Item Response Scaling Laws: A Measurement Theory Approach for Efficient and Generalizable Neural Scaling Estimation

Anonymous Authors[1]

Abstract

Scaling laws provide a fundamental framework for understanding the performance of Large Language Models (LLMs), yet deriving them requires prohibitively expensive evaluations across thousands of checkpoints or millions of inference samples. To address this, we introduce Item Response Scaling Laws (IRSL), a unified framework that integrates Item Response Theory (IRT) within scaling law formulation. Unlike traditional approaches that treat each model-benchmark pair in isolation, IRSL disentangles latent model ability from question characteristics, factorizing the scaling law estimation for M models and N questions to significantly reduce parameter complexity from $O(M \times N)$ to $O(M+N)$. We propose Beta-IRT, a novel extension that leverages the empirical probability responses of LLMs, such as token probabilities in pre-training and pass rates in test-time sampling, to capture richer signals than binary responses. We validate our approach across two prevalent scaling paradigms: (1) pre-training downstream scaling, using 6,612 LLM checkpoints and 37,682 questions from 10 benchmarks; and (2) test-time scaling, using 12 LLMs and 120 questions from 4 benchmarks with up to 2,500 samples per question. In both cases, we demonstrate that IRSL yields more reliable scaling estimates under limited query budgets. Furthermore, we show that the estimated latent model abilities are generalizable, enabling accurate performance forecasting across benchmarks that share the same measurement objective.

1. Introduction

Scaling laws provide a principled framework for predicting performance and allocating resources in Large Language Models (LLMs). We focus on two primary forms: pre-training downstream scaling, which characterizes how performance on downstream tasks improves with pre-training compute (Biderman et al., 2023; Grattafiori et al., 2024), and test-time scaling, which describes how performance improves with the number of independent inference samples (Brown et al., 2024; Hughes et al., 2024).

Deriving these laws is computationally expensive. A pre-training scaling study typically requires evaluating thousands of model checkpoints across tens of thousands of questions. Similarly, establishing test-time scaling laws requires a massive number of queries: number of models $\times$ number of questions $\times$ number of samples per question (typically $10^2 \times 10^5 \times 10^4$). Consequently, practical studies are often constrained to small experimental scales (Chen et al., 2024; Brown et al., 2024). The laws derived from such limited scales can exhibit unintuitive behaviors. For example, Brown et al. (2024) empirically find a power-law test-time scaling relationship that, as Schaeffer et al. (2025) demonstrates, holds only for specific, ill-structured distributions of single-sample success rates.

To address the cost of evaluation, we turn to Item Response Theory (IRT). Originating in psychology and human testing, IRT is a probabilistic framework that models the interaction between test takers and questions, known for significantly reducing the number of queries required to reliably estimate the ability of test takers. It has been highly successful in both human testing (Lord, 1980) and recent LLM leaderboard evaluations (Mishra et al., 2024; Magnusson et al., 2025; Gadre et al., 2024). These applications typically rely on binary responses[1]. However, unlike human testing, LLMs provide empirical probability responses. In pre-training, LLMs yield token probabilities that offer smoother scaling signals than discrete accuracy (Schaeffer et al., 2024; Magnusson et al., 2025). In test-time sampling, LLMs provide per-attempt success rates averaged from many independent inferences. Such empirical probability responses convey richer information than binary responses. To leverage this information, we propose Beta-IRT, which uses a Beta loss to model these

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

[1]Where the response of a test taker to a question is either correct or incorrect.

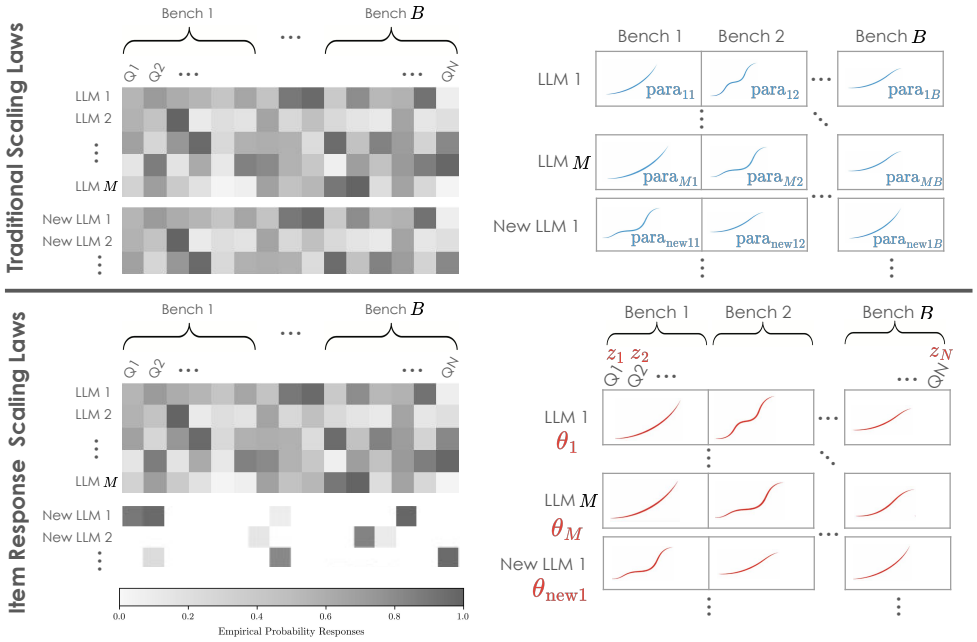

Figure 1. Overview of Item Response Scaling Laws (IRSL). The left panel displays the empirical probability response matrix across LLMs and benchmark questions. The sparse rows for "New LLMs" in IRSL demonstrate the query efficiency of IRSL, which can reliably estimate the scaling laws for new LLMs using a significantly smaller query budget. The right panel illustrates the parameterization difference. Traditional scaling laws (blue) fit separate parameters for each LLM-benchmark pair (e.g., $para_{11}$), resulting in a complexity of $O(M \times N)$. In contrast, IRSL (red) disentangles performance into LLM abilities θ and question characteristics z, factorizing the estimation to reduce parameter complexity to $O(M + N)$.

empirical probability responses.

Building on this, we introduce Item Response Scaling Laws (IRSL), a methodology that integrates IRT into the scaling law formulation. By modeling empirical probability responses with Beta-IRT, we can obtain more reliable scaling law estimates with a limited query budget. Furthermore, IRSL facilitates generalization. Traditional approaches treat each LLM-benchmark pair independently; if evaluating M models on N question sets, this requires fitting $M \times N$ separate scaling curves. In contrast, IRSL leverages the property of IRT to disentangle the ability of LLMs from the characteristics of the questions. This factorizes the problem into M sets of LLM-specific parameters and N sets of question-specific parameters, reducing the complexity to $O(M+N)$. This factorization allows the estimated ability to be transferred across benchmarks that share the same measurement objective. Our contributions are as follows:

- We conduct a large-scale study on 6,612 LLM checkpoints and 37,682 questions from 10 benchmarks to demonstrate the effectiveness of our pre-training downstream IRSL. We show that it yields generalizable and robust estimates of scaling behavior with limited query budgets.
- On 12 LLMs across 120 questions from 4 benchmarks with up to 2,500 samples per question, preliminary evidence suggests that IRSL similarly applies to test-time scaling.

By embedding the scaling law within the Beta-IRT framework, our approach provides a theoretically principled and empirically validated alternative to traditional aggregate performance scaling. Our code is released at `anonymous.4open.science/r/irsl-7560`. See Section A for related works.

2. Method

Item Response Theory (IRT) provides an elegant mathematical framework to model the interaction of LLMs and benchmark questions. We show how, under this framework, various known scaling laws arise naturally, and how the framework facilitates efficient and generalizable scaling laws estimation. We show the definitions, traditional fitting approaches, and IRT-based fitting approaches of the scaling laws in Table 1.

2.1. Traditional Binary-IRT

Item Response Theory refers to a class of probabilistic latent variable models that explain the relationship between the test taker's latent ability, the question's characteristics (e.g., difficulty), and the observed response from the test taker to the questions (Baker, 2001; Van der Linden et al.,

	Definition	Traditional Fitting Approach	IRT-based Fitting Approach
Pre-training Acc	$\mathrm{Acc}(i,\mathcal{D}) = \frac{1}{N}\sum_{j=1}^{N} Y_{ij}$	$a \cdot \sigma(b \cdot (\alpha \cdot \mathrm{FLOP}^{-\beta} + \gamma - l_0)) + c$	$\frac{1}{N}\sum_{j=1}^{N} \sigma(d_j \cdot (a \cdot \log(\mathrm{FLOP}_i) + b - z_j))$
Pre-training $\mathrm{p}_{\text{Correct Choice}}$	$\mathrm{p}_{\text{Correct Choice}}(i,\mathcal{D}) = \frac{1}{N}\sum_{j=1}^{N} \mathrm{p}_{\text{Correct Choice}}(i,j)$	$a \cdot \sigma(b \cdot (\alpha \cdot \mathrm{FLOP}^{-\beta} + \gamma - l_0)) + c$	$\frac{1}{N}\sum_{j=1}^{N} \sigma(d_j \cdot (a \cdot \log(\mathrm{FLOP}_i) + b - z_j))$
Test-time pass@k	$\mathrm{pass@k}(i,\mathcal{D}) = \frac{1}{N}\sum_{j=1}^{N} \mathrm{pass@k}(i,j)$	$\frac{1}{N}\sum_{j=1}^{N}(1-(1-\mathrm{pass@1}(i,j))^k)$	$\frac{1}{N}\sum_{j=1}^{N}(1-(1-\sigma(d_j \cdot (\theta_i - z_j)))^k)$

Table 1. The definitions, traditional fitting approach, and IRT-based fitting approach for Acc, $\mathrm{p}_{\text{Correct Choice}}$ (pre-training downstream scaling law), and pass@k (test-time scaling law). We use the 2PL model as a demonstration. Traditional fitting approaches fit parameters specific to LLMs and benchmarks. Our IRT-based fitting approach learns question-level parameters (rather than benchmark-level ones), thus being able to generalize across question sets with the same measurement objective.

2000). A central model in IRT is the 1PL model (Rasch, 1993), where each test taker has an ability parameter θ, and each question has a difficulty parameter z. A higher θ denotes greater ability, and a higher z denotes a more difficult question. Let y denote the binary response of the test taker to the question, where $y = 1$ if the response is correct and 0 otherwise. The probability of a correct response is modeled by $p(y = 1 \mid \theta, z) = \sigma(\theta - z)$, where σ is the sigmoid function. Another widely adopted model in IRT is the 2PL model (Lord, 1952; Birnbaum, 1968), which adds a discrimination parameter d to capture how sharply a question differentiates between test takers of different abilities, modeling the probability of a correct response as $p(y = 1 \mid \theta, z, d) = \sigma(d \cdot (\theta - z))$. The difficulty z and the discrimination d are collectively referred to as the item parameters. The use of IRT consists of two phases: calibration, which estimates the item parameters, and adaptive testing, which enables efficient ability estimation for new test takers.

During calibration, a binary response matrix Y of size $M \times N$ is collected, where M and N denote the number of test takers and questions, respectively. Entry Y_{ij} represents the response of test taker i to question j. With the binary response matrix, the item parameters can be estimated via either MLE or EM by minimizing the Bernoulli loss between the IRT predicted probabilities and the observed binary responses $\mathcal{L}_{\text{Bernoulli}} = -\sum_{i=1}^{M}\sum_{j=1}^{N}\left[Y_{ij}\log p_{ij} + (1 - Y_{ij})\log(1 - p_{ij})\right]$ (Bock & Aitkin, 1981; Chalmers, 2012; Wu et al., 2020).

During adaptive testing, the ability of a new test taker is efficiently estimated through an iterative procedure that alternates between ability update and question selection. In the ability update step, the test taker's ability is estimated from its responses to all previously asked questions. In the question selection step, the most informative question is selected for query based on the current ability estimate. Consequently, significantly fewer questions are required to obtain a reliable estimate of the new test taker's ability (Meijer & Nering, 1999; Chang, 2015; Magis et al., 2017).

2.2. Traditional Scaling Laws

We investigate two scaling laws: the pre-training downstream scaling law and the test-time scaling law. The pre-training downstream scaling law characterizes how the performance of an LLM i on a benchmark $\mathcal{D}$ scales with the pre-training compute FLOP. Traditional approach involves a two-step fitting process: first modeling the relationship between pre-training loss L and compute FLOP, and subsequently mapping the loss L to benchmark performance $\mathrm{Performance}(i,\mathcal{D})$ (Bhagia et al., 2024):

$$\begin{aligned} L &\approx \alpha \cdot \mathrm{FLOP}^{-\beta} + \gamma, \\ \mathrm{Performance}(i,\mathcal{D}) &\approx a \cdot \sigma(b \cdot (L - l_0)) + c, \end{aligned} \tag{1}$$

where σ denotes the sigmoid function, and $\alpha, \beta, \gamma, a, b, c$, and l_0 are learnable parameters. Following Bhagia et al. (2024), we use the benchmark-specific loss[2] as L. Consequently, all scaling law parameters are benchmark- and LLM-specific, implying that parameters derived for one LLM-benchmark pair do not generalize to others. $\mathrm{Performance}(i,\mathcal{D})$ can be quantified using metrics such as accuracy (Acc) or the average probability of the correct choice ($\mathrm{p}_{\text{Correct Choice}}$). Previous work notes that discrete metrics like Acc can exhibit performance jumps across scales, whereas continuous metrics like $\mathrm{p}_{\text{Correct Choice}}$ often reveal more predictable improvements (Schaeffer et al., 2024; Magnusson et al., 2025). See Appendix B for the calculation details of L, Acc, and $\mathrm{p}_{\text{Correct Choice}}$.

The test-time scaling law characterizes the relationship between the success rate of an LLM i on a benchmark $\mathcal{D}$ and the number of independent inference samples k. For an LLM i and a question j, $\mathrm{pass@1}(i,j)$ is defined as the probability that a single sample from LLM i correctly answers question j. The question-level success rate, $\mathrm{pass@k}(i,j)$, is defined as the probability that at least one of the k generated responses is correct. The benchmark-level success rate $\mathrm{pass@k}(i,\mathcal{D})$ is computed by averaging the probabilities over all benchmark questions $\mathrm{pass@k}(i,\mathcal{D}) = \frac{1}{N}\sum_{j=1}^{N}\mathrm{pass@k}(i,j)$. Previous stud-

[2]Can be understood as the pre-training validation loss on benchmark questions.

ies empirically find that $-\log \text{pass@k}$ exhibits a power-law decay with respect to k (Brown et al., 2024; Hughes et al., 2024): $-\log \text{pass@k}(i, \mathcal{D}) \approx uk^{-v}$, where u and v are scaling law parameters. Schaeffer et al. (2025) note that while the question-level success rate theoretically scales exponentially with k, the benchmark-level power law emerges because the distribution of $\text{pass@1}(i, j)$ is heavy-tailed towards extremely difficult questions. The relationship between $\text{pass@k}(i, \mathcal{D})$ and $\text{pass@1}(i, j)$ can be expressed as:

$$\text{pass@k}(i, \mathcal{D}) = \frac{1}{N}\sum_{j=1}^{N}(1 - (1 - \text{pass@1}(i, j))^k), \quad (2)$$

where pass@1 is benchmark- and LLM-specific.

2.3. Beta-IRT

Unlike human testing, LLM provides empirical probability responses that convey richer information than binary responses, such as $\text{p}_{\text{Correct Choice}}$ from pre-training downstream scaling and pass@1 from test-time scaling. Drawing on insights from Beta regression (Ferrari & Cribari-Neto, 2004), we propose Beta-IRT, which replaces the standard Bernoulli loss with the Beta loss: $\mathcal{L}_{\text{Beta}} = -\sum_{i=1}^{M}\sum_{j=1}^{N}\log p(P_{ij}; p_{ij}, \phi)$, where P_{ij} denotes the empirical response probability, p_{ij} denotes IRT predicted probability, and ϕ is a precision parameter. We empirically find that Beta-IRT achieves reliable calibration with significantly fewer test takers than Binary-IRT, substantially reducing calibration costs.

2.4. Item Response Scaling Laws

The core idea is to model $\text{p}_{\text{Correct Choice}}$ and pass@1 within the IRT framework. For the pre-training downstream scaling law, we employ a two-stage fitting procedure: first mapping pre-training compute FLOP to the ability θ, and subsequently mapping θ to the benchmark performance $\text{Performance}(i, \mathcal{D})$. Empirically, we observe that the θ scales linearly with $\log \text{FLOP}$ (Figure 12):

$$\begin{aligned} \theta_i &\approx a \cdot \log(\text{FLOP}_i) + b, \\ \text{Performance}(i, \mathcal{D}) &\approx \frac{1}{N}\sum_{j=1}^{N}\sigma(d_j \cdot (\theta_i - z_j)), \end{aligned} \quad (3)$$

where a, b, and θ_i are LLM-specific parameters, and d_j and z_j are question-specific parameters. Specifically, for the baseline scenario where $\text{Performance}(i, \mathcal{D})$ is measured by accuracy, we employ Binary-IRT with binary responses. For our approach, where $\text{Performance}(i, \mathcal{D})$ is measured by $\text{p}_{\text{Correct Choice}}$, we employ Beta-IRT with empirical probability responses.

With calibrated item parameters, adaptive testing enables the efficient estimation of a new LLM's ability using fewer questions, facilitating the rapid derivation of its pre-training downstream scaling law. Furthermore, IRSL offers generalizability across benchmarks. For a target benchmark $\mathcal{D}'$ sharing the same measurement objective as $\mathcal{D}$, the θ estimated from $\mathcal{D}$ is transferable. This allows for the prediction of performance on $\mathcal{D}'$ via $\text{Performance}(i, \mathcal{D}') \approx \frac{1}{N'}\sum_{j=1}^{N'}\sigma(d'_j \cdot (\theta_i - z'_j))$, obviating the need to collect empirical responses from LLM i on $\mathcal{D}'$.

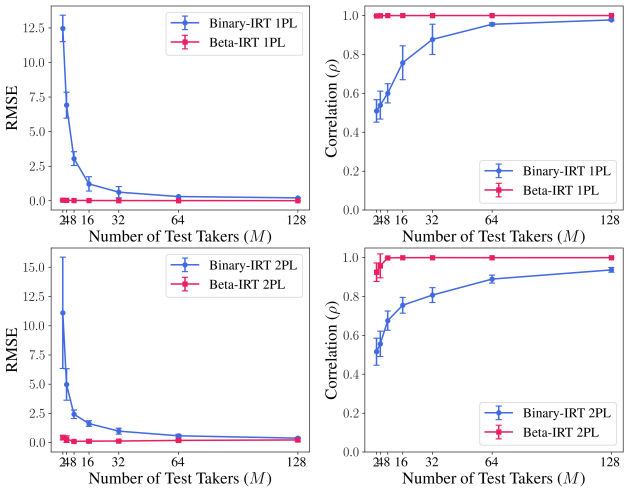

Figure 2. Sample efficiency comparison between Binary-IRT and Beta-IRT calibration. We report RMSE (Left) and Correlation (Right) for both the 1PL model (Top) and the 2PL model (Bottom) as a function of the number of test takers M. Error bars indicate ± 1 standard deviation over 10 trials. Beta-IRT significantly improves sample efficiency for calibration.

For the test-time scaling law, we model the benchmark-level success rate by substituting the Beta-IRT predicted single-attempt probability for $\text{pass@1}(i, j)$:

$$\text{pass@k}(i, \mathcal{D}) = \frac{1}{N}\sum_{j=1}^{N}(1 - (1 - \sigma(d_j \cdot (\theta_i - z_j)))^k), \quad (4)$$

where θ_i is LLM-specific parameter, and d_j and z_j are question-specific parameters. Similar to pre-training downstream scaling, our approach enables efficient estimation of a new LLM's ability using fewer questions, and the ability can generalize across different benchmarks sharing the measurement objective. Furthermore, in test-time scaling, a binary response tensor of shape $M \times N \times K$ is first collected, where K denotes the total number of samples. This tensor is averaged across the sample dimension to yield an empirical probability response matrix. In this setting, we empirically find that Beta-IRT facilitates the efficient estimation of a new LLM's ability using significantly fewer samples, further enhancing query efficiency.

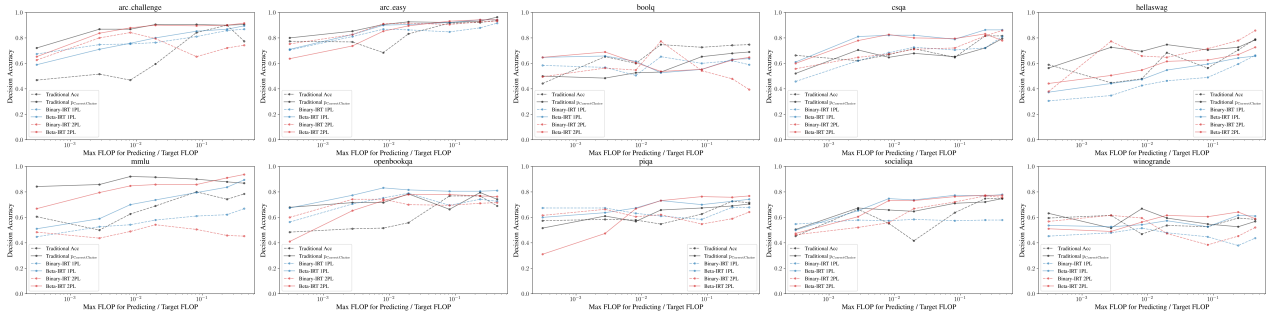

Figure 3. Decision Accuracy vs. Proportion of Target FLOPs across 10 Benchmarks. We iteratively fit scaling laws by including larger models and extrapolating to the target size to predict benchmark accuracy rankings. Black lines denote Traditional Scaling; Blue and Red lines denote IRSL 1PL and 2PL, respectively. Dashed lines indicate binary responses (Acc), while solid lines indicate empirical probability responses ($p_{\text{Correct Choice}}$). We conclude that Beta-IRT provides a more robust estimate of the scaling law curve, especially for benchmarks with lower quality.

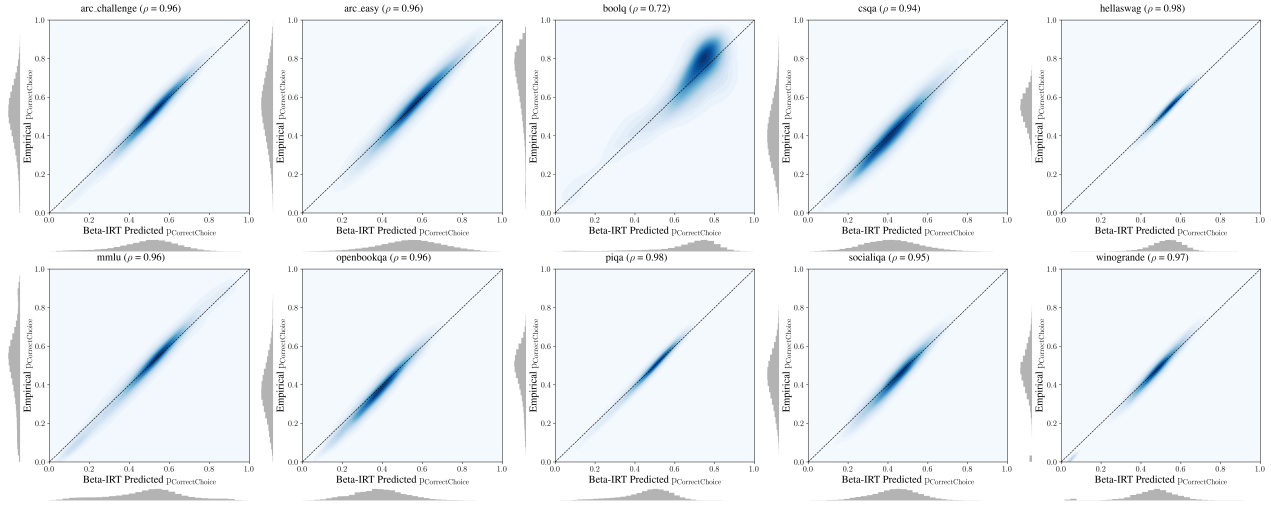

Figure 4. Correlation between Beta-IRT 2PL predicted $p_{\text{Correct Choice}}$ (x-axis) and empirical $p_{\text{Correct Choice}}$ (y-axis), visualized using 2-D Kernel Density Estimation (KDE) contour plots. The Pearson correlation coefficient (ρ) is reported for each benchmark, with marginal histograms showing the $p_{\text{Correct Choice}}$ distribution. The corresponding results for 1PL are provided in Figure 16 in Appendix C. We conclude that Beta-IRT effectively captures the underlying response structure.

3. Experiments

In Section 3.1, we conduct a simulation study to demonstrate the superior sample efficiency of Beta-IRT. In Section 3.2, we demonstrate the advantages of the Item Response Scaling Law (IRSL) for pre-training downstream scaling, and in Section 3.3, we preliminarily validate its effectiveness for test-time scaling.

3.1. Sample Efficiency of Beta-IRT

To quantify the information gain provided by empirical response probabilities, we conduct controlled simulations comparing the standard Binary-IRT with our proposed Beta-IRT for both 1PL and 2PL models. We generate true abilities $\theta_i \sim \mathcal{N}(0,1)$ for M test takers and question difficulties $z_j \sim \mathcal{N}(0,1)$ for $N = 100$ questions. For the 2PL model, question discriminations are sampled from $d_j \sim \text{LogNormal}(0, 0.5)$. We simulate binary response matrices $Y_{ij} \sim \text{Bernoulli}(p_{ij})$ and empirical probability matrices $P_{ij} = p_{ij} + \varepsilon_{ij}$, where the noise term $\varepsilon_{ij} \sim \mathcal{N}(0, 0.01^2)$ mimics empirical uncertainty.

We vary the number of test takers M across the set $\{2, 4, 8, 16, 32, 64, 128\}$, a range chosen to reflect the typical availability of test takers in LLM evaluation. We report the Root Mean Square Error (RMSE) and Pearson correlation coefficient (ρ) between the estimated and true item parameters, averaged over 10 independent trials. Figure 2 illustrates the substantial sample efficiency advantage of Beta-IRT. In the 1PL setting, Beta-IRT achieves near-perfect parameter recovery (RMSE < 0.05, $\rho > 0.999$) with as few as $M = 2$ test takers. In contrast, Binary-IRT requires significantly larger sample sizes ($M \geq 64$) to attain comparable accuracy. The 2PL model exhibits a similar trend: Beta-IRT 2PL maintains an RMSE < 0.7 across all sample sizes, while Binary-IRT 2PL begins with

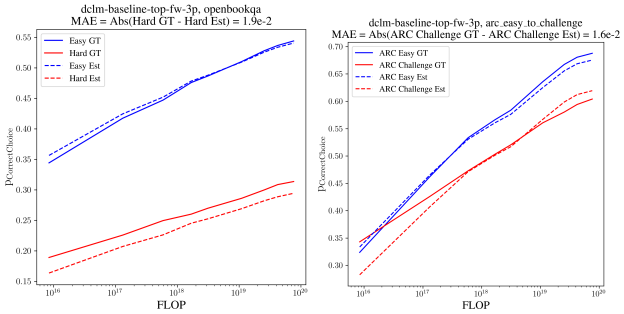

Figure 5. Generalizability of IRSL from easy to hard sets. (Left) Within-benchmark transfer on OpenBookQA. (Right) Cross-benchmark transfer from ARC Easy to ARC Challenge. Solid lines represent the Ground Truth (GT) scaling curves, while dashed lines represent the estimated curves where LLM ability is derived solely from the easy set. The close alignment between the Hard GT (red solid) and Hard Est (red dashed) demonstrates that IRSL can accurately predict the scaling trend on harder sets.

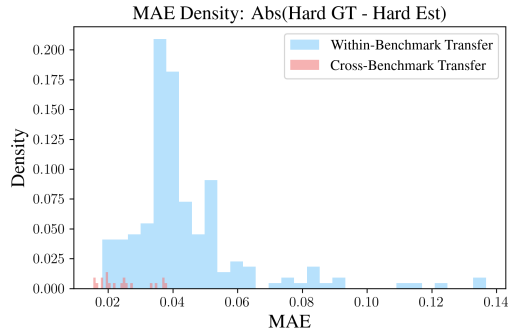

Figure 6. MAE distribution for hard set estimation across all benchmarks and LLM data mixtures. We report the MAE between the ground truth scaling curve and the estimated curve for two settings: Within-Benchmark Transfer (blue) and Cross-Benchmark Transfer (red). The consistently low MAE values indicate that the ability θ estimated by IRSL is robustly transferable, enabling reliable performance forecasting on benchmark sets that share the same measurement objective. See Figure 19 for the full results.

a high error and only approaches the performance of Beta-IRT 2PL at $M = 128$. These findings confirm that Beta-IRT significantly improves sample efficiency for calibration, reducing the high computational costs associated with large-scale LLM benchmarking.

3.2. Pre-training Downstream IRSL

We use the data suite from DataDecide (Magnusson et al., 2025), a large-scale controlled experiment on pre-training downstream scaling. The objective is to identify which of the 25 pre-training data mixtures yields the highest benchmark accuracy for the target model size (here, 1B). Because LLMs are expensive to pretrain, standard practice involves fitting scaling laws on smaller models and extrapolating to the target size. The suite comprises models pre-trained on 25 data mixtures across 14 model sizes, ranging from 4M to 1B parameters. Each run includes 6 to 30 checkpoints depending on the model size, resulting in a total of 6,612 model checkpoints. All checkpoints are evaluated on 10 multiple-choice benchmarks, totaling 37,682 questions. From this, we extract two response matrices of shape 6612×37682: a binary response matrix and an empirical probability response matrix. We randomly select 5 data mixtures to serve as the train set for calibration. The remaining 20 data mixtures constitute the test set for adaptive testing, where we estimate the ability θ using a budget of only 50 questions per benchmark.

We evaluate the effectiveness of a scaling law method using Decision Accuracy, a metric that quantifies rank consistency. Let $\mathcal{P}$ denote the set of all pairs of data mixtures (A, B) in the test set. Let y and $\hat{y}$ represent the ground truth benchmark accuracy at the 1B target size and the predicted performance extrapolated from the scaling law, respectively. Decision Accuracy is defined as:

$$\text{Decision Accuracy} = \frac{1}{|\mathcal{P}|} \sum_{(A,B)\in\mathcal{P}} \mathbb{I}(\text{sign}(\hat{y}_A - \hat{y}_B) = \text{sign}(y_A - y_B)). \quad (5)$$

We iteratively include larger models for the scaling law fitting and extrapolate to the target size to predict the benchmark accuracy rankings. Figure 3 reports the Decision Accuracy against the proportion of target FLOPs across 10 benchmarks. We compare six scaling law methods: traditional scaling law (using Acc or $\text{p}_{\text{Correct Choice}}$ via Equation 1) and IRSL (Binary-IRT and Beta-IRT, using 1PL and 2PL variants via Equation 3). On ARC Challenge, ARC Easy, and MMLU, Beta-IRT matches the strong performance of Traditional $\text{p}_{\text{Correct Choice}}$, and they outperform other methods. On CommonsenseQA, OpenBookQA, PIQA, SocialIQA, and WinoGrande, Beta-IRT demonstrates superior reliability, outperforming other methods. On BoolQ and HellaSwag, Beta-IRT fails to capture a predictive trend. These observations align with the findings of Heineman et al. (2025), a follow-up study on DataDecide that introduces a signal-to-noise ratio to assess benchmark quality in downstream scaling. Specifically, we find that Beta-IRT ties with Traditional $\text{p}_{\text{Correct Choice}}$ on high-quality benchmarks, outperforms other methods on lower-quality benchmarks, and fails to capture a trend on extremely noisy benchmarks. We conclude that Beta-IRT provides a more robust estimate of the scaling law curve with limited query budget, especially for benchmarks with lower quality. We report the scaling curve fitting for the six methods in Figure 13, 14, and 15 in Appendix C.

We report the strong correlation between Beta-IRT predicted $\text{p}_{\text{Correct Choice}}$ and the empirical $\text{p}_{\text{Correct Choice}}$ on the test set, as illustrated in Figure 4 for the 2PL variant and Figure 16 for the 1PL variant. We conclude that Beta-IRT effectively captures the underlying response structure. We

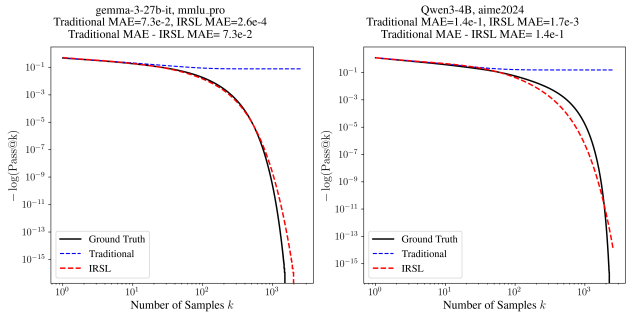

Figure 7. Comparison of three test-time scaling curves: Ground Truth, Traditional scaling law, and IRSL, for two representative LLM-Benchmark pairs in the test set. We plot $-\log \text{pass@k}$ against the number of samples k. We conclude that IRSL yields more reliable scaling estimates given a limited query budget.

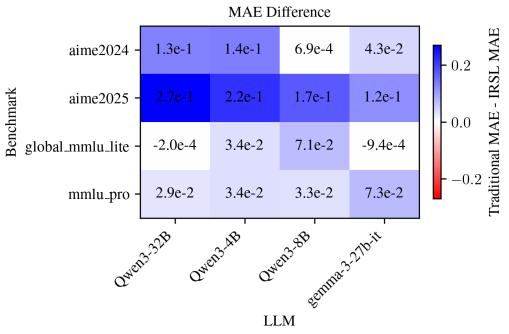

Figure 8. Reliability comparison of IRSL vs. Traditional scaling. The heatmap displays the performance gap Traditional MAE − IRSL MAE. Positive values (blue) indicate that IRSL achieves a lower MAE and thus provides a more accurate estimate. The consistent positive values across nearly all benchmarks and LLMs demonstrate the superiority of IRSL under a limited query budget. The corresponding results for the 2PL variant are provided in Figure 20 in Appendix D.

further report the Beta-IRT curve on single questions in Figure 17 and 18 in Appendix C.

Next, we demonstrate the generalizability of IRSL across benchmark sets with different difficulties. We partition each benchmark into an easy and a hard subset based on the mean $p_{\text{Correct Choice}}$ of each question across all LLM checkpoints. We estimate the ability θ of each LLM checkpoint using only the easy subset. Then, using these θ estimates alongside the calibrated item parameters of the hard subset, we generate the scaling curve for the hard subset without accessing the responses. Figure 5 (Left) illustrates this within-benchmark transfer for OpenBookQA on a representative LLM data mixture. We further demonstrate cross-benchmark transfer in Figure 5 (Right), showing that θ estimated on ARC Easy effectively predicts the scaling curve on ARC Challenge. Figure 19 reports the distribution of Mean Absolute Error (MAE) between the ground truth and the estimated scaling curve on the hard sets across all benchmarks and data mixtures (Full results at Figure 19. We conclude that the ability estimated by IRSL is transferable, enabling reliable performance forecasting on benchmark sets with the same measurement objective.

3.3. Test-time IRSL

We collect a binary response tensor of shape $12 \times 120 \times 2500$ (12 LLMs, 120 questions from 4 benchmarks, 2500 samples). We obtain the empirical pass@1 response matrix by averaging over the last dimension. We filter out questions with extremely low pass@1 as they offer no discriminatory power. We randomly select 8 LLMs to serve as the training set for calibration, while the remaining 4 LLMs constitute the test set for adaptive testing with a query budget of 50 samples per question. Given the limited number of LLMs for calibration, we report the 1PL model as our primary result and present the 2PL findings in

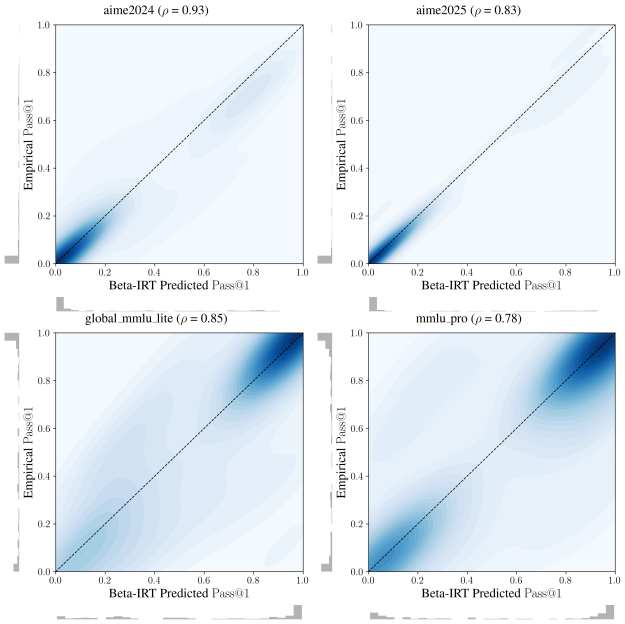

Figure 9. Correlation between Beta-IRT 1PL predicted pass@1 (x-axis) and empirical pass@1 (y-axis), visualized using 2-D Kernel Density Estimation (KDE) contour plots. The Pearson correlation coefficient (ρ) is reported for each benchmark, with marginal histograms showing the pass@1 distribution. The corresponding results for the 2PL variant are provided in Figure 21.

Appendix D[3].

We report three scaling curves ($-\log \text{pass@k}$ versus the number of samples k) for LLMs in the test set in Figure 7: (1) The Ground Truth curve, where pass@k is estimated from all available samples using the unbiased and numerically stable estimator proposed by Chen et al. (2021):

[3]The 2PL model typically requires more test takers to achieve reliable calibration.

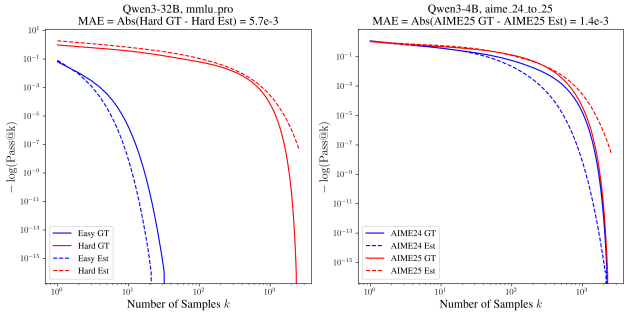

Figure 10. Generalizability of Test-time IRSL from easy to hard sets. (Left) Within-benchmark transfer on MMLU Pro. (Right) Cross-benchmark transfer from AIME 2024 to AIME 2025. The close alignment between Hard GT and Hard Est demonstrates that the test-time scaling trend on harder sets can be reliably forecasted using ability parameters estimated from the easy sets.

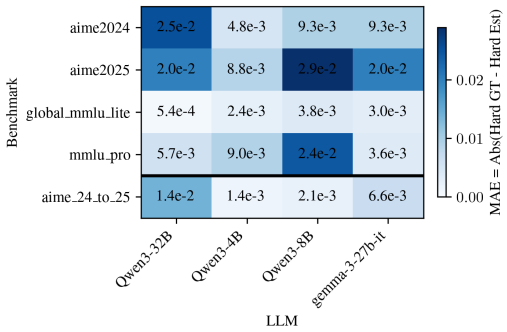

Figure 11. MAE of hard set estimation for test-time scaling. The last row specifically corresponds to the cross-benchmark transfer from AIME 24 to AIME 25. The consistently low MAE values indicate that the ability θ estimated by IRSL is transferable, enabling reliable performance forecasting on benchmark sets with the same measurement objective.

$\text{pass@k}(i,j) \approx 1 - \binom{H-c_{ij}}{k} / \binom{H}{k}$, where H is the total number of samples and c_{ij} is the number of correct samples by LLM i on question j. (2) The traditional scaling curve, where pass@k is estimated from the limited query budget via Equation 2. (3) The IRSL curve, where the ability θ is estimated from the same limited query budget, and pass@k is subsequently derived using Equation 4. As shown in Figure 7, there is a high alignment between the IRSL curve and the Ground Truth curve. To quantify the superiority of IRSL against traditional scaling law, we compute the MAE of $-\log \text{pass@k}$ for both methods relative to the ground truth. We visualize the performance gap Traditional MAE − IRSL MAE in Figure 8 across all benchmarks and test LLMs. The consistent positive values (indicated in blue) demonstrate that IRSL yields more reliable test-time scaling law estimates given a limited query budget. The corresponding results for the 2PL variants are presented in Figure 20.

We report the strong correlation between Beta-IRT predicted pass@1 and the empirical pass@1 on the test set, as illustrated in Figure 9 for the 1PL variant and Figure 21 for the 2PL variant. We further report the Beta-IRT curve on single questions in Figure 22 and 23 in Appendix D.

Next, we validate the generalizability of test-time IRSL across benchmark sets with different difficulty levels, following the same partitioning strategy used in our pre-training analysis. We estimate the ability θ using only the easy subset and transfer it to predict the scaling curve of the hard subset (or a harder benchmark) without accessing the response data. Figure 10 illustrates this capability: the left panel shows within-benchmark transfer for MMLU Pro using Qwen3-32B, while the right panel demonstrates cross-benchmark transfer, where θ estimated on AIME 2024 effectively predicts performance on AIME 2025. To quantify robustness across all settings, Figure 11 reports the MAE between the ground truth and the estimated scaling curves for the hard sets across all benchmarks and LLMs. The consistently low errors confirm that the ability parameters θ estimated by IRSL are robustly transferable, enabling reliable test-time forecasting on harder tasks sharing the same measurement objective.

4. Limitations and Future Work

IRSL excels when benchmarks have heterogeneous question difficulty, evaluation budgets are limited, and cross-question generalization is needed. However, traditional scaling with $\text{p}_{\text{Correct Choice}}$ already performs well on high-quality benchmarks with smooth probability responses (e.g., ARC Challenge, MMLU). In such cases, IRSL offers comparable accuracy with added interpretability but may not justify calibration overhead if only aggregate metrics are needed. On extremely noisy benchmarks (e.g., BoolQ, HellaSwag), neither approach captures reliable trends. Unlike classical power-law models that extrapolate to unseen compute regimes, IRSL requires pre-calibrated item difficulties from prior model responses, limiting applicability to established benchmarks. Difficulties calibrated under one evaluation setup may also not transfer to different conditions. IRSL is thus best viewed as complementary to classical scaling laws.

A primary limitation is the restricted data scale for test-time scaling analysis. Future work includes scaling up the test-time experimental setup, fitting shared latent abilities across benchmarks (Truong et al., 2025; Kipnis et al., 2025), exploring alternative probabilistic models (e.g., Beta-Binomial, zero-inflated models), extending to other scaling laws (Ruan et al., 2024; Kaplan et al., 2020; Arora et al., 2025), and polytomous IRT (Ostini & Nering, 2006).

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Arora, A., Jurafsky, D., Potts, C., and Goodman, N. D. Bayesian scaling laws for in-context learning, 2025. URL https://arxiv.org/abs/2410.16531.

Bahri, Y., Dyer, E., Kaplan, J., Lee, J., and Sharma, U. Explaining neural scaling laws. *arXiv preprint arXiv:2102.06701*, 2021.

Baker, F. B. *The basics of item response theory*. ERIC, 2001.

Bhagia, A., Liu, J., Wettig, A., Heineman, D., Tafjord, O., Jha, A. H., Soldaini, L., Smith, N. A., Groeneveld, D., Koh, P. W., et al. Establishing task scaling laws via compute-efficient model ladders. *arXiv preprint arXiv:2412.04403*, 2024.

Biderman, S., Schoelkopf, H., Anthony, Q. G., Bradley, H., O'Brien, K., Hallahan, E., Khan, M. A., Purohit, S., Prashanth, U. S., Raff, E., et al. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pp. 2397–2430. PMLR, 2023.

Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability. In Lord, F. M. and Novick, M. (eds.), *Statistical Theories of Mental Test Scores*, pp. 392–479. Addison-Wesley, Reading, MA, 1968.

Bock, R. D. and Aitkin, M. Marginal maximum likelihood estimation of item parameters: Application of an em algorithm. *Psychometrika*, 46(4):443–459, 1981.

Brown, B., Juravsky, J., Ehrlich, R., Clark, R., Le, Q. V., Ré, C., and Mirhoseini, A. Large language monkeys: Scaling inference compute with repeated sampling. *arXiv preprint arXiv:2407.21787*, 2024.

Chalmers, R. P. mirt: A multidimensional item response theory package for the r environment. *Journal of Statistical Software*, 48(6):1–29, 2012. doi: 10.18637/jss.v048.i06. URL https://www.jstatsoft.org/index.php/jss/article/view/v048i06.

Chang, H.-H. Psychometrics behind computerized adaptive testing. *Psychometrika*, 80(1):1–20, 2015.

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. Evaluating large language models trained on code, 2021. URL https://arxiv.org/abs/2107.03374.

Chen, Y., Huang, B., Gao, Y., Wang, Z., Yang, J., and Ji, H. Scaling laws for predicting downstream performance in llms. *arXiv preprint arXiv:2410.08527*, 2024.

Ferrari, S. and Cribari-Neto, F. Beta regression for modelling rates and proportions. *Journal of Applied Statistics*, 31(7):799–815, 2004. doi: 10.1080/0266476042000214501. URL https://doi.org/10.1080/0266476042000214501.

Gadre, S. Y., Smyrnis, G., Shankar, V., Gururangan, S., Wortsman, M., Shao, R., Mercat, J., Fang, A., Li, J., Keh, S., Xin, R., Nezhurina, M., Vasiljevic, I., Jitsev, J., Soldaini, L., Dimakis, A. G., Ilharco, G., Koh, P. W., Song, S., Kollar, T., Carmon, Y., Dave, A., Heckel, R., Muennighoff, N., and Schmidt, L. Language models scale reliably with over-training and on downstream tasks, 2024. URL https://arxiv.org/abs/2403.08540.

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Heineman, D., Hofmann, V., Magnusson, I., Gu, Y., Smith, N. A., Hajishirzi, H., Lo, K., and Dodge, J. Signal and noise: A framework for reducing uncertainty in language model evaluation, 2025. URL https://arxiv.org/abs/2508.13144.

Hernandez, D., Kaplan, J., Henighan, T., and McCandlish, S. Scaling laws for transfer. *arXiv preprint arXiv:2102.01293*, 2021.

Hestness, J., Narang, S., Ardalani, N., Diamos, G., Jun, H., Kianinejad, H., Patwary, M. M. A., Yang, Y., and Zhou, Y. Deep learning scaling is predictable, empirically. *arXiv preprint arXiv:1712.00409*, 2017.

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., Casas, D. d. L., Hendricks, L. A., Welbl, J., Clark, A., et al. Training

compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

Hofmann, V., Heineman, D., Magnusson, I., Lo, K., Dodge, J., Sap, M., Koh, P. W., Wang, C., Hajishirzi, H., and Smith, N. A. Fluid language model benchmarking, 2025. URL https://arxiv.org/abs/2509.11106.

Hughes, J., Price, S., Lynch, A., Schaeffer, R., Barez, F., Koyejo, S., Sleight, H., Jones, E., Perez, E., and Sharma, M. Best-of-n jailbreaking, 2024. URL https://arxiv.org/abs/2412.03556.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Kipnis, A., Voudouris, K., Buschoff, L. M. S., and Schulz, E. metabench – a sparse benchmark of reasoning and knowledge in large language models, 2025. URL https://arxiv.org/abs/2407.12844.

Lord, F. M. *A Theory of Test Scores*. Psychometric Corporation, Richmond, VA, 1952.

Lord, F. M. *Applications of Item Response Theory to Practical Testing Problems*. Routledge, 1st edition, 1980. doi: 10.4324/9780203056615.

Lourie, N., Hu, M. Y., and Cho, K. Scaling laws are unreliable for downstream tasks: A reality check. *arXiv preprint arXiv:2507.00885*, 2025.

Madaan, L., Singh, A. K., Schaeffer, R., Poulton, A., Koyejo, S., Stenetorp, P., Narang, S., and Hupkes, D. Quantifying variance in evaluation benchmarks, 2024. URL https://arxiv.org/abs/2406.10229.

Magis, D., Yan, D., and von Davier, A. A. *Computerized Adaptive and Multistage Testing with R*. Springer, Cham, 2017.

Magnusson, I., Tai, N., Bogin, B., Heineman, D., Hwang, J. D., Soldaini, L., Bhagia, A., Liu, J., Groeneveld, D., Tafjord, O., Smith, N. A., Koh, P. W., and Dodge, J. Datadecide: How to predict best pretraining data with small experiments, 2025. URL https://arxiv.org/abs/2504.11393.

Meijer, R. R. and Nering, M. L. Computerized adaptive testing: Overview and introduction. *Applied Psychological Measurement*, 23(3):187–194, 1999.

Mishra, S., Poesia, G., Mo, B., and Goodman, N. D. Mathcamps: Fine-grained synthesis of mathematical problems from human curricula. *arXiv preprint arXiv:2407.00900*, 2024.

Muennighoff, N., Rush, A., Barak, B., Le Scao, T., Tazi, N., Piktus, A., Pyysalo, S., Wolf, T., and Raffel, C. A. Scaling data-constrained language models. *Advances in Neural Information Processing Systems*, 36, 2024.

Ostini, R. and Nering, M. *Polytomous Item Response Theory Models*. Polytomous Item Response Theory Models. SAGE Publications, 2006. ISBN 9780761930686. URL https://books.google.com.hk/books?id=wS8VEMtJ3UYC.

Polo, F. M., Weber, L., Choshen, L., Sun, Y., Xu, G., and Yurochkin, M. tinybenchmarks: evaluating llms with fewer examples, 2024. URL https://arxiv.org/abs/2402.14992.

Rasch, G. *Probabilistic models for some intelligence and attainment tests.* ERIC, 1993.

Ruan, Y., Maddison, C. J., and Hashimoto, T. Observational scaling laws and the predictability of language model performance. *arXiv preprint arXiv:2405.10938*, 2024.

Schaeffer, R., Schoelkopf, H., Miranda, B., Mukobi, G., Madan, V., Ibrahim, A., Bradley, H., Biderman, S., and Koyejo, S. Why has predicting downstream capabilities of frontier ai models with scale remained elusive? *arXiv preprint arXiv:2406.04391*, 2024.

Schaeffer, R., Kazdan, J., Hughes, J., Juravsky, J., Price, S., Lynch, A., Jones, E., Kirk, R., Mirhoseini, A., and Koyejo, S. How do large language monkeys get their power (laws)?, 2025. URL https://arxiv.org/abs/2502.17578.

Snell, C., Lee, J., Xu, K., and Kumar, A. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*, 2024.

Truong, S., Tu, Y., Liang, P., Li, B., and Koyejo, S. Reliable and efficient amortized model-based evaluation. *arXiv preprint arXiv:2503.13335*, 2025.

Van der Linden, W. J., Glas, C. A., et al. *Computerized adaptive testing: Theory and practice*, volume 13. Springer, 2000.

Wu, M., Davis, R. L., Domingue, B. W., Piech, C., and Goodman, N. Variational item response theory: Fast, accurate, and expressive. *arXiv preprint arXiv:2002.00276*, 2020.

A. Related Work

Pre-training Downstream Scaling Law Many neural networks exhibit power-law scaling for the pre-training loss as a function of compute, data, or parameters (Hestness et al., 2017; Kaplan et al., 2020; Bahri et al., 2021; Hernandez et al., 2021; Hoffmann et al., 2022; Muennighoff et al., 2024). Unlike predicting loss, predicting downstream performance from scale is generally harder (Lourie et al., 2025; Schaeffer et al., 2024). However, recent work has demonstrated that it can be done based on a two-step prediction that chains together predictions from scale to loss and loss to downstream performance (Biderman et al., 2023; Magnusson et al., 2025; Gadre et al., 2024).

Test-time Scaling Law Test-time scaling laws characterize how a model's performance on a benchmark (e.g., success rate) improves as the number of stochastic samples drawn at inference increases, typically following a power law (Brown et al., 2024; Snell et al., 2024; Hughes et al., 2024). Schaeffer et al. (2025) demonstrate that such a power relationship holds only for ill-structured response distributions in single-sample success rates.

Measurement Theory-based AI evaluation Several recent works adopt Item Response Theory (IRT) as a foundation for AI evaluation using binary responses and Bernoulli loss (Truong et al., 2025; Hofmann et al., 2025; Kipnis et al., 2025; Madaan et al., 2024; Polo et al., 2024). Magnusson et al. (2025) introduce DataDecide, a fully-open large-scale controlled experiment designed to rigorously evaluate scaling laws for pre-training data mixtures. In a follow-up study, Heineman et al. (2025) propose a signal-to-noise ratio to quantify benchmark reliability, demonstrating that higher ratios correlate with better scaling predictions.

B. Pre-training Downstream Scaling Law Metrics Calculation Details

In this section, we explain the calculation of the benchmark-specific loss L, accuracy Acc, and the average probability of the correct choice $\mathrm{p}_{\text{Correct Choice}}$. Consider a question from a multiple-choice benchmark:

[Question Content]
A. [Choice A Content]
B. [Choice B Content]
C. [Choice C Content]
D. [Choice D Content]

Assuming the correct answer is C, the metrics are calculated as follows:

- **Benchmark-specific Loss:** Also known as bits per byte (BPB). For an individual question, this is calculated as the negative log-likelihood of the token sequence corresponding to the correct choice content (i.e., [Choice C Content]) conditioned on the question content (i.e., [Question Content]), normalized by the length of the correct choice content in bytes. The benchmarked-level value is averaged across all questions.
- **Average Probability of Correct Choice:** For an individual question, this measures the probability of the token sequence representing the correct choice content, conditioned on the question content, normalized by the character length of the choice. The benchmarked-level value is averaged across all questions.
- **Accuracy:** Also known as cloze formulation accuracy or RC format accuracy. This is determined by computing the probability of the token sequence for each choice content given the question content, normalized by the character length of each choice. The choice with the highest probability is selected as the predicted answer. The question is assigned a score of 1 if the prediction matches the correct choice, and 0 otherwise. The benchmarked-level value is averaged across all questions.

C. Additional Results for Pre-training Downstream IRSL

Figure 12 shows the empirical observation of the linear relationship between θ and $\log \mathrm{FLOP}$ for Beta-IRT 2PL. The trend is similar for Binary-IRT and 1PL variants.

Figure 13 shows the scaling curve fitting for traditional scaling law step 1. Figure 14 shows the scaling curve fitting for traditional scaling law step 2. Figure 15 shows the scaling curve fitting for IRSL step 1. Following Bhagia et al. (2024), we fit step 1 only on final checkpoints for each model size, as the learning rate schedule prevents accurate FLOP estimation on intermediate checkpoints.

Figure 16 shows the correlation between Beta-IRT 1PL predicted $p_{\text{Correct Choice}}$ and empirical $p_{\text{Correct Choice}}$. Figure 17 and 18 show the Beta-IRT curve on a randomly sampled question for 2PL and 1PL, respectively.

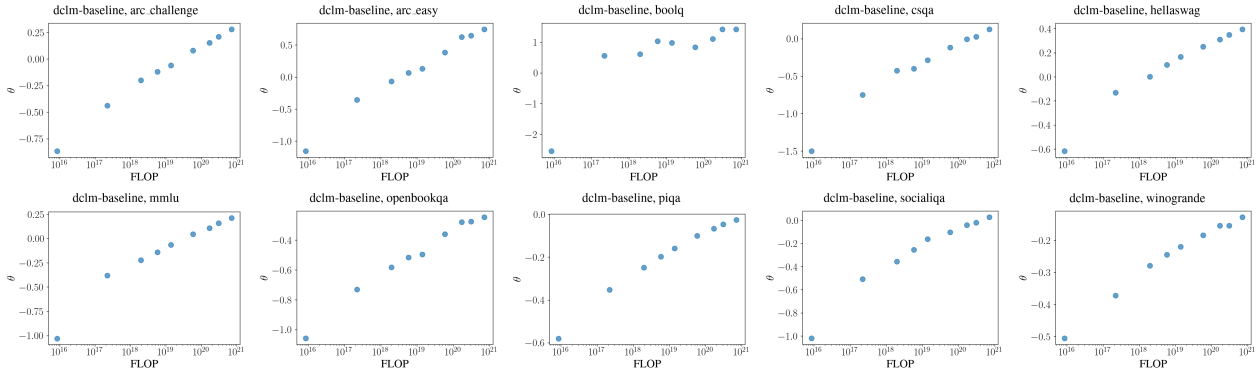

Figure 12. Linear relationship between θ and $\log \text{FLOP}$ for Beta-IRT 2PL on the test set. We display a representative LLM data mixture across all 10 benchmarks. This linear trend is consistent across other data mixtures, as well as Binary-IRT and 1PL variants.

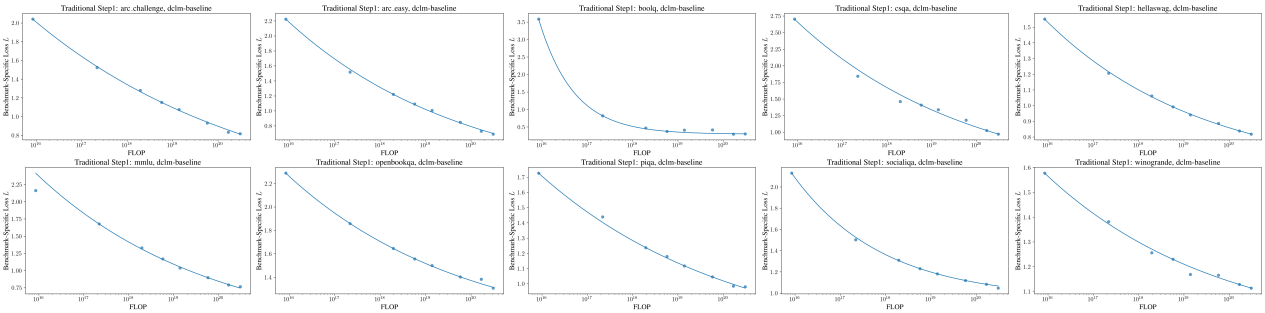

Figure 13. The scaling curve fitting for traditional scaling law step 1: $L \approx \alpha \cdot \text{FLOP}^{-\beta} + \gamma$. We display a representative LLM data mixture across all 10 benchmarks. The trend is consistent across other data mixtures.

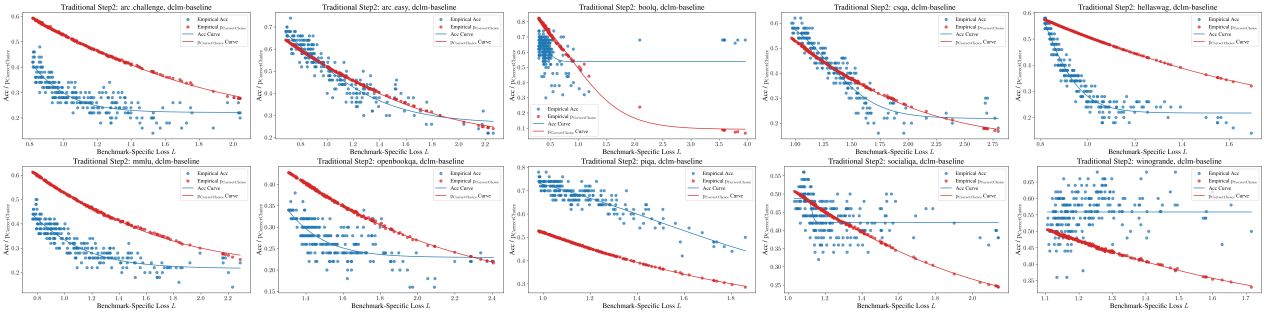

Figure 14. The scaling curve fitting for traditional scaling law step 2: $\text{Performance}(i, \mathcal{D}) \approx a \cdot \sigma(b \cdot (L - l_0)) + c$. We display a representative LLM data mixture across all 10 benchmarks. The trend is consistent across other data mixtures.

D. Additional Results for Test-time IRSL

Figure 20 shows the reliability comparison of IRSL vs. Traditional scaling for the 2PL model. Figure 21 shows the correlation between Beta-IRT 2PL predicted pass@1 and empirical pass@1. Figure 22 and Figure 21 show the Beta-IRT curve on a randomly sampled question for 1PL and 2PL, respectively.

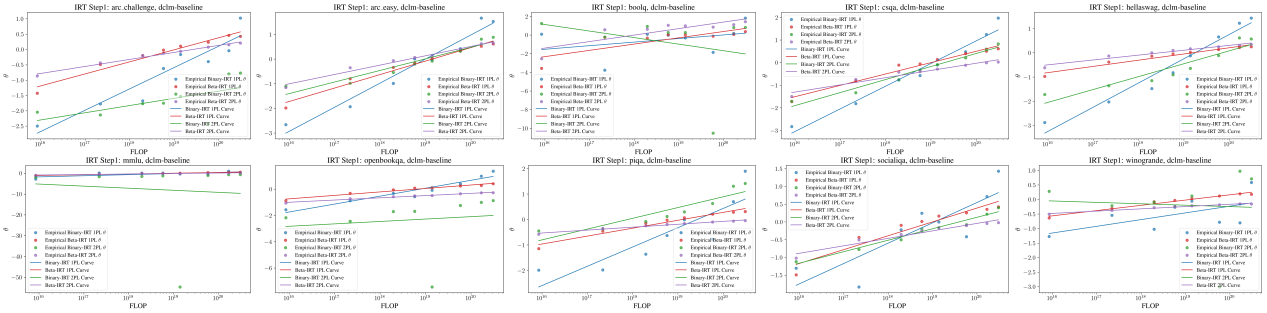

Figure 15. The scaling curve fitting for IRSL step 1: $\theta_i \approx a \cdot \log(\text{FLOP}_i) + b$. We display a representative LLM data mixture across all 10 benchmarks. The trend is consistent across other data mixtures.

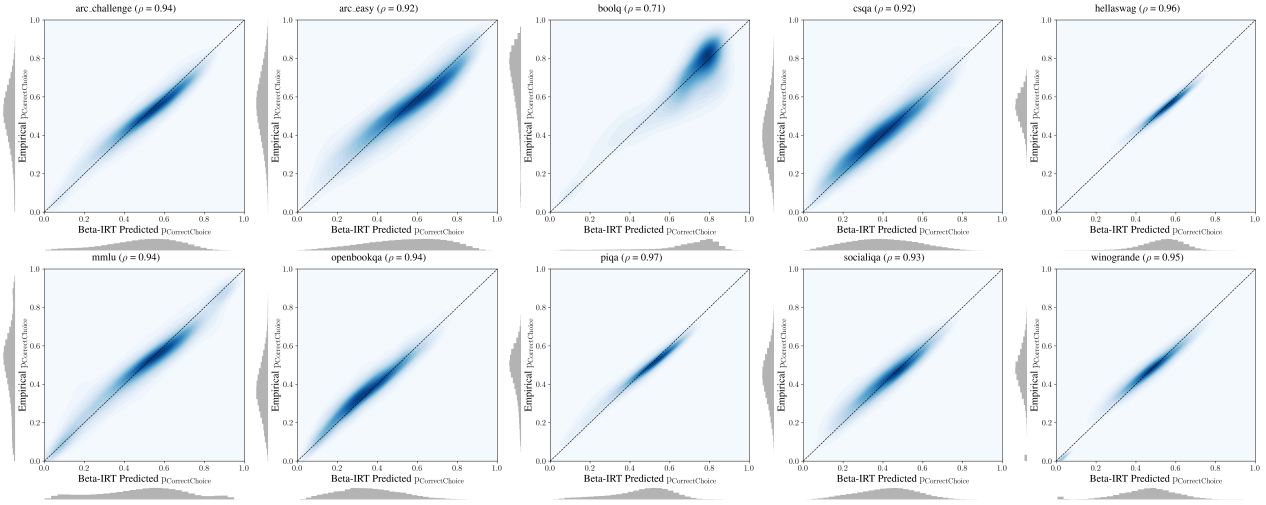

Figure 16. Correlation between Beta-IRT 1PL predicted $\text{p}_{\text{Correct Choice}}$ and empirical $\text{p}_{\text{Correct Choice}}$.

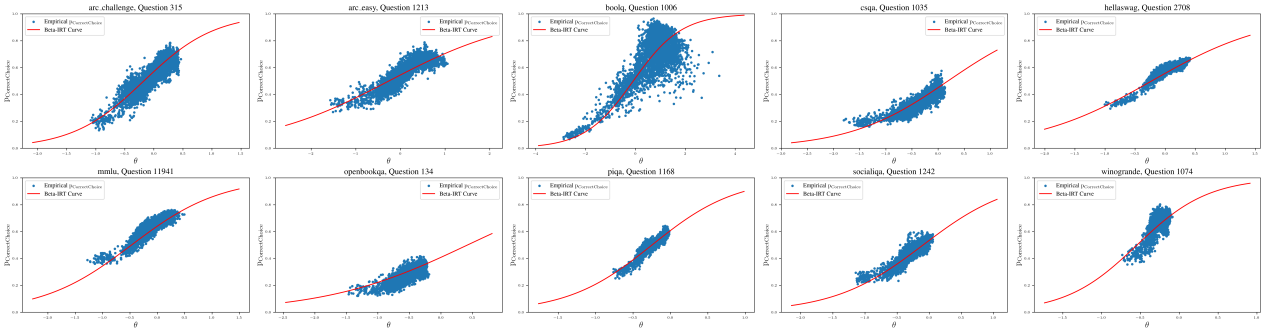

Figure 17. Beta-IRT 2PL curve on a single question for each benchmark. The x-axis is the ability parameter θ, and the y-axis is $\text{p}_{\text{Correct Choice}}$. The red line checkpoints the fitted Beta-IRT curve. The blue dots represent the empirical $\text{p}_{\text{Correct Choice}}$; each dot corresponds to an LLM checkpoint in the test set.

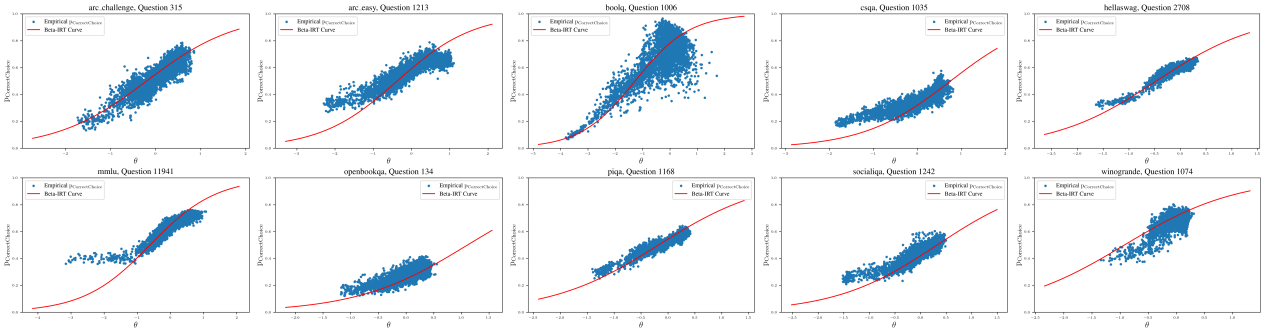

Figure 18. Beta-IRT 1PL curve on a single question for each benchmark.

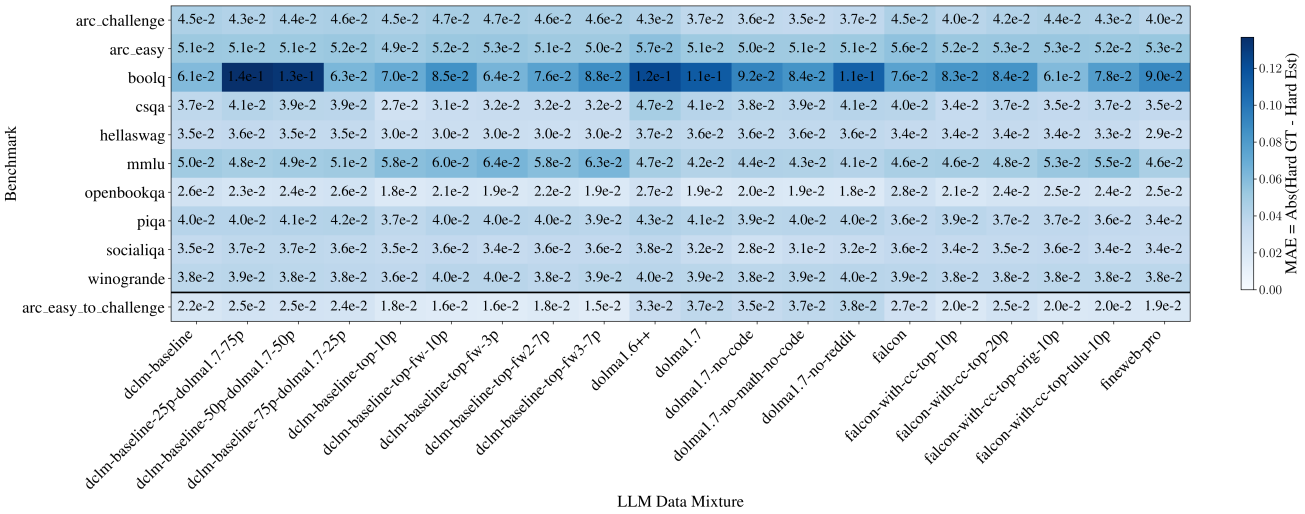

Figure 19. MAE of hard set estimation across all benchmarks and LLM data mixtures. We report the MAE between the ground truth scaling curve and the estimated curve on the hard sets. The last row specifically corresponds to the cross-benchmark transfer from ARC Easy to ARC Challenge.

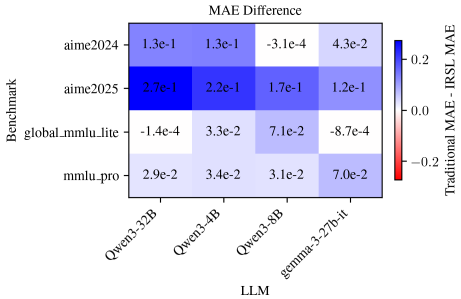

Figure 20. Reliability comparison of IRSL vs. Traditional scaling for the 2PL model.

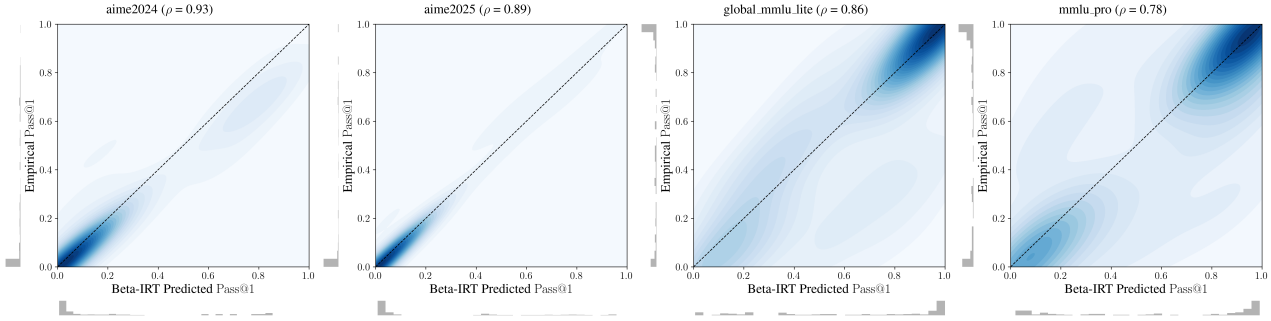

Figure 21. Correlation between Beta-IRT 2PL predicted pass@1 and empirical pass@1.

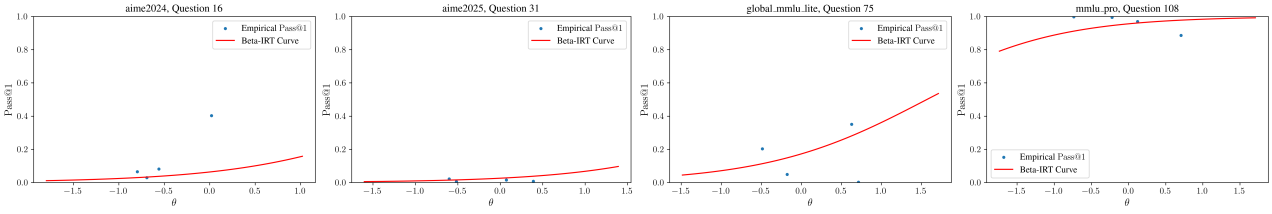

Figure 22. Beta-IRT 1PL curve on a single question for each benchmark. The x-axis is the ability parameter θ, and the y-axis is pass@1. The red line checkpoints the fitted Beta-IRT curve. The blue dots represent the empirical pass@1; each dot corresponds to an LLM in the test set.

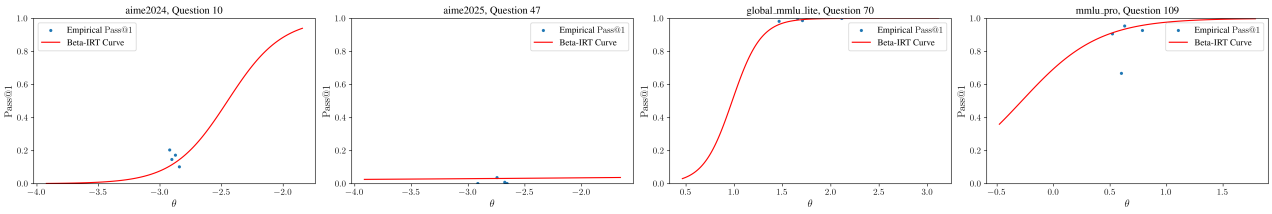

Figure 23. Beta-IRT 2PL curve on a single question for each benchmark.