ell^{\mathrm{whole}}$ to assess accuracy, (2) **Average Number of Items Administered** to measure efficiency, (3) **Item Exposure Rate** quantifying how frequently each item appears across evaluations, (4) **Test Overlap Rate** $\bar{Q}$ measuring the average proportion of shared items between any two model evaluations, and (5) **Runtime** per session to assess computational efficiency. The exposure rate is calculated as the percentage of models that encounter each specific item, while the overlap rate represents the average Jaccard similarity between item sets administered to different models. Additional experimental configurations and psychometric interpretations are provided in Appendix C.

4.2 Performance and Reliability Analysis

Table 1 reveals ATLAS's consistent efficiency advantages across all benchmarks. On HellaSwag, ATLAS achieves the best accuracy (0.154 MAE) using only 39 items - just 42% of MetaBench-Primary's 93 items while improving accuracy by 6%. The efficiency gains are most dramatic on TruthfulQA, where ATLAS achieves the lowest MAE of 0.067 using only 33% of MetaBench-Primary's items (51 vs 154) and outperforms both MetaBench variants. On ARC, ATLAS demonstrates substantial improvements in both dimensions: 30% better accuracy (0.099 vs 0.142 MAE) using just 53% of the items (77 vs 145). Even on GSM8K where MetaBench-Primary achieves marginally lower error,

Table 1: Baseline and adaptive evaluation results (MAE and Avg. Item). Lower is better ↓.

Method	WinoGrande		TruthfulQA		HellaSwag		GSM8K		ARC	
	MAE	Items	MAE	Items	MAE	Items	MAE	Items	MAE	Items
Random	0.176	100	0.105	100	0.231	100	0.164	100	0.147	100
TinyBenchmarks	0.247	100	0.161	100	0.299	100	0.173	100	0.165	100
MetaBench-P	**0.148**	133	0.081	154	0.164	93	**0.154**	100	0.142	145
MetaBench-S	0.210	106	0.074	136	0.238	58	0.162	100	0.133	100
$\text{ATLAS}_{0.1}$	0.149	78	**0.067**	51	**0.154**	39	0.159	73	**0.099**	77
$\text{ATLAS}_{0.2}$	0.181	39	0.074	**30**	0.159	**30**	0.181	39	0.126	36
$\text{ATLAS}_{0.3}$	0.183	**32**	0.072	30	0.161	30	0.186	**32**	0.128	**30**

Table 2: Adaptive evaluation efficiency and diversity. ATLAS maintains low item exposure rates ($< 10\%$) and moderate test overlap ($16 - 27\%$), providing natural contamination resistance while achieving fast selection times (< 1 minute per model). Lower values indicate better performance for all metrics.

Benchmark (Item #)	Method	Test Overlap ↓ Rate (%)	Avg. Item ↓ Exposure Rate (%)	Avg. Selection ↓ Time (s)
WinoGrande (1046)	$\text{ATLAS}_{SE\leq 0.1}$	18.85	9.28	40.99
	$\text{ATLAS}_{SE\leq 0.2}$	**17.32**	5.07	19.92
	$\text{ATLAS}_{SE\leq 0.3}$	20.28	**4.24**	**16.74**
TruthfulQA (628)	$\text{ATLAS}_{SE\leq 0.1}$	**18.93**	**8.17**	15.97
	$\text{ATLAS}_{SE\leq 0.2}$	21.69	10.56	**9.37**
	$\text{ATLAS}_{SE\leq 0.3}$	21.67	10.68	9.72
HellaSwag (5608)	$\text{ATLAS}_{SE\leq 0.1}$	**15.76**	**3.90**	75.52
	$\text{ATLAS}_{SE\leq 0.2}$	19.28	6.23	**56.93**
	$\text{ATLAS}_{SE\leq 0.3}$	19.25	6.85	57.06
GSM8K (1307)	$\text{ATLAS}_{SE\leq 0.1}$	22.43	8.32	45.69
	$\text{ATLAS}_{SE\leq 0.2}$	**22.16**	**5.63**	24.08
	$\text{ATLAS}_{SE\leq 0.3}$	26.89	8.24	**19.06**
ARC (842)	$\text{ATLAS}_{SE\leq 0.1}$	19.50	10.08	30.99
	$\text{ATLAS}_{SE\leq 0.2}$	**19.25**	**5.42**	13.98
	$\text{ATLAS}_{SE\leq 0.3}$	21.91	9.77	**11.82**

ATLAS maintains competitive performance (0.159 vs 0.154 MAE) while requiring 27% fewer items. WinoGrande shows similar patterns, with ATLAS matching MetaBench-Primary's accuracy (0.149 vs 0.148 MAE) using only 59% of the items.

Figure 1 compares ability estimates from subsets against whole-bank references. ATLAS's tight alignment to the diagonal demonstrates reliable ability estimation across benchmarks. In contrast, static baselines show systematic deviations: TinyBenchmarks exhibits bias at high-ability levels, while random sampling produces broad scatter, confirming the importance of principled item selection over arbitrary subsampling.

4.3 Contamination Resistance and Efficiency Benefits

Table 2 demonstrates ATLAS's contamination resistance and efficiency advantages across multiple dimensions. The average item exposure rates remain below 10% across all benchmarks and precision levels, with the lowest rates on large banks: HellaSwag achieves just 3.9% exposure with SE ≤ 0.1, meaning each item appears in fewer than 4% of model evaluations. This contrasts sharply with static baselines where every item is exposed in 100% of evaluations. Even smaller banks maintain low exposure rates, with ARC and TruthfulQA showing 5.4-10.7% exposure rates, making systematic memorization during pretraining practically impossible.

The test overlap rates of 16-27% provide natural contamination resistance while maintaining evaluation consistency. On HellaSwag, the 15.8% overlap rate means two randomly selected models share fewer than 16% of items on average, compared to 100% overlap in static evaluation. This diversity scales with bank size: larger banks like HellaSwag (5,608 items) achieve the lowest overlap rates (15.8-19.3%), while smaller banks like GSM8K show higher but still

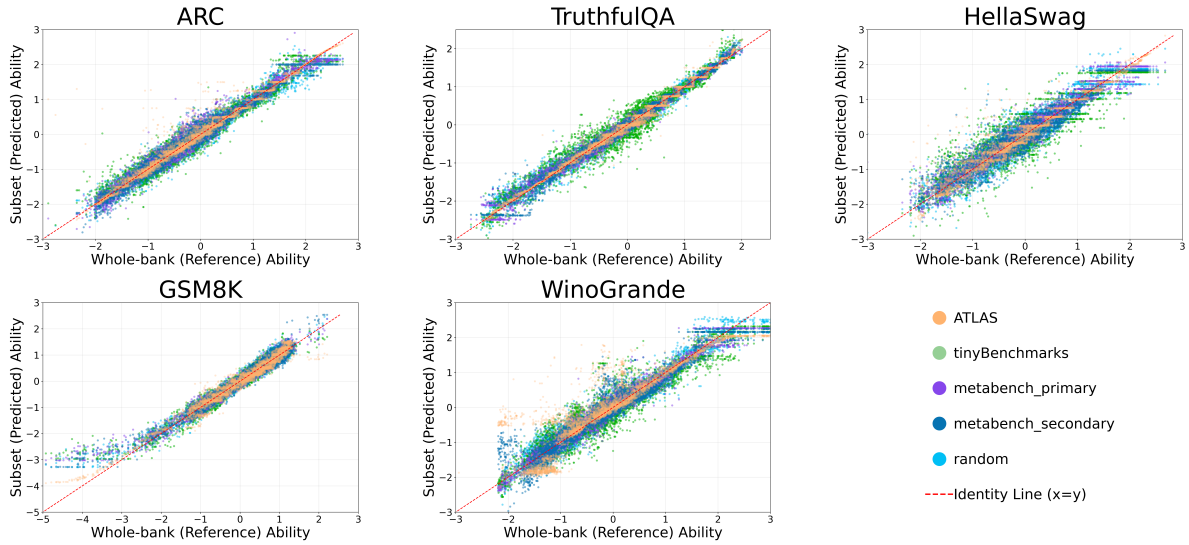

Figure 1: Comparison of subset (predicted) ability estimates against whole-bank (reference) abilities across five benchmarks. Points along the diagonal indicate perfect agreement. ATLAS maintains the closest alignment overall, particularly on TruthfulQA, ARC, HellaSwag and in the low-ability regime of GSM8K, while static baselines such as TinyBenchmarks and random show greater variance and systematic deviation.

protective overlap rates (22.2-26.9%). The precision threshold affects this trade-off: stricter thresholds (SE ≤ 0.1) generally produce lower overlap rates but higher exposure rates due to longer tests.

Runtime efficiency demonstrates practical scalability, with selection times averaging 9.4-75.5 seconds per model across benchmarks. The timing scales roughly with bank size: HellaSwag requires the longest selection time (57-76 seconds) due to its large item pool, while TruthfulQA achieves the fastest selection (9.4-16.0 seconds). Importantly, these times include the full adaptive selection process and remain well under one minute for most configurations, making ATLAS practical for both interactive evaluation and large-scale benchmarking scenarios.

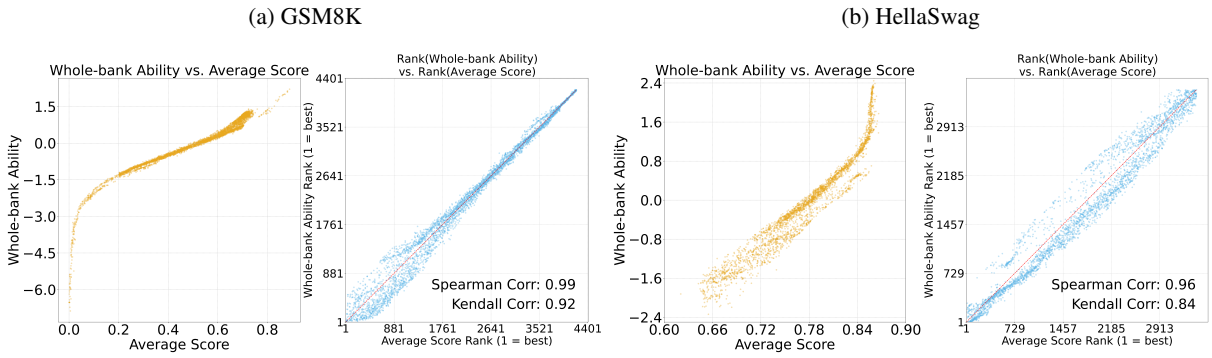

Figure 2: Comparison of IRT ability estimates $\hat{\theta}_\ell^{\text{whole}}$ with raw accuracy. Left: Ability vs accuracy reveals strong correlation but critical differences at performance extremes where accuracy collapses. Right: Rank comparison shows systematic reordering, with 23% (GSM8K) and 31% (HellaSwag) of models shifting > 10 positions. IRT separates models with identical accuracies by accounting for which items they solve correctly.

4.4 Distinguishing Low- and High-Performing Models

Figure 2 demonstrates IRT's key advantage: separating models with similar accuracy scores through ability estimates. Despite strong correlations (0.99 for GSM8K, 0.96 for HellaSwag), systematic differences emerge at performance extremes. In low-performing regimes, accuracy collapses into narrow bands (0.10-0.15) while IRT spans $\theta \approx -3$ to -1, distinguishing models that solve easy versus challenging items. In high-performing regimes, ceiling effects compress accuracy differences, but IRT maintains discrimination across $\theta \approx 1.5$ to 2.5. Similar patterns are observed

across TruthfulQA, WinoGrande, and ARC benchmarks, with additional score-versus-theta comparisons provided in Appendix G.

The right panels show systematic rank reordering: 23-31% of models shift > 10 positions when ranked by IRT versus accuracy. Models performing well on hard items receive higher IRT ranks despite moderate accuracy, while those succeeding on easy items are appropriately downweighted. This enhanced discrimination provides more reliable model comparisons, especially critical in saturated performance regions where accuracy-based evaluation fails. Similar patterns of IRT superiority in distinguishing models are observed across TruthfulQA, WinoGrande, and ARC benchmarks, with additional analysis provided in Appendix G.

4.5 Items Are Not Equally Informative

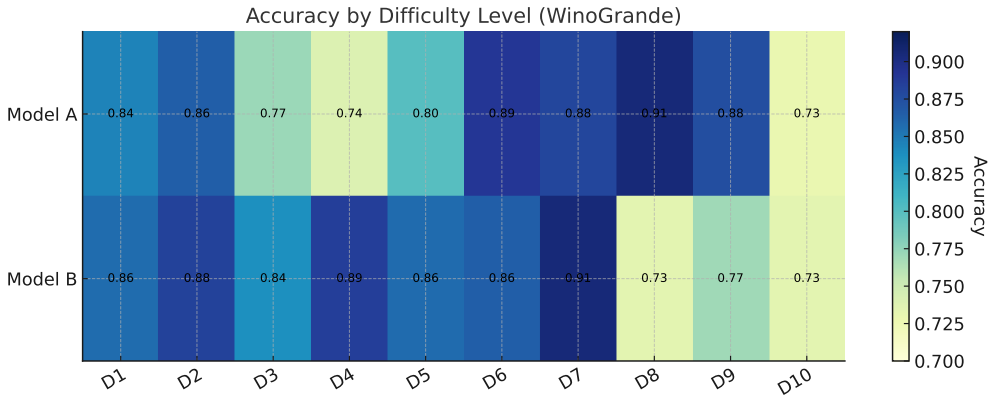

Figure 3: Two models with identical accuracy (0.833) on WinoGrande receive different ability estimates ($\hat{\theta}_A = 1.2$ vs $\hat{\theta}_B = 0.6$). Model A succeeds on harder items (darker cells on right), while Model B answers easier items (darker cells on left). IRT captures these item difficulty patterns that raw accuracy cannot.

Figure 3 demonstrates IRT's key advantage: models with identical accuracy (0.833) receive different ability estimates ($\hat{\theta}_A = 1.2$ vs $\hat{\theta}_B = 0.6$) based on which items they solve correctly. Under the 3PL IRT model, items differ in discrimination a_i, difficulty b_i, and guessing c_i parameters. Model A succeeds on high-difficulty items ($b_i > 0.5$) with strong discrimination ($a_i > 1.5$), yielding 2.3× more Fisher information than Model B, which mainly answers easier, less discriminative items ($b_i < -0.5$, $a_i < 0.8$).

IRT provides automatic quality control by weighting items according to their empirical contribution to distinguishing model abilities. In our calibrated banks, 3.2-5.7% of items exhibit negative discrimination ($a_i < 0$), indicating systematic flaws where stronger models perform worse. For example, WinoGrande item #247 achieves 0.89 accuracy but $a_i = -0.43$ due to exploitable linguistic artifacts. Under raw accuracy, this flawed item contributes equally (weight = 1/N) to all scores, potentially inflating weak models. Under IRT, negative discrimination automatically down-weights its contribution by 82% (effective weight $\propto a_i^2 \approx 0.18$), reducing measurement contamination and providing more reliable ability estimates than accuracy alone. Similar patterns of identical accuracy leading to different ability estimates are observed across ARC, HellaSwag, and other benchmarks, with additional heatmap visualizations provided in Appendix G.

5 Discussion

5.1 Implications for Benchmark Design

Our findings underscore several critical considerations for the design of future large language model (LLM) benchmarks. First, reduced item sets necessitate **content balancing** to preserve assessment validity. Content balancing entails maintaining proportional representation across distinct skill domains or cognitive competencies to ensure that the benchmark reflects the intended construct rather than overemphasizing specific subskills. Without such balance, benchmarks risk producing biased or unrepresentative evaluations. The proposed adaptive framework can be readily extended to optimize both domain coverage and measurement precision simultaneously [22]. Achieving comparable balance using static, reduced item sets would be substantially more difficult and would require extensive manual tuning or domain expertise.

Second, **item quality** exerts a direct influence on evaluation reliability. Our analysis reveals that 3.2–5.7% of items display negative discrimination parameters, indicating systematic flaws that undermine accuracy-based performance estimates. Although IRT automatically attenuates the influence of such items through probabilistic weighting, many existing benchmarks lack analogous quality control mechanisms. Consequently, poorly constructed or ambiguous items can distort aggregate scores and obscure genuine model differences. Future benchmark design should therefore adopt psychometric validation as standard practice, incorporating procedures such as discrimination screening and content alignment checks to enhance the interpretability and fairness of evaluation outcomes.

Third, item exposure rate and test overlap rate are essential metrics for **benchmark robustness and reusability**. A low exposure rate ensures that individual items are rarely reused across assessments, reducing the risk of model memorization or contamination of test content. Similarly, a low test overlap rate—the proportion of shared items between distinct benchmark forms—preserves evaluation independence and minimizes redundancy. Adaptive benchmarking intrinsically achieves lower exposure and overlap rates by tailoring item selection to each model's estimated proficiency, generating diverse yet psychometrically comparable test forms. This adaptivity confers a distinct advantage over static benchmarks, enabling more secure, scalable, and sustainable evaluation practices in iterative model development contexts.

5.2 Limitations and Future Work

Despite notable efficiency gains, our framework has several limitations. Initial calibration relies on a representative model population, which may become outdated as architectures evolve. Nonetheless, adaptive testing supports incremental updates: new model responses can be incorporated to refine item parameters and maintain calibration accuracy over time. The current implementation remains limited to multiple-choice formats, while generative or open-ended tasks require alternative scoring and modeling strategies. Moreover, our framework is unidimensional, assuming a single latent proficiency dimension across all items, whereas LLM capabilities are inherently multidimensional. Future work should extend the framework to multidimensional IRT formulations that jointly model reasoning, factuality, and linguistic ability.

To facilitate broader adoption, we have implemented a modular scoring interface that allows users to define custom evaluation functions (e.g., determining whether a model answers a selected item correctly). Future work will explore online or Bayesian calibration techniques for continuous item updating, multidimensional modeling to capture diverse LLM capabilities, and hybrid adaptive–generative evaluation for open-ended tasks. We also plan to extend the system with interfaces for cross-model benchmarking, enhanced item exposure control, and visualization tools that improve interpretability and diagnostic insight.

6 Conclusion

Adaptive testing replaces exhaustive benchmarking with efficient, contamination-resistant measurement. Our ATLAS framework reduces items by 90% while maintaining precision, separates models with identical accuracies using IRT-based ability estimates, and limits contamination through $< 10\%$ item exposure rates. With benchmark saturation and contamination risks increasing, adaptive testing provides a rigorous alternative, applying measurement science principles instead of static averaging.

References

[1] Frederic M. Lord. Applications of Item Response Theory to Practical Testing Problems. Lawrence Erlbaum Associates, Hillsdale, NJ, 1980.

[2] Howard Wainer, Neil J Dorans, Ronald Flaugher, Bert F Green, and Robert J Mislevy. Computerized adaptive testing: A primer. Routledge, 2000.

[3] David J Weiss. Improving measurement quality and efficiency with adaptive testing. Applied psychological measurement, 6(4):473–492, 1982.

[4] Allan Birnbaum. Some latent trait models. Statistical theories of mental test scores, 1968.

[5] Ronald K Hambleton, Hariharan Swaminathan, and H Jane Rogers. Fundamentals of item response theory, volume 2. Sage, 1991.

[6] John P Lalor, Pedro Rodriguez, João Sedoc, and Jose Hernandez-Orallo. Item response theory for natural language processing. In Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Tutorial Abstracts, pages 9–13, 2024.

[7] Gauthier Guinet et al. Evaluating large language models with psychometrics. OpenReview, 2025.

[8] Felipe Maia Polo, Lucas Weber, Leshem Choshen, Yuekai Sun, Gongjun Xu, and Mikhail Yurochkin. tinybenchmarks: evaluating llms with fewer examples. In International Conference on Machine Learning (ICML), 2024.

[9] Alex Kipnis, Konstantinos Voudouris, Luca M. Schulze Buschoff, and Eric Schulz. metabench – a sparse benchmark of reasoning and knowledge in large language models. In The Thirteenth International Conference on Learning Representations (ICLR), 2025.

[10] April Zenisky, Ronald K Hambleton, and Richard M Luecht. Multistage testing: Issues, designs, and research. In Elements of adaptive testing, pages 355–372. Springer, 2009.

[11] Yi Zheng and Hua-Hua Chang. On-the-fly assembled multistage adaptive testing. Applied Psychological Measurement, 39(2):104–118, 2015.

[12] Xiuxiu Tang, Yi Zheng, Tong Wu, Kit-Tai Hau, and Hua-Hua Chang. Utilizing response time for item selection in on-the-fly multistage adaptive testing for pisa assessment. Journal of Educational Measurement, 2024.

[13] Yan Zhuang, Qi Liu, Yuting Ning, Weizhe Huang, Rui Lv, Zhenya Huang, Guanhao Zhao, Zheng Zhang, Qingyang Mao, Shijin Wang, et al. Efficiently measuring the cognitive ability of llms: An adaptive testing perspective. arXiv preprint arXiv:2306.10512, 2023.

[14] Yan Zhuang, Qi Liu, Zachary A. Pardos, Patrick C. Kyllonen, Jiyun Zu, Zhenya Huang, Shijin Wang, and Enhong Chen. Position: Ai evaluation should learn from how we test humans. In International Conference on Machine Learning (ICML), 2025.

[15] Michael J. Kolen and Robert L. Brennan. Test Equating, Scaling, and Linking: Methods and Practices. Springer, New York, NY, 3rd edition, 2014.

[16] R Philip Chalmers. mirt: A multidimensional item response theory package for the r environment. Journal of statistical Software, 48:1–29, 2012.

[17] R Darrell Bock and Robert J Mislevy. Adaptive eap estimation of ability in a microcomputer environment. Applied psychological measurement, 6(4):431–444, 1982.

[18] Thomas A Warm. Weighted likelihood estimation of ability in item response theory. Psychometrika, 54(3):427–450, 1989.

[19] Frank B Baker and Seock-Ho Kim. Item response theory: Parameter estimation techniques. CRC press, 2004.

[20] Alberto Maydeu-Olivares. Making sense of item fit statistics. Educational and Psychological Measurement, 75(3):365–387, 2015.

[21] G Gage Kingsbury and Anthony R Zara. Procedures for selecting items for computerized adaptive tests. Applied measurement in education, 2(4):359–375, 1989.

[22] Ying Cheng and Hua-Hua Chang. The maximum priority index method for severely constrained item selection in computerized adaptive testing. British journal of mathematical and statistical psychology, 62(2):369–383, 2009.

[23] Mary J Allen and Wendy M Yen. Introduction to measurement theory. Waveland Press, 2001.

[24] Shyh-Yueh Chen. A simplified procedure for estimating item exposure rates in computerized adaptive testing. Applied Psychological Measurement, 29(3):209–227, 2005.

[25] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. arXiv preprint arXiv:2009.03300, 2020.

A Detailed Analysis of Average Score Limitations

Average score (percent correct) remains the most widely reported metric for evaluating LLMs. While it provides a convenient ordinal indicator for fixed forms, it is a shaky measure of underlying ability.

First, average scores are form-dependent: changing the mix or difficulty of items alters percent correct, even if the model's true ability is unchanged. Second, the metric has a nonlinear scale: improvements at the extremes (e.g., $98\% \rightarrow 100\%$) do not reflect the same underlying gain as improvements in the middle (e.g., $50\% \rightarrow 52\%$). Third, it assumes equal informativeness across items, allowing easy or guessable items to influence the mean as much as highly discriminative ones. Fourth, it is subject to coverage bias: the observed score reflects the content blueprint of the test rather than ability across domains. Fifth, average scores offer no measure of uncertainty, making it unclear whether differences are statistically meaningful. Finally, they are highly sensitive to contamination: memorized items from pretraining can artificially inflate percent correct without reflecting genuine reasoning or generalization.

In contrast, IRT-based ability estimates (θ) provide form-invariant, uncertainty-aware measures that adjust for item difficulty and discrimination. Reporting $\theta \pm \mathrm{SE}(\theta)$ offers a psychometrically principled alternative. For communication purposes, reconstructed percent scores may be shown alongside, but θ should serve as the primary indicator of model capability.

B Comparison of IRT-Based Benchmark Methods

Table 3: Comparison of IRT-Based Benchmark Methods for LLMs

Factor	**TinyBenchmarks (Static)**	**MetaBench (Static)**	**ATLAS**
IRT Calibration	Required	Required	Required
Adaptivity	None (same items)	None (same items)	High (items vary by ability)
Test Length	Fixed	Fixed	Variable, stopping rules
Exposure control	High (same items reused)	High (same items reused)	Low (rotating pool)
Pool Sensitivity	Subset dependent	Subset dependent	Robust to large pools
Fairness	Biased if mis-targeted	Biased if mis-targeted	Balanced across abilities
Score Precision	Low at extremes	Low at extremes	High, SEs available
Model Fit	Rarely checked	Rarely checked	Possible fit checks
Efficiency	Fixed (100 items)	Fixed (100 items)	Fewer items, comparable accuracy
Saturation Risk	High	High	Low

C Detailed Experimental Setup and Metrics

C.1 Benchmarks and Datasets

We conduct experiments on five diverse benchmarks covering different cognitive domains:

- **WinoGrande**: Commonsense reasoning with pronoun resolution
- **TruthfulQA**: Factual consistency and truthfulness evaluation
- **HellaSwag**: Procedural inference and common sense completion
- **GSM8K**: Mathematical word problems requiring multi-step reasoning
- **ARC**: Scientific question answering across multiple domains

All experiments use the calibrated item banks from Section 3.3, ensuring consistent filtering and parameter quality across datasets. Items with poor psychometric properties (negative discrimination $a_i < 0$, extreme difficulty $|b_i| > 4$, or unrealistic guessing $c_i > 0.5$) were excluded during calibration.

C.2 Baseline Configurations

We compare ATLAS against four fixed, non-adaptive strategies:

Random Baseline: Samples 100 items uniformly from the full bank without consideration of item parameters or model ability.

TinyBenchmarks: Uses the predetermined subset from [8], selected via clustering methods but without explicit Fisher information optimization for ability estimation.

MetaBench-Primary and MetaBench-Secondary: Curated splits from [9] that require computationally expensive iterations to identify stable subsets. These splits emphasize predictive accuracy over psychometric validity.

All baseline data is available on Hugging Face: tinyBenchmarks, HCAI/metabench.

Unlike ATLAS, these approaches do not adapt to individual test-takers and serve only as static reference points for accuracy–efficiency tradeoffs.

C.3 ATLAS Configuration Details

For each model ℓ, we run ATLAS under three precision-based stopping thresholds:

- $\mathrm{SE}(\hat{\theta}) \leq 0.1$: High precision, suitable for fine-grained model comparison
- $\mathrm{SE}(\hat{\theta}) \leq 0.2$: Moderate precision, balancing accuracy and efficiency
- $\mathrm{SE}(\hat{\theta}) \leq 0.3$: Lower precision, maximizing efficiency for rapid screening

A minimum of 30 items is enforced to prevent premature termination due to lucky guesses or initial high-information items, while the maximum is capped at 500 items to ensure computational feasibility. This setup balances precision and budget constraints, simulating realistic conditions for adaptive evaluation in production environments.

C.4 Detailed Metric Definitions

Average Mean Absolute Error (MAE):

$$\text{MAE} = \frac{1}{|\mathcal{M}|} \sum_{\ell \in \mathcal{M}} |\hat{\theta}_\ell - \hat{\theta}_\ell^{\text{whole}}|$$

where $\mathcal{M}$ is the set of evaluated models, $\hat{\theta}_\ell$ is the ATLAS-derived ability estimate, and $\hat{\theta}_\ell^{\text{whole}}$ is the reference estimate from the complete item bank.

Average Item Exposure Rate:

$$\text{Exposure} = \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \frac{\text{count}(i)}{|\mathcal{M}|}$$

where $\text{count}(i)$ is the number of times item i was administered across all models $\mathcal{M}$, and $|\mathcal{I}|$ is the total number of items in the bank.

Test Overlap Rate:

$$\bar{Q} = \frac{1}{\binom{|\mathcal{M}|}{2}} \sum_{\ell_1 < \ell_2} \frac{|\mathcal{T}_{\ell_1} \cap \mathcal{T}_{\ell_2}|}{\min(|\mathcal{T}_{\ell_1}|, |\mathcal{T}_{\ell_2}|)}$$

where $\mathcal{T}_\ell$ represents the set of items administered to model ℓ.

C.5 Psychometric Interpretation Guidelines

Item Exposure Rate Benchmarks: Values below 20–30% are considered desirable in large-scale adaptive testing, as they indicate that no small subset of items dominates the evaluation, reducing security risks and promoting item bank longevity.

Test Overlap Rate Benchmarks: Values under 10–15% are typically considered good in adaptive testing practice, providing sufficient form diversity to resist contamination while maintaining measurement consistency.

Runtime Considerations: All measurements reflect computational efficiency in fully automated settings. Runtime includes item selection via Fisher information maximization but excludes model inference time, which varies significantly across architectures.

D Model Fit Evaluation across Datasets

Table 4 summarizes the average Root Mean Square Error of Approximation (RMSEA) from the MIRT M2 model fit across multiple benchmark datasets. Overall, all datasets exhibit low RMSEA values (ranging from 0.0436 to 0.0690), indicating a generally good fit between the latent multidimensional item response model and the observed response data. Among the evaluated datasets, *GSM8K* achieves the lowest average RMSEA (0.0436), suggesting that the underlying latent trait structure aligns particularly well with the response patterns, likely due to its consistent problem-solving format. In contrast, *TruthfulQA* shows a slightly higher RMSEA (0.0690), reflecting greater variability or multidimensionality in the response space.

The narrow gaps between the 5% and 95% RMSEA bounds across datasets indicate model stability and relatively low dispersion of fit indices among individual files. This consistency across diverse benchmarks such as *Winogrande*, *HellaSwag*, and *ARC* supports the reliability of the MIRT calibration process. Taken together, these results suggest that the applied MIRT M2 model provides a satisfactory and robust representation of item-level relationships across heterogeneous reasoning and language understanding tasks.

Table 4: Summary of MIRT M2 Average RMSEA across datasets

Dataset	Avg. RMSEA	Avg. 5% RMSEA	Avg. 95% RMSEA	Num. Files
Winogrande	0.0565	0.0561	0.0568	10
TruthfulQA	0.0690	0.0686	0.0693	6
HellaSwag	0.0482	0.0478	0.0486	50
GSM8K	0.0436	0.0433	0.0440	12
ARC	0.0588	0.0585	0.0592	8

Note: 'Avg. RMSEA' is the mean RMSEA across all files for each dataset. 'Avg. 5% RMSEA' and 'Avg. 95% RMSEA' are the averages of the file-level 5th and 95th percentile RMSEA bounds, respectively.

E Data Preprocessing Details

E.1 Point-Biserial Correlation Formula

The point-biserial correlation [23] for item i is defined as:

$$r_{pb}(i) \;=\; \frac{\bar{T}_{\ell|Y_{i\ell}=1} - \bar{T}_{\ell|Y_{i\ell}=0}}{s_T} \cdot \sqrt{p_i q_i},$$

where $\bar{T}_{\ell|Y_{i\ell}=1}$ and $\bar{T}_{\ell|Y_{i\ell}=0}$ are the mean total scores of models that answered item i correctly and incorrectly, respectively; s_T is the standard deviation of total scores $\{T_\ell\}$; $p_i = \frac{1}{|\mathcal{L}|}\sum_{\ell\in\mathcal{L}} Y_{i\ell}$ is the proportion of models that answered item i correctly; $q_i = 1 - p_i$; and $|\mathcal{L}|$ is the number of models.

E.2 Filtering Results

F Evaluation Metric Definitions

This section provides the exact mathematical definitions of the evaluation metrics introduced in Section 4, along with brief interpretations.

	WinoGrande	TruthfulQA	HellaSwag	GSM8K	ARC
Num. Selected Models	4680	4635	3467	4195	4162
Num. Selected Item	1046	628	5608	1306	842
Avg. M_2 RMSEA	0.0565	0.0690	0.0482	0.0436	0.0588

Table 5: Summary of model and item filtering per benchmark. RMSEA values reflect average fit from M_2 statistics under the 3PL IRT model. Calibration and ability estimation were performed using the `mirt` package in `R`, with WLE as the estimator. Items were filtered based on variability, ceiling effects, and point-biserial discrimination.

Average Mean Absolute Error (MAE). For each model ℓ, let $\hat{\theta}_\ell$ denote the CAT-derived ability estimate and $\hat{\theta}_\ell^{\text{whole}}$ the full-bank reference estimate. The average MAE is:

$$\text{MAE} = \frac{1}{|\mathcal{L}|} \sum_{\ell \in \mathcal{L}} \left| \hat{\theta}_\ell - \hat{\theta}_\ell^{\text{whole}} \right|, \tag{4}$$

where $|\mathcal{L}|$ is the number of models evaluated. *Interpretation:* Lower MAE indicates higher fidelity, i.e., CAT ability estimates more closely match the whole-bank reference.

Average Item Exposure Rate. Let h_i denote the number of models administered item i, with $|\mathcal{I}|$ total items and $|\mathcal{L}|$ total models. The item exposure probability for item i is

$$P(A_i) = \frac{h_i}{|\mathcal{L}|}. \tag{5}$$

The average item exposure rate is then

$$\bar{P}(A_i) = \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} P(A_i). \tag{6}$$

Interpretation: Lower values indicate higher adaptivity and greater item diversity, while higher values suggest uniform or repetitive item usage across models.

Test Overlap Rate. Following **(author?)** [24], the expected proportion of common items between two randomly selected test forms is given by

$$\bar{Q} = \frac{|\mathcal{L}| \sum_{i=1}^{|\mathcal{I}|} P(A_i)^2}{\bar{L}(|\mathcal{L}| - 1)} - \frac{1}{|\mathcal{L}| - 1}, \tag{7}$$

where $\bar{L}$ is the average test length. *Interpretation:* Lower values of $\bar{Q}$ imply greater test form diversity, which reduces risks of collusion and item memorization.

Correlation Metrics For completeness, we provide the definitions of the rank-based correlation coefficients used in Section 4.

Spearman correlation.

$$\rho = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n(n^2 - 1)},$$

where d_i is the rank difference for observation i across the two measures.

Kendall correlation.

$$\tau = \frac{(\#\text{concordant pairs}) - (\#\text{discordant pairs})}{\frac{1}{2}n(n-1)}.$$

G Additional Experimental Results

While both *TinyBenchmarks* [8] and *MetaBench* [9] include MMLU [25] as part of their evaluation suites, they treat it as a single unified dataset by aggregating all 57 subject areas. In contrast, we perform evaluation on a per-subject basis. This design choice acknowledges the heterogeneous nature of MMLU, where each subject represents a distinct knowledge domain with varying linguistic characteristics, content distributions, and difficulty levels. Aggregating across all subjects can obscure these domain-specific patterns and limit interpretability in adaptive assessment.

Since each MMLU subject contains approximately 150 items, we randomly selected several representative subjects—*Anatomy*, *Astronomy*, and *High-School Macroeconomics*—to illustrate ATLAS's behavior under data-sparse conditions. The corresponding results are reported in Table 6. Despite the small number of items per subject, ATLAS consistently demonstrates robust adaptive evaluation performance. As the selection threshold is relaxed ($SE \leq 0.1 \rightarrow 0.3$), the mean absolute error (MAE) increases moderately (e.g., from 0.099 to 0.235 in *Anatomy*), while the number of evaluated items is substantially reduced (approximately 80%). This indicates that ATLAS effectively balances efficiency and accuracy, even in limited-data regimes.

Moreover, reductions in test overlap and exposure rates across the evaluated subjects suggest that the adaptive mechanism achieves broader item coverage and mitigates redundancy. Evaluation time also decreases proportionally with the number of items, confirming the computational efficiency of the adaptive process.

Table 6: Adaptive evaluation results on MMLU subsets (item range: 10–100). Lower is better ↓.

Benchmark	Method	MAE ↓	Avg. Item ↓	Test Overlap ↓	Exposure Rate ↓	Avg. Time (s) ↓
MMLU-Anatomy (113 items)	$\text{ATLAS}_{SE \leq 0.1}$	**0.099**	100.00	0.94	0.88	4.93
	$\text{ATLAS}_{SE \leq 0.2}$	0.149	53.29	0.54	0.47	2.70
	$\text{ATLAS}_{SE \leq 0.3}$	0.235	**20.49**	**0.32**	**0.18**	**0.98**
MMLU-Astronomy (136 items)	$\text{ATLAS}_{SE \leq 0.1}$	**0.099**	93.22	0.82	0.69	6.30
	$\text{ATLAS}_{SE \leq 0.2}$	0.157	48.21	0.48	0.35	3.29
	$\text{ATLAS}_{SE \leq 0.3}$	0.235	**11.80**	**0.40**	**0.11**	**0.79**
MMLU-High School Macroeconomics (281 items)	$\text{ATLAS}_{SE \leq 0.1}$	**0.089**	87.94	0.48	0.31	10.49
	$\text{ATLAS}_{SE \leq 0.2}$	0.197	27.64	0.22	0.13	3.46
	$\text{ATLAS}_{SE \leq 0.3}$	0.240	**18.26**	**0.20**	**0.09**	**2.34**

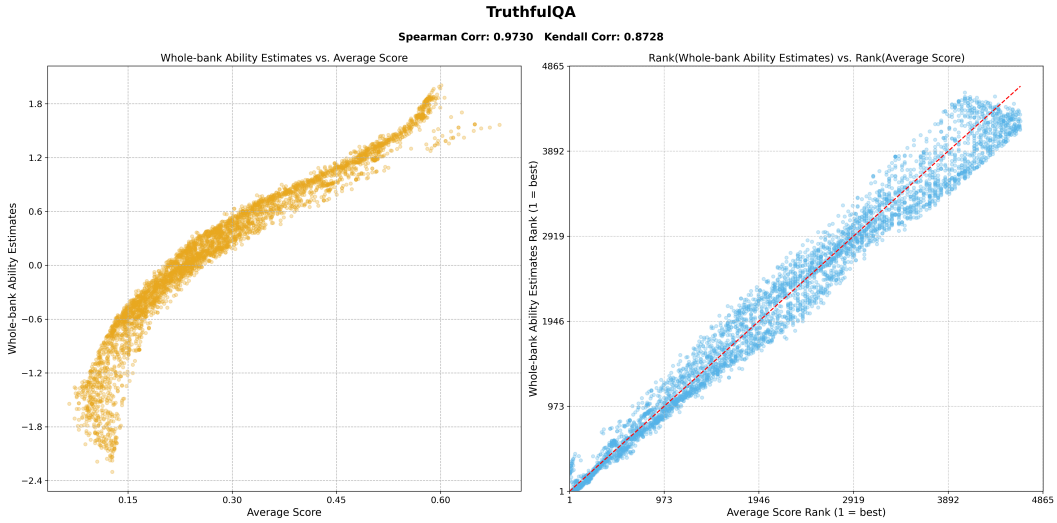

Figure 4: **Comparison of raw average scores and whole-bank ability estimates on TruthfulQA.** (Left) While average scores compress performance at the extremes, whole-bank ability estimates reveal clearer separation among both low- and high-performing models, reflecting sensitivity to item difficulty and discrimination. (Right) Rank comparison shows strong consistency between the two measures (Spearman $\rho = 0.97$, Kendall $\tau = 0.87$), but ability-based ranking provides finer resolution, especially in distinguishing weaker and stronger models beyond what raw accuracy captures.

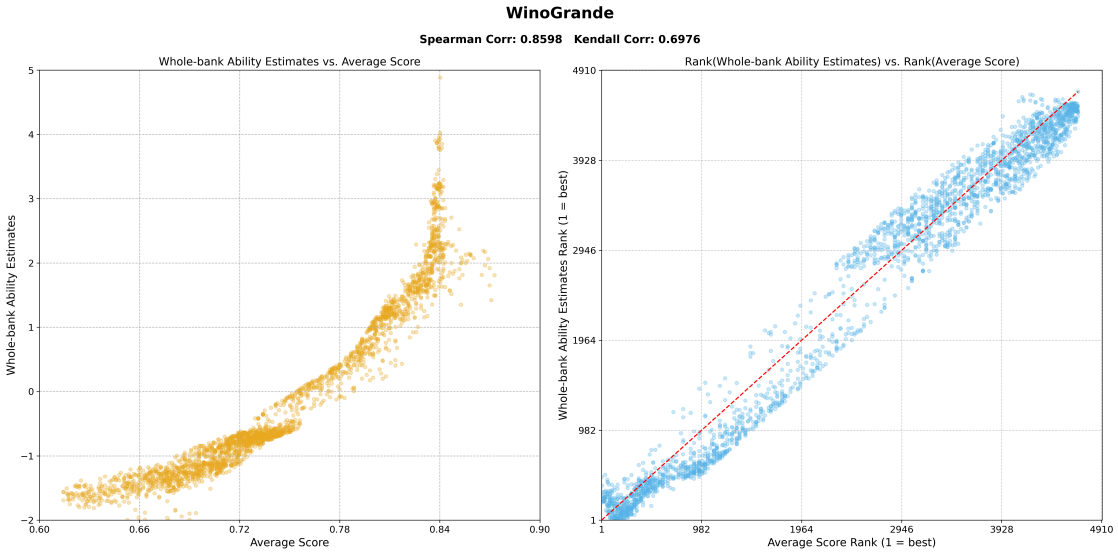

Figure 5: **Comparison of raw average scores and whole-bank ability estimates on WinoGrande.** (Left) Whole-bank estimates show a non-linear relationship with average score and reveal clearer separation on high-performing models, highlighting that ability captures relative item difficulty and provides finer differentiation beyond raw accuracy. (Right) Rank comparison indicates strong but imperfect alignment (Spearman $\rho = 0.86$, Kendall $\tau = 0.70$), with deviations from the diagonal reflecting cases where ability-based ranking distinguishes models more effectively than accuracy alone.

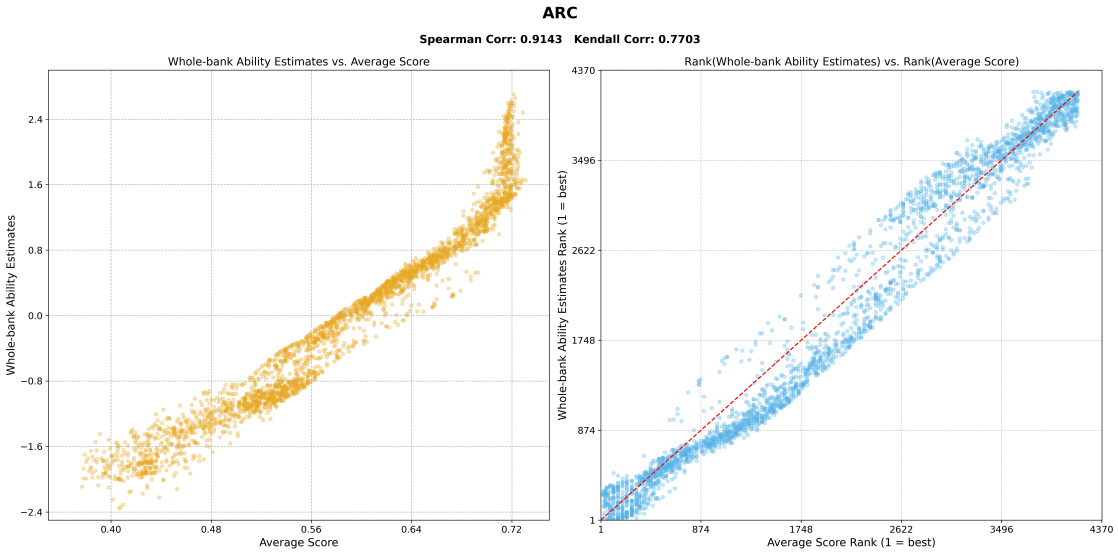

Figure 6: **Comparison of raw average scores and whole-bank ability estimates on ARC.** (Left) Whole-bank estimates exhibit a non-linear relationship with average scores, providing clearer separation on high-performing models by accounting for item difficulty and discrimination. (Right) Rank comparison shows strong but not perfect alignment between the two metrics (Spearman $\rho = 0.91$, Kendall $\tau = 0.77$), with deviations from the diagonal highlighting cases where ability-based ranking offers more informative distinctions than raw accuracy alone.

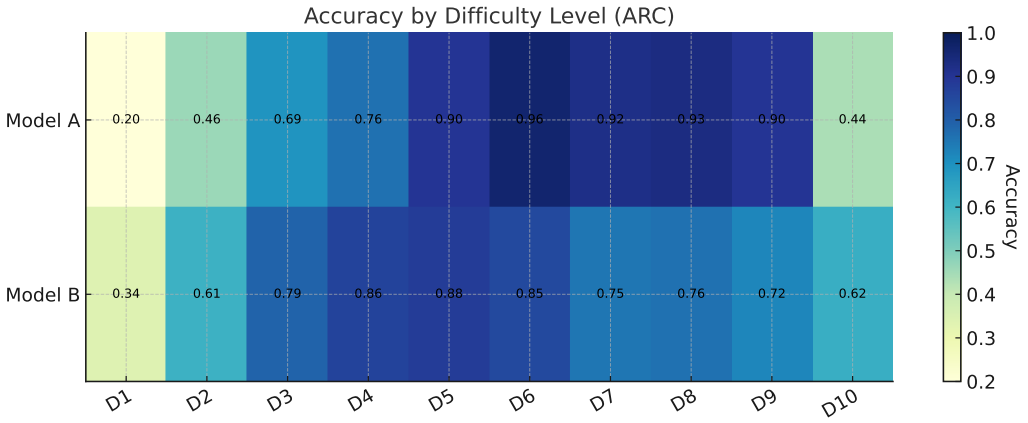

Figure 7: Two models with the similar average accuracy (0.713) and (0.714) on ARC nevertheless receive very different whole-bank ability estimates. Model A (`mera-mix-4x7B`) attains a whole-bank ability rank of 270 because its correct responses are concentrated on more difficult items. In contrast, Model B (`LLaMAAntino-3-ANITA-8B-Inst-DPO-ITA`) is assigned a much lower whole-bank ability rank of 2612, as its successes occur primarily on easier items. This divergence shows how IRT-based ability estimation can distinguish models that appear identical under raw accuracy by accounting for item difficulty.

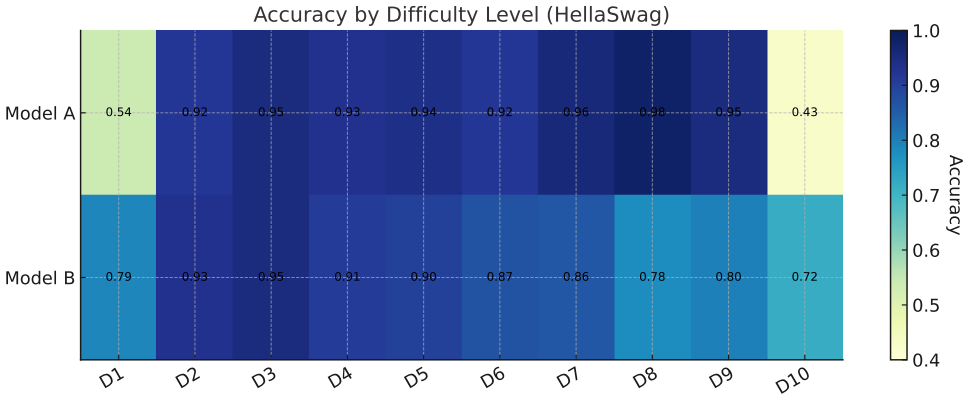

Figure 8: Two models with the same average accuracy (0.853) on HellaSwag nevertheless receive very different whole-bank ability estimates. Model A (`supermario_v1`) attains a whole-bank ability rank of 347 because its correct responses are concentrated on more difficult items. In contrast, Model B (`contaminated_proof_7b_v1.0_safetensor`) is assigned a much lower whole-bank ability rank of 3074, as its successes occur primarily on easier items. This divergence shows how IRT-based ability estimation can distinguish models that appear identical under raw accuracy by accounting for item difficulty.