RETHINKING ITEM RESPONSE THEORY FOR LLM EVALUATION WITH ITEM-MARGINALIZED MAXIMUM LIKELIHOOD

Anonymous authors
Paper under double-blind review

ABSTRACT

Item Response Theory (IRT) is increasingly used to rank language models, analyze benchmarks, and reduce evaluation cost. Yet conventional IRT estimation was developed for a response matrix geometry with many examinees and comparatively few items. Modern LLM evaluation often reverses this regime, comparing a small number of models on hundreds or thousands of benchmark items. We study the consequences of this mismatch and develop item-marginalized IRT estimation as an alternative formulation which treats evaluated model abilities as fixed inferential targets and benchmark items as random effects. Across simulations calibrated to six benchmarks, we find a systematic geometry-dependent crossover. Conventional person-marginalized IRT performs better when models are abundant, whereas item marginalization becomes increasingly advantageous as the response matrix becomes item rich. With only 12 models, item marginalization reduces average Kendall ranking loss across six benchmarks by 56% relative to person marginalization. This advantage translates directly to efficient LLM evaluation. With only 12–40 calibration models, item-marginalized item selection reduces held-out ranking loss by 18%, enabling IRT-based benchmark distillation from only a few dozen fully evaluated models rather than requiring large historical response matrices. On real sparse response matrices, item marginalization also consistently improves ranking recovery over conventional IRT. Mechanism and robustness experiments attribute these gains to improved population generalization of item characteristics and show that the crossover persists across alternative IRT models and strongly non-Gaussian latent populations. These results suggest that the direction of marginalization should reflect the geometry and generalization target of the evaluation, making IRT better aligned with modern LLM evaluation.

1 INTRODUCTION

Item Response Theory (IRT) refers to a family of psychometric modeling approaches aimed at estimating the latent ability of test takers and item-specific properties such as latent difficulty in assessments. Concretely, IRT models the probability of a participant responding correctly as a function of both a respondent's latent ability and item characteristics, such as difficulty and, depending on the model, discrimination. While it originated in psychometric research, it is increasingly used in language model evaluation to support model comparison, benchmark analysis, item selection, and efficient evaluation (Lalor et al., 2016; Sedoc & Ungar, 2020; Vania et al., 2021; Kipnis et al., 2025; Polo et al., 2024; Choi et al., 2026; Cong et al., 2026). In language model evaluation, models take the role of respondents while datapoints in a benchmark take the role of items, allowing model capabilities and the properties of benchmark items to be estimated jointly from observed responses. Yet many practical LLM evaluation settings, where a limited number of LLMs is evaluated on a large set of items, depart sharply from the sampling regime that motivated classical IRT where a small set of items is used to assess a large number of human participants (Hambleton et al., 1991; Jiang et al., 2026), as illustrated in Figure 1.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

The basis for estimating IRT models is the response matrix, a matrix storing the item-wise responses of each participant. Conventional marginal maximum-likelihood IRT treats the two axes of the response matrix (participants and their item-wise responses) asymmetrically, modeling respondent abilities as random effects while treating item parameters as fixed parameters (Bock & Aitkin, 1981; De Boeck, 2008; De Boeck et al., 2011). We refer to this conventional orientation as person-marginalized (PM) IRT. This is well suited to settings with many respondents per item, e.g. large-scale assessments such as PISA. However, it is problematic in settings where only data of a few respondents are available, since every additional benchmark item introduces new item-specific parameters that must then be estimated from the same limited model panel. Recent work identifies this few-model, many-item regime as a central challenge for IRT-based evaluation of LLMs (Jiang et al., 2026).

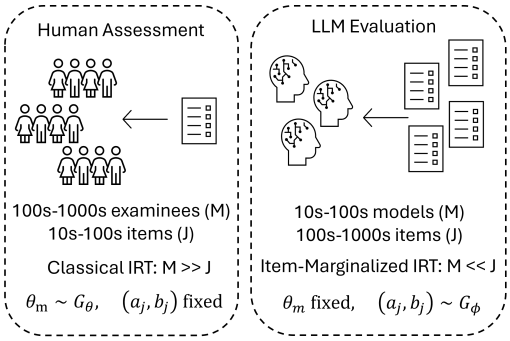


Figure 1: Contrasting response-matrix geometries in classical human assessment and item-rich LLM evaluation. Classical IRT is naturally aligned with respondent-rich settings ($M \gg J$), whereas many practical LLM evaluations involve a limited set of models and a large number of items ($M \ll J$), motivating an item-marginalized formulation.

In this work, we therefore study the opposite marginalization direction. Following the fixed-person, random-item perspective in random-item IRT (De Boeck, 2008), we treat the evaluated model abilities as fixed inferential targets and benchmark items as draws from a population distribution. The resulting item-marginalized (IM) estimator pools information across items, allowing additional benchmark items to provide repeated measurements of the same finite set of model abilities. Across simulations calibrated using real benchmark data (Kipnis et al., 2025), we find a systematic crossover. PM IRT performs better when models are abundant, the two approaches are similar near balanced response matrices, and IM increasingly outperforms PM as the evaluation becomes item-rich. In an evaluation setting with only 12 models, IM reduces Kendall ranking loss by 56% on average relative to PM. Mechanism experiments show that, with few models PM spuriously concentrates Fisher information on a small subset of items and overfits the realized response matrix, whereas IM remains closer to the population optimum and achieves lower expected response negative log-likelihood. Robustness experiments further show that the crossover persists across IRT models, and strongly non-Gaussian latent populations.

The geometry of the response matrix also matters for efficient benchmark construction. IRT can help with selecting an appropriate set of items for constructing a benchmark. Existing IRT-based compression methods can substantially reduce the number of items required for future model evaluation, but they often rely on a large historical response matrix for initial item calibration (Truong et al., 2025; Polo et al., 2024; Kipnis et al., 2025; Wang et al., 2025; Zhang et al., 2026a). In MetaBench, for example, item calibration draws on item-level responses from more than 5,000 models (Kipnis et al., 2025). We instead ask how reliably a benchmark can be constructed when only a small model panel is available. With 12–40 calibration models, IM-based item selection achieves a mean held-out Kendall's τ of 0.919, reducing held-out ranking loss by 18% relative to PM. On real sparse response matrices, IM also improves ranking recovery over conventional PM when models share only a small anchor set and answer partially overlapping subsets of the remaining benchmark. These results suggest that item marginalization can reduce the dependence of LLM evaluation on large warm-start response matrices.

2 Background

2.1 IRT for LLM Evaluation

Item Response Theory (IRT) is a family of latent-variable models for jointly characterizing respondents and test items (van der Linden, 2016). IRT has a growing history in NLP and language-model evaluation, including evaluation-scale construction (Lalor et al., 2016), efficient chatbot evaluation (Sedoc & Ungar, 2020), test-set analysis (Vania et al., 2021), benchmark compression (Kipnis et al.,

2025; Polo et al., 2024), LLM-as-a-judge reliability analysis (Choi et al., 2026), and educational assessment with LLMs (Cong et al., 2026).

Conventional IRT estimation is typically deployed in person-rich settings, with many respondents relative to the number of items (Hambleton et al., 1991; De Ayala, 2009). LLM benchmarking often exhibits the opposite geometry, with relatively few models evaluated on hundreds or thousands of items (Zhou et al., 2026; Zhang et al., 2026b). Jiang et al. (2026) show that this mismatch can substantially degrade parameter recovery and downstream ranking inferences. However, their study is primarily diagnostic and does not modify the estimation framework for this reversed geometry.

2.2 Marginal Maximum Likelihood

Marginal maximum likelihood (MML) is a standard approach to IRT estimation (Bock & Aitkin, 1981; Chalmers, 2012). Conventional MML treats person abilities as draws from a population distribution, integrates them out, and estimates item parameters as fixed effects. This person-marginalized formulation is well matched to the classical person-rich setting.

Likelihood-based applications of conventional IRT to LLM evaluation generally retain this person-marginalized orientation (Kipnis et al., 2025; Lalor et al., 2016; Li et al., 2026b; Cong et al., 2026). This orientation is also embedded in widely used IRT software such as `mirt`, whose standard marginal maximum-likelihood estimator integrates over the latent respondent distribution (Chalmers, 2012). When the respondent panel is small while the number of items grows, however, this formulation introduces an increasing number of item-specific parameters, each informed by only the few available responses, a setting closely related to the classical incidental-parameter problem (Neyman & Scott, 1948; Lancaster, 2000). To our knowledge, transposing the marginalized axis within MML has not been systematically studied for LLM evaluation.

2.3 Random-Item IRT

De Boeck (2008) develops random-item IRT and considers fixed-person, random-item models for cases where inference concerns a specific set of persons but should generalize over a broader population of items. This perspective is closely related to broader mixed effects and generalizability formulations of measurement, in which item effects may themselves be random and person and item effects can be modeled jointly (Van den Noortgate et al., 2003; De Boeck & Wilson, 2004; Brennan, 1992).

We argue that LLM evaluation provides a particularly fitting setting for this random-item perspective. In LLM evaluation, the evaluated models are typically from a small, concrete set, whereas benchmark items are numerous and can be viewed as samples from a broader population of admissible evaluation benchmarks. This suggests treating model abilities as finite inferential targets while modeling items as a population.

3 Method

3.1 Person-marginalized IRT for LLM Evaluation

Consider a response matrix $\mathbf{Y} = \{y_{mj}\}_{m=1,j=1}^{M,J}$, where M models are evaluated on J items. Let θ_m denote the latent ability of model m, and let $\boldsymbol{\phi}_j$ denote the parameters of item j. A broad class of IRT models can be written as

$$p(y_{mj} = 1 \mid \theta_m, \boldsymbol{\phi}_j) = f(\theta_m, \boldsymbol{\phi}_j), \tag{1}$$

where the choice of response function f yields, for example, the Rasch (Rasch, 1960), 2PL (Lord & Novick, 2008), 3PL (Lord, 1980), or 4PL (Barton & Lord, 1981) model.

In classical human assessment, the typical geometry of Y is $M \gg J$, based on many examinees responding to a comparatively small set of items. Conventional marginal maximum-likelihood IRT therefore models abilities as draws from a population distribution,

$$\theta_m \overset{\text{iid}}{\sim} G_\theta, \tag{2}$$

while treating the item parameters $\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_J$ as fixed effects.

The corresponding person-marginalized likelihood is

$$\ell_{\mathrm{PM}}(\boldsymbol{\phi}, \boldsymbol{\eta}) = \sum_{m=1}^{M} \log \int \prod_{j=1}^{J} p(y_{mj} \mid \theta, \boldsymbol{\phi}_j)\, dG_\theta(\theta; \boldsymbol{\eta}). \tag{3}$$

Modern LLM evaluation often exhibits the opposite geometry. A relatively small set of models is evaluated on hundreds or thousands of benchmark items, such that $M \ll J$. In the item-rich regime $J \to \infty$ with M fixed, conventional PM introduces $O(J)$ item-specific nuisance parameters whose per-item information remains bounded, whereas the model abilities remain a fixed-dimensional inferential target. We therefore reverse the conventional random-effects structure. We treat the finite collection of model abilities $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_M)$ as the inferential targets, and model items as independent draws from an item population,

$$\boldsymbol{\phi}_j \overset{\text{iid}}{\sim} G_\phi(\cdot; \boldsymbol{\eta}). \tag{4}$$

Marginalizing over the item parameters gives

$$\ell_{\mathrm{IM}}(\boldsymbol{\theta}, \boldsymbol{\eta}) = \sum_{j=1}^{J} \log \int \prod_{m=1}^{M} p(y_{mj} \mid \theta_m, \boldsymbol{\phi})\, dG_\phi(\boldsymbol{\phi}; \boldsymbol{\eta}). \tag{5}$$

We refer to this formulation as *item-marginalized IRT*. Equation 3 and equation 5 make the symmetry explicit. PM marginalizes over the model axis, whereas IM marginalizes over the item axis. Under IM, each additional item contributes another independent marginal likelihood term for the same finite set of model abilities, whereas PM introduces new item-specific parameters as J grows.

3.2 2PL Instantiation

We instantiate the general formulation using the two-parameter logistic (2PL) response model,

$$p(y_{mj} = 1 \mid \theta_m, a_j, b_j) = \sigma(a_j(\theta_m - b_j))\,, \tag{6}$$

where $b_j \in \mathbb{R}$ denotes item difficulty, $a_j > 0$ item discrimination, and $\sigma(x) = (1 + e^{-x})^{-1}$. We parameterize discrimination as $\alpha_j = \log a_j$ and define

$$\boldsymbol{\phi}_j = (b_j, \alpha_j)^\top. \tag{7}$$

For IM-2PL, the item population is modeled as

$$\boldsymbol{\phi}_j \overset{\text{iid}}{\sim} \mathcal{N}(\mathbf{0}, \Sigma_\phi), \tag{8}$$

with

$$\Sigma_\phi = \begin{pmatrix} \sigma_b^2 & \rho\sigma_b\sigma_\alpha \\ \rho\sigma_b\sigma_\alpha & \sigma_\alpha^2 \end{pmatrix}. \tag{9}$$

The zero means identify the location and scale of the latent model, while σ_b, σ_α, and ρ characterize variation in item difficulty, discrimination, and their dependence.

Substituting equation 8 into equation 5, we jointly estimate

$$\left(\widehat{\boldsymbol{\theta}}, \widehat{\sigma}_b, \widehat{\sigma}_\alpha, \widehat{\rho}\right) = \underset{\boldsymbol{\theta}, \sigma_b, \sigma_\alpha, \rho}{\arg\max}\, \ell_{\mathrm{IM}}. \tag{10}$$

The estimated abilities $\widehat{\boldsymbol{\theta}}$ induce the model ranking.

We evaluate the marginal likelihood using Gauss–Hermite quadrature, with optimization details provided in Appendix A. For sparse response matrices, both likelihoods include only observed model–item pairs. We focus on the two one-sided marginalizations because a fully crossed random-effects formulation, with both models and items treated as random, loses the low-dimensional factorization exploited by PM and IM and requires substantially more expensive approximate inference at benchmark scale. We additionally evaluate 3PL and 4PL variants in Appendix B, a two-component Gaussian mixing distribution in Appendix C, matched variational-inference implementations of PM and IM in Appendix D, statistical significance in Appendix E, and strongly non-Gaussian latent populations in Appendix F.

3.3 EVALUATION

Our evaluation addresses three complementary questions. First, we examine whether the relative performance of PM and IM is governed by the response-matrix geometry that motivates our formulation. Second, we test whether IM improves the identification of informative items for efficient benchmark construction, a canonical downstream use of IRT. Finally, we consider cold-start evaluation with no pre-calibrated item bank or large historical response matrix, testing whether IM can recover model rankings from a small model panel and sparse, partially overlapping responses.

We use the item-level response matrices released with MetaBench (Kipnis et al., 2025; Kipnis, 2024), originally collected from the Open LLM Leaderboard. The data cover ARC (Clark et al., 2018), GSM8K (Cobbe et al., 2021), HellaSwag (Zellers et al., 2019), MMLU (Hendrycks et al., 2021), TruthfulQA (Lin et al., 2022), and WinoGrande (Sakaguchi et al., 2019). For the first two experiments, we fit benchmark-specific population distributions to the real response matrices and use them to generate controlled benchmark-calibrated simulations. The calibration procedure and resulting population fits are reported in Appendix G. This preserves the empirical ability and item distributions of each benchmark while allowing us to vary the response-matrix geometry and evaluate against known latent ground truth. The third experiment instead operates directly on MetaBench response matrices. We evaluate model-ranking recovery using Kendall's τ_b (denoted τ for brevity) and item-parameter recovery using Pearson correlation, and define ranking loss as $1 - \tau$. Experimental details are provided in Appendix H.

Response-matrix geometry. Using benchmark-calibrated simulations, we vary (M, J) from strongly model rich to strongly item rich regimes while keeping $MJ \approx 48{,}000$. Model abilities and joint $(b, \log a)$ parameters are resampled from benchmark-specific empirical distributions. We compare IM and PM in parameter and ranking recovery across 50 replications.

Efficient benchmark construction. Using $M \in \{12, 24, 40, 120, 219, 400\}$ calibration models, we fit IM and PM and retain the top 10% of items by estimated Fisher information, with random selection as a baseline. Held-out models are ranked by raw accuracy on the selected subset and compared with their full-pool raw-accuracy ranking.

Cold-start sparse model evaluation. On real MetaBench responses, we sample panels of $M = 24$ models. Each model answers 5% of the full item pool and a limited number of additional items under a balanced sparse design in which each fresh item is observed by $r = 3$ models. We compare the resulting IM and PM rankings, together with a RAW baseline that ranks models by mean accuracy over their observed responses, against the full-benchmark ranking.

4 RESULTS

4.1 RESPONSE-MATRIX GEOMETRY

Figure 2 shows a consistent geometry dependent shift in the relative performance of IM and PM. When the response matrix is model rich, PM is generally competitive or better, whereas the advantage progressively shifts toward IM as J/M increases. Around the balanced regime ($J \approx M$), the two estimators perform similarly, while in strongly item-rich settings IM yields higher model-ranking and item parameter recovery. The geometry trend is statistically significant for all six benchmark-calibrated populations (Appendix E).

Table 1: Mean fitting time (s) across six datasets.

M	J	PM	IM
4,000	12	**0.365**	4.083
2,000	24	**0.375**	1.506
1,200	40	**0.386**	1.037
400	120	**0.416**	0.442
219	219	0.445	**0.228**
120	400	0.545	**0.252**
40	1,200	0.896	**0.296**
24	2,000	1.099	**0.392**
12	4,000	1.337	**0.784**

The trend is clearest for model ranking and item difficulty, whose paired differences increase smoothly with $\log_{10}(J/M)$. Log-discrimination recovery exhibits greater variability at the most extreme geometries. In the most item-rich configuration $(M, J) = (12, 4000)$, only 12 model responses are available per item, which limits the information available for estimating discrimination and leads to weaker recovery of $\log a$ even under IM. Nevertheless, the relative advantage shifts consistently from PM to IM as J/M increases.

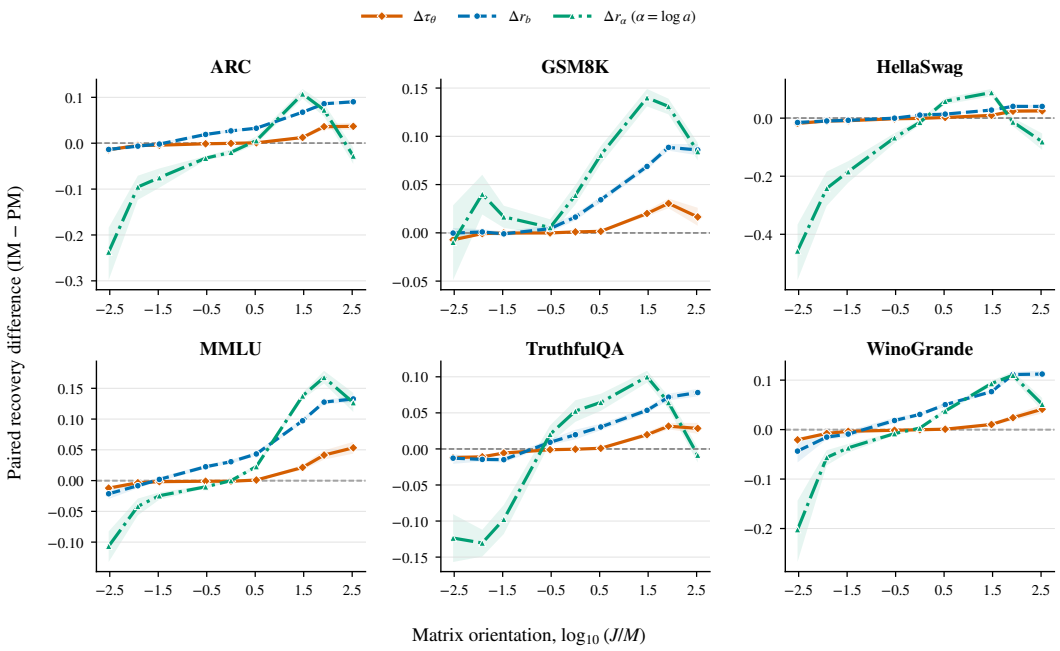


Figure 2: Effect of response-matrix geometry on parameter and ranking recovery. Each panel shows the paired performance difference between IM and PM as a function of $\log_{10}(J/M)$ for one MetaBench-calibrated simulation. We report model-ranking recovery using Kendall's τ and recovery of item-difficulty and log-discrimination using Pearson's r. Positive values favor IM and negative values favor PM.

Figure 3 provides a mechanism for this transition. With only a few responses per item, PM assigns highly concentrated information to a small subset of items, although these estimates exhibit substantially higher population negative log-likelihood. The loss landscapes show the same effect geometrically. PM moves toward a sample-specific optimum of the realized response matrix that lies far from the population optimum, whereas IM sacrifices some realized fit while remaining much closer to the population loss minimum. This suggests that item marginalization reduces the influence of poorly identified, sample-specific item effects on model ability estimation.

The computational advantage shifts with the same geometry. PM is faster when $M \gg J$, while IM becomes faster around the balanced regime and remains faster throughout the item-rich settings in which its statistical advantage is largest (Table 1).

4.2 Efficient benchmark construction

Figure 4 shows a clear difference in calibration efficiency between IM and PM. With only 12–40 calibration models, IM maintains strong held-out ranking preservation across benchmarks, whereas PM is more sensitive to calibration-panel size. In this regime, IM-based item selection achieves a mean held-out Kendall's τ of 0.919 and reduces ranking loss by 18% relative to PM. This advantage holds in 35 of 36 benchmark–calibration conditions (Appendix E). With 40 models, IM already approaches or exceeds PM trained on hundreds of calibration models on several benchmarks. As M increases, PM catches up and the difference narrows.

The PM curves also exhibit noticeable non-monotonicity at small M, most prominently on HellaSwag, where held-out ranking preservation drops sharply at $M = 40$ before re-

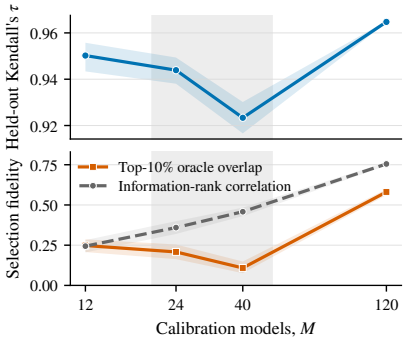


Figure 5: PM selection diagnostics on HellaSwag.

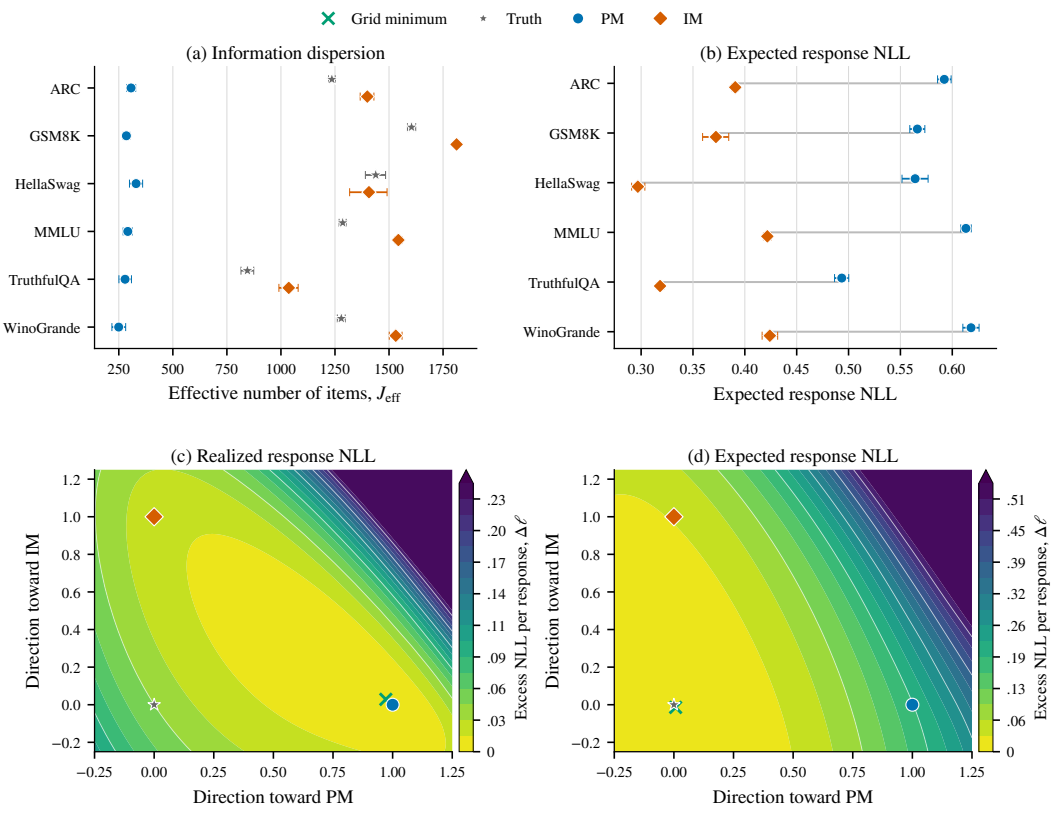


Figure 3: Item marginalization limits spurious information concentration in the item-rich regime. Top: across six MetaBench-calibrated simulations with $M = 24$ and $J = 2{,}000$, PM concentrates estimated item information and incurs higher population expected negative log-likelihood (NLL), whereas IM more closely matches the generating population. Information concentration is summarized by $J_{\text{eff}} = (\sum_j \overline{\mathcal{I}}_j)^2 / \sum_j \overline{\mathcal{I}}_j^2$, where $\overline{\mathcal{I}}_j$ is the average Fisher information of item j. Bottom: conditional-NLL landscapes for a representative ARC replication on the two-dimensional plane defined by the directions from the generating parameters to the linked PM and IM estimates, placing truth, PM, and IM at $(0, 0)$, $(1, 0)$, and $(0, 1)$, respectively. Top-panel results average 30 replications per benchmark; error bars show 95% bootstrap confidence intervals.

covering at $M = 120$. To better understand this behavior, Figure 5 examines PM item selection on HellaSwag. Agreement with the ground-truth information ranking improves steadily as more calibration models are added, yet overlap with the oracle top-10% item set falls from approximately 0.25 at $M = 12$ to 0.11 at $M = 40$. The held-out ranking performance deteriorates at the same point and recovers once the top-10% overlap improves at $M = 120$. This suggests that the dip is not driven by a deterioration of the overall information ranking, but by instability in identifying the highly informative tail used for top-K selection. Even as the global item ranking improves, estimation errors around the selection boundary can substantially change which items enter the distilled benchmark.

4.3 Cold-start sparse model evaluation

We next use $M = 24$ real models and a shared 5% anchor set. We vary the additional sparse-response budget from the anchor-only setting to the maximum feasible budget, with intermediate levels at 10%, 25%, 50%, and 75%. Across the six benchmarks, the maximum budget ranges from 76 to 1,485 additional responses per model.

As shown in Figure 6, ranking recovery generally improves as additional sparse responses become available. IM consistently outperforms PM once additional responses are introduced, across all

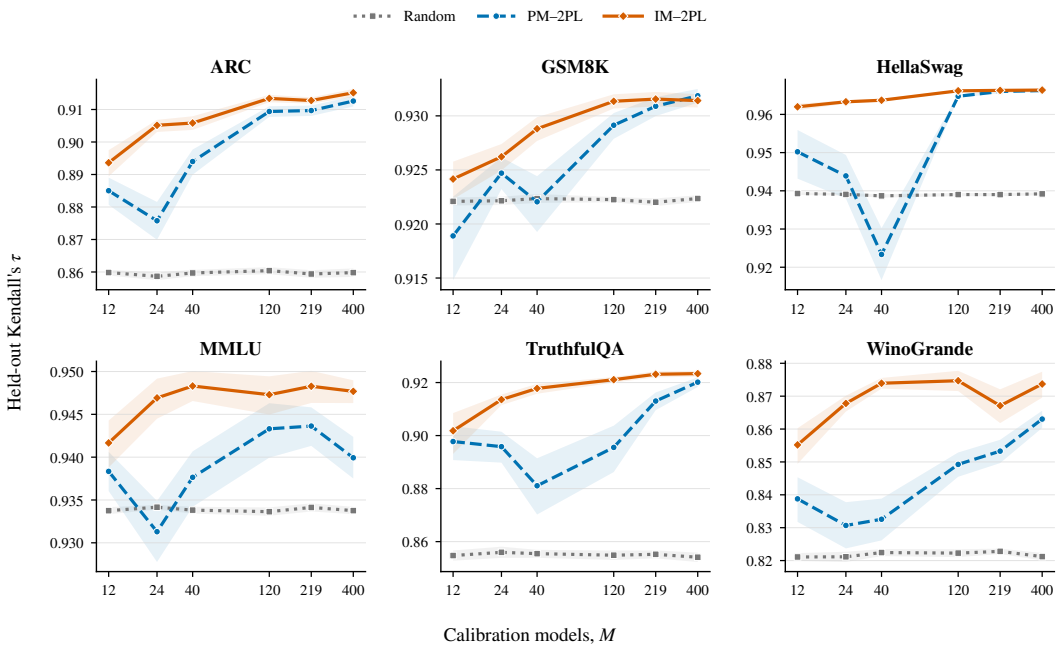


Figure 4: Efficient benchmark construction under limited calibration data. For each benchmark, IM and PM are fitted using M calibration models and select the top 10% of items by estimated average Fisher information. Ranking preservation is evaluated on 4,000 independently sampled test models using Kendall's τ between the selected-subset and full-item-pool rankings. Random subsets of the same size serve as a baseline.

30 nonzero-budget conditions (Appendix E), with particularly clear gains on GSM8K, HellaSwag, MMLU, and TruthfulQA. The advantage remains visible even at larger budgets, indicating that the benefit of item marginalization is not confined to the most sparse setting.

Compared with naive averaging over the observed responses, IM is competitive and often superior at moderate to large budgets, outperforming raw scoring in 21 of 30 conditions (Appendix E), although raw scoring remains strong on several benchmarks. Because the observation pattern is randomized and balanced, raw averaging is design-unbiased for the full-benchmark mean score in expectation over the assignment, which partly explains its strong performance in this controlled setting. This is consistent with MetaBench, which found that even random item subsampling can retain substantial information about full-benchmark performance (Kipnis et al., 2025). Overall, the results show that IM can recover reliable model rankings from a small model panel and partially overlapping responses without requiring a pre-calibrated item bank or dense historical response matrix.

5 Discussion

Our results suggest that marginalization direction should be viewed as part of IRT estimation design. Conventional MML is naturally aligned with settings where respondents are sampled from a population and items are fixed (Bock & Aitkin, 1981; Holland, 1990), whereas random-item IRT applies when inference concerns a finite set of respondents and measurements generalize over a population of admissible items (De Boeck, 2008; Briggs & Wilson, 2007). LLM evaluation often more closely resembles the latter setting, with a small model panel and a large benchmark item pool (Keller et al., 2026). The observed crossover therefore suggests that the consequences of marginalization depend on response-matrix geometry.

When M is small, PM introduces more item-specific parameters as J grows while information per item remains bounded, creating a tension related to the incidental parameter problem (Neyman & Scott, 1948; Lancaster, 2000). The analyses in the appendix further show that this geometry effect is not specific to a particular population assumption or estimation procedure. A more flexible ability

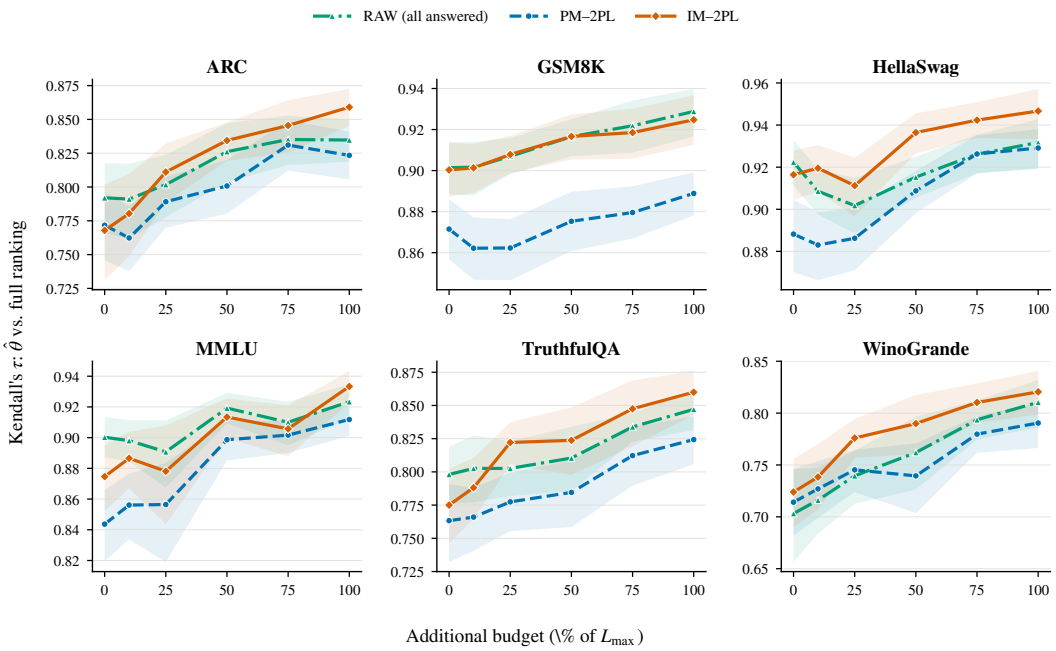


Figure 6: Ranking recovery under sparse evaluation. We sample $M = 24$ models and use 5% of each benchmark as shared anchor items. The remaining evaluation budget is expressed as a fraction of the maximum feasible per-model budget $L_{\max}$ under a balanced design in which each additional item is answered by $r = 3$ models. We report Kendall's τ between the estimated ranking and the full-benchmark score ranking. Shaded regions denote 95% confidence intervals over replications.

distribution attenuates but does not eliminate the crossover (Appendix C), and the same qualitative pattern persists under matched variational inference, alternative response models, and strongly non-Gaussian latent populations (Appendices B, D, and F). However, discrimination remains difficult to estimate when few models answer each item, and PM can be preferable when many models are available or discrimination is the primary inferential target.

This distinction also clarifies the role of IM in efficient evaluation. Compact benchmarks can be constructed through IRT, adaptive testing, learned prediction, and direct subset optimization (Vivek et al., 2024; Polo et al., 2024; Truong et al., 2025; Wang et al., 2025; Li et al., 2026b; Zhang et al., 2026a), including recent methods using only a small number of source models (Zhang et al., 2026b; Das et al., 2026). Our contribution is complementary to improving the ultimate compression ratio. We show that changing the marginalization direction can substantially reduce the amount of model-response data required for reliable IRT-based selection. The cold-start experiment extends this result to partially overlapping response matrices. Although naive scoring remains competitive on several benchmarks (Kipnis et al., 2025), IM substantially improves the reliability of IRT-based ranking under sparse historical response data.

Several limitations define the scope of these conclusions. Our experiments use six binary-response benchmarks and primarily isolate marginalization direction within a unidimensional IRT framework. Real evaluation may also involve continuously refreshed item pools (Jain et al., 2025), multidimensional or context-dependent item characteristics (Shang et al., 2026), language-specific effects (Lior et al., 2026), and subjective or pairwise observations (Chiang et al., 2024; Li et al., 2026a). Extending item marginalization to hierarchical or context-conditioned item populations and more general observation and missingness models is therefore an important direction for future work. Overall, our results argue against treating person marginalization as an invariant default. PM remains well suited to model-rich settings and item-focused inference, whereas IM is better aligned with settings in which a finite model panel is evaluated through repeated samples from a broader item population. The appropriate marginalization direction should therefore reflect both the geometry of the response matrix and the intended direction of generalization.

AI USE STATEMENT

We have not used generative AI tools for generating synthetic datasets, helping develop theoretical models or conceptual frameworks, formulating mathematical claims, proposing or refining hypotheses, designing or providing feedback on research methodology or experiments, implementing methods, cleaning and reformatting datasets, or interpreting results. Providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, assisting with translation, and supporting qualitative and thematic data analysis are not applicable to this work. Additionally, we used generative AI tools to create or edit software code and to edit a research paper to improve readability. We have reviewed all AI-assisted work. LLM-generated code was verified and tested for correctness. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We provide the methodological and experimental details needed to reproduce our results in the main text and appendix, including the estimation procedures, data construction, sampling schemes, experimental configurations, evaluation metrics, and implementation details. The supplementary material contains the scripts used to run the experiments and generate the reported results.

REFERENCES

Mark A Barton and Frederic M Lord. An upper asymptote for the three-parameter logistic item-response model. *ETS Research Report Series*, 1981(1):i–8, 1981.

R. Darrell Bock and Murray Aitkin. Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, 46(4):443–459, 1981. doi: 10.1007/BF02293801.

Robert L. Brennan. Generalizability theory. *Educational Measurement: Issues and Practice*, 11(4):27–34, 1992. doi: https://doi.org/10.1111/j.1745-3992.1992.tb00260.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1745-3992.1992.tb00260.x.

Derek C Briggs and Mark Wilson. Generalizability in item response modeling. *Journal of Educational Measurement*, 44(2):131–155, 2007. doi: https://doi.org/10.1111/j.1745-3984.2007.00031.x.

R. Philip Chalmers. 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.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. Chatbot arena: An open platform for evaluating LLMs by human preference. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=3MW8GKNyzI.

Junhyuk Choi, Sohhyung Park, Chanhee Cho, Hyeonchu Park, and Bugeun Kim. Diagnosing the reliability of llm-as-a-judge via item response theory, 2026. URL https://arxiv.org/abs/2602.00521.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge, 2018. URL https://arxiv.org/abs/1803.05457.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems, 2021. URL https://arxiv.org/abs/2110.14168.

Longwei Cong, Sonja Hahn, Sebastian Gombert, Leon Camus, Hendrik Drachsler, and Ulf Kroehne. Estimating LLM grading ability and response difficulty in automatic short answer grading via item response theory. In Ekaterina Kochmar, Bashar Alhafni, Stefano Bannò, Marie Bexte, Jill Burstein, Andrea Horbach, Ronja Laarmann-Quante, Anais Tack, Victoria Yaneva, and Zheng Yuan (eds.), *Proceedings of the 21st Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2026)*, pp. 259–271, San Diego, California, USA, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-409-5. doi: 10.18653/v1/2026.bea-1.19. URL https://aclanthology.org/2026.bea-1.19/.

Devleena Das, Rajeev Patwari, Vikram Kumar Bukka, Nithin Kumar Guggilla, Elliott Delaye, and Ashish Sirasao. Mince: Shrinking llm evaluation datasets via few-model monte carlo calibration, 2026. URL https://arxiv.org/abs/2606.22826.

Rafael Jaime De Ayala. *The theory and practice of item response theory*. Guilford Press New York, NY, 2009.

Paul De Boeck. Random item irt models. *Psychometrika*, 73(4):533–559, 2008. doi: https://doi.org/10.1007/s11336-008-9092-x.

Paul De Boeck and Mark Wilson (eds.). *Explanatory item response models: A generalized linear and nonlinear approach*. Springer, 2004. doi: https://doi.org/10.1007/978-1-4757-3990-9.

Paul De Boeck, Marjan Bakker, Robert Zwitser, Michel Nivard, Abe Hofman, Francis Tuerlinckx, and Ivailo Partchev. The estimation of item response models with the lmer function from the lme4 package in r. *Journal of Statistical Software*, 39(12):1–28, 2011. doi: 10.18637/jss.v039.i12. URL https://www.jstatsoft.org/index.php/jss/article/view/v039i12.

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

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding, 2021. URL https://arxiv.org/abs/2009.03300.

Paul W Holland. On the sampling theory foundations of item response theory models. *Psychometrika*, 55(4):577–601, 1990.

Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. Livecodebench: Holistic and contamination free evaluation of large language models for code. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=chfJJYC3iL.

Han Jiang, Sunbeom Kwon, Jinwen Luo, Ziang Xiao, and Susu Zhang. Can we trust item response theory for ai evaluation?, 2026. URL https://arxiv.org/abs/2607.15190.

Andrew Keller, Kweku Kwegyir-Aggrey, Ryan Steed, Anita Rao, Julia Sharp, and Amanda Bergman. Expanding the ai evaluation toolbox with statistical models, 2026. URL https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=961314.

Alex Kipnis. metabench - paper data, July 2024. URL https://doi.org/10.5281/zenodo.12819251.

Alex Kipnis, Konstantinos Voudouris, Luca Schulze Buschoff, and Eric Schulz. metabench - a sparse benchmark of reasoning and knowledge in large language models. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 31734–31770, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/4ebc26584810a189ef1e4f173aba4319-Paper-Conference.pdf.

John P. Lalor, Hao Wu, and Hong Yu. Building an evaluation scale using item response theory. In Jian Su, Kevin Duh, and Xavier Carreras (eds.), *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 648–657, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1062. URL https://aclanthology.org/D16-1062/.

John Patrick Lalor and Pedro Rodriguez. py-irt: A scalable item response theory library for python. *INFORMS Journal on Computing*, 35(1):5–13, January 2023. ISSN 1526-5528. doi: 10.1287/ijoc.2022.1250. URL http://dx.doi.org/10.1287/ijoc.2022.1250.

Tony Lancaster. The incidental parameter problem since 1948. *Journal of Econometrics*, 95(2):391–413, 2000. ISSN 0304-4076. doi: https://doi.org/10.1016/S0304-4076(99)00044-5. URL https://www.sciencedirect.com/science/article/pii/S0304407699000445.

Aaron J. Li, Hao Huang, Youngmin Park, Yitong Ma, Wei-Lin Chiang, Li Chen, Cho-Jui Hsieh, Bin Yu, and Ion Stoica. Dualeval: Joint model-item calibration for unified llm evaluation, 2026a. URL https://arxiv.org/abs/2606.26429.

Peiyu Li, Xiuxiu Tang, Si Chen, Ying Cheng, Ronald Metoyer, Ting Hua, and Nitesh V. Chawla. Adaptive testing for llm evaluation: A psychometric alternative to static benchmarks, 2026b. URL https://arxiv.org/abs/2511.04689.

Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3214–3252, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.229. URL https://aclanthology.org/2022.acl-long.229/.

Gili Lior, Tzviel Frostig, Gabriel Stanovsky, and Matan Eyal. Extending item response theory for efficient and meaningful multilingual evaluation, 2026. URL https://arxiv.org/abs/2606.15643.

Frederic M Lord. *Applications of item response theory to practical testing problems*. Routledge, 1980.

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

Jerzy Neyman and Elizabeth L Scott. Consistent estimates based on partially consistent observations. *Econometrica: journal of the Econometric Society*, pp. 1–32, 1948. doi: https://doi.org/10.2307/1914288.

Felipe Maia Polo, Lucas Weber, Leshem Choshen, Yuekai Sun, Gongjun Xu, and Mikhail Yurochkin. tinybenchmarks: evaluating llms with fewer examples. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

G. Rasch. *Probabilistic Models for Some Intelligence and Attainment Tests*. Studies in mathematical psychology. Danmarks Paedagogiske Institut, 1960. ISBN 9780598554512. URL https://books.google.de/books?id=XnlevwEACAAJ.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial winograd schema challenge at scale, 2019. URL https://arxiv.org/abs/1907.10641.

João Sedoc and Lyle Ungar. Item response theory for efficient human evaluation of chatbots. In Steffen Eger, Yang Gao, Maxime Peyrard, Wei Zhao, and Eduard Hovy (eds.), *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*, pp. 21–33, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.eval4nlp-1.3. URL https://aclanthology.org/2020.eval4nlp-1.3/.

Ergan Shang, Weijing Tang, and Yinqiu He. Llm evaluation on unseen questions: Contextual multidimensional irt model, 2026. URL https://arxiv.org/abs/2608.22295.

Sang T. Truong, Yuheng Tu, Percy Liang, Bo Li, and Sanmi Koyejo. Reliable and efficient amortized model-based evaluation. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 60238–60265. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/truong25c.html.

Wim Van den Noortgate, Paul De Boeck, and Michel Meulders. Cross-classification multilevel logistic models in psychometrics. *Journal of Educational and Behavioral Statistics*, 28(4): 369–386, 2003. doi: 10.3102/10769986028004369. URL https://doi.org/10.3102/10769986028004369.

W.J. van der Linden (ed.). *Handbook of Item Response Theory: Volume 1: Models*. Behavioral Sciences. Chapman and Hall/CRC, 2016. ISBN 9781315374512. doi: 10.1201/9781315374512.

Clara Vania, Phu Mon Htut, William Huang, Dhara Mungra, Richard Yuanzhe Pang, Jason Phang, Haokun Liu, Kyunghyun Cho, and Samuel R. Bowman. Comparing test sets with item response theory. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 1141–1158, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.92. URL https://aclanthology.org/2021.acl-long.92/.

Rajan Vivek, Kawin Ethayarajh, Diyi Yang, and Douwe Kiela. Anchor points: Benchmarking models with much fewer examples. In Yvette Graham and Matthew Purver (eds.), *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1576–1601, St. Julian's, Malta, March 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.eacl-long.95. URL https://aclanthology.org/2024.eacl-long.95/.

Shaobo Wang, Cong Wang, Wenjie Fu, Yue Min, Mingquan Feng, Isabel Guan, Xuming Hu, Conghui He, Cunxiang Wang, Kexin Yang, Xingzhang Ren, Fei Huang, Dayiheng Liu, and Linfeng Zhang. Rethinking llm evaluation: Can we evaluate llms with 200x less data?, 2025. URL https://arxiv.org/abs/2510.10457.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: Can a machine really finish your sentence? In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4791–4800, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1472. URL https://aclanthology.org/P19-1472/.

Taolin Zhang, Hang Guo, Wang Lu, Tao Dai, Shu-Tao Xia, and Jindong Wang. Sparseeval: Efficient evaluation of large language models by sparse optimization, 2026a. URL https://arxiv.org/abs/2602.07909.

Yueqi Zhang, Jin Hu, Shaoxiong Feng, Peiwen Yuan, Xinglin Wang, Yiwei Li, Jiayi Shi, Chuyi Tan, Ji Zhang, Boyuan Pan, Yao Hu, and Kan Li. Learning more from less: Unlocking internal representations for benchmark compression, 2026b. URL https://arxiv.org/abs/2602.00710.

Hongli Zhou, Hui Huang, Ziqing Zhao, Lvyuan Han, Huicheng Wang, Kehai Chen, Muyun Yang, Wei Bao, Jian Dong, Bing Xu, et al. Lost in benchmarks? rethinking large language model benchmarking with item response theory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 35085–35093, 2026. doi: https://doi.org/10.1609/aaai.v40i41.40814.

A Estimation Details

Identification and parameterization. For PM-2PL, we use the conventional standard-normal ability population $\theta_m \sim \mathcal{N}(0,1)$ and estimate (b_j, α_j) as fixed item-specific parameters, where $\alpha_j = \log a_j$. Model abilities are reported as posterior EAP scores after fitting. For IM-2PL, the evaluated abilities are fixed parameters and the item parameters follow

$$\begin{pmatrix} b_j \\ \alpha_j \end{pmatrix} \sim \mathcal{N}_2 \left(\mathbf{0}, \begin{pmatrix} \sigma_b^2 & \rho\sigma_b\sigma_\alpha \\ \rho\sigma_b\sigma_\alpha & \sigma_\alpha^2 \end{pmatrix} \right). \tag{11}$$

The zero means identify the latent location and scale, while σ_b, σ_α, and ρ are estimated jointly with the M model abilities. PM therefore estimates $2J$ item-specific parameters, whereas IM estimates $M + 3$ parameters in the marginal likelihood.

Marginal-likelihood evaluation. PM evaluates its one dimensional ability integral using 21-point Gauss–Hermite quadrature. IM evaluates the joint (b, α) integral using a 15×15 product Gauss–Hermite rule, with nodes transformed according to the current item population covariance matrix. All likelihood calculations are performed in the log domain using stable log-sum-exp operations. Both estimators are optimized with L-BFGS-B under bounded parameterizations and use the same response matrix in every paired comparison. For sparse response matrices, the posterior likelihood product is restricted to models for which item j is observed.

Empirical-Bayes item estimates. We recover item difficulty and discrimination from their empirical-Bayes posteriors conditional on the fitted abilities,

$$p\Big(\boldsymbol{\phi}_j \mid \mathbf{y}_{\cdot j}, \widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\eta}}\Big) \propto \left[\prod_m p\Big(y_{mj} \mid \widehat{\theta}_m, \boldsymbol{\phi}_j\Big)\right] g_\phi(\boldsymbol{\phi}_j; \widehat{\boldsymbol{\eta}})\,. \tag{12}$$

Posterior means of b_j and a_j are computed using the same quadrature representation and are used whenever item specific quantities are required. PM instead uses its directly estimated fixed item parameters.

Scale linking. For simulation-based parameter recovery, ability estimates are mapped to the generating scale using a positive affine transformation fitted between the estimated and true abilities. Item difficulties are transformed accordingly, whereas item discriminations are adjusted inversely by the scale factor. These transformations leave fitted response probabilities unchanged and, because the transformation is strictly increasing, preserve rank-based ability metrics.

Complete optimization settings, convergence diagnostics, and implementation details are provided with the supplementary code.

B 3PL AND 4PL INSTANTIATIONS

We test whether the geometry-dependent PM–IM crossover persists beyond the 2PL response model by repeating the benchmark calibration and response-matrix geometry experiment with 3PL and 4PL variants.

Benchmark calibration. For each of the six complete MetaBench datasets, we fit the native 3PL and 4PL implementations in `py-irt` 0.7.1 using full-data variational inference. The 3PL response function is

$$p_{mj} = c_j + (1 - c_j)\,\sigma[a_j(\theta_m - b_j)]\,, \qquad 0 < c_j < 1. \tag{13}$$

The model called `4pl` in `py-irt` instead uses

$$p_{mj} = u_j\,\sigma[a_j(\theta_m - b_j)]\,, \qquad 0 < u_j < 1, \tag{14}$$

which introduces an item-specific upper asymptote but no lower asymptote. We therefore refer to this specification as U4PL. For each benchmark, the fitted empirical ability distribution and the joint item-parameter rows $(b_j, \log a_j, c_j)$ or $(b_j, \log a_j, u_j)$ define the corresponding simulation population. For IM-3PL and IM-U4PL, (b_j, α_j) follows the same zero-mean correlated Gaussian population as in IM-2PL. The bounded asymptote is modeled independently on the logit scale, $\mathrm{logit}(c_j) \sim \mathcal{N}(\mu_c, \sigma_c^2)$ for 3PL and $\mathrm{logit}(u_j) \sim \mathcal{N}(\mu_u, \sigma_u^2)$ for U4PL.

Response-matrix geometry. We use the same (M, J) geometries, approximately fixed response budget $MJ \approx 48{,}000$, and 50 paired replications as in the 2PL experiment. Within each replication, abilities and complete item-parameter rows are sampled from the corresponding benchmark-calibrated population. PM and IM are fitted to the same generated response matrix and linked to the true latent scale before evaluation. PM uses 21-point Gauss–Hermite quadrature for the one-dimensional ability integral, while IM uses 512 scrambled Sobol points for the higher-dimensional item integral.

Figures 7 and 8 reproduce the orientation-dependent pattern observed under 2PL. Across the six benchmark-calibrated populations, the Spearman correlation between $\log(J/M)$ and the mean IM–PM recovery difference ranges from 0.95 to 0.98 under 3PL and from 0.95 to 1.00 under U4PL. In

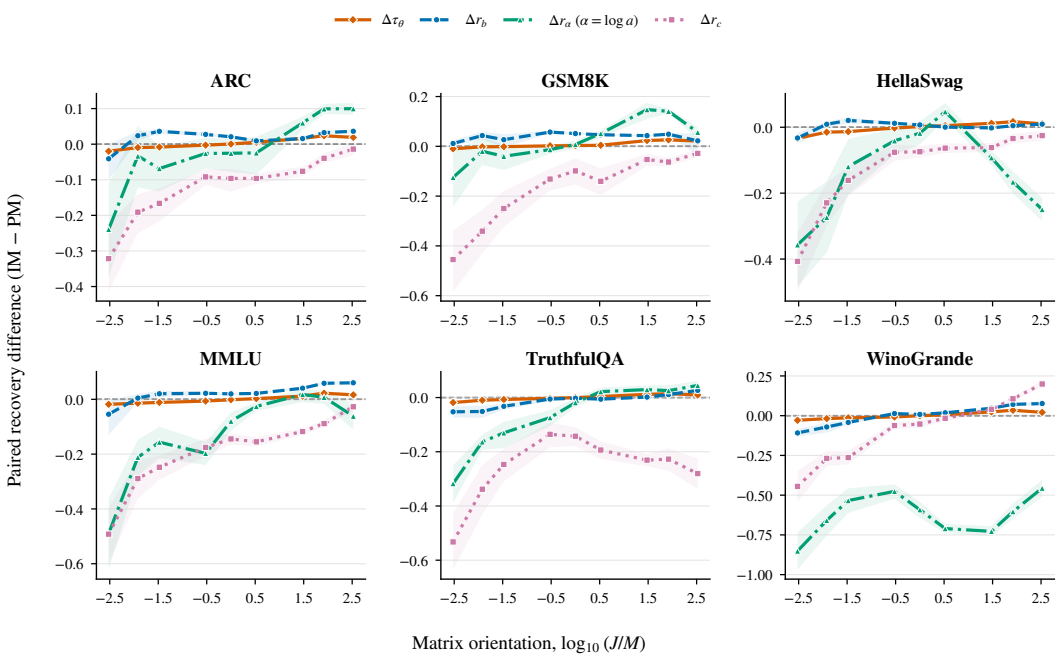


Figure 7: Paired IM–PM recovery differences under the 3PL. Curves report recovery of ability ranking and the item parameters b, $\alpha = \log a$, and c. Lines show paired means and bands show 95% bootstrap confidence intervals across 50 replications. Positive values favor IM.

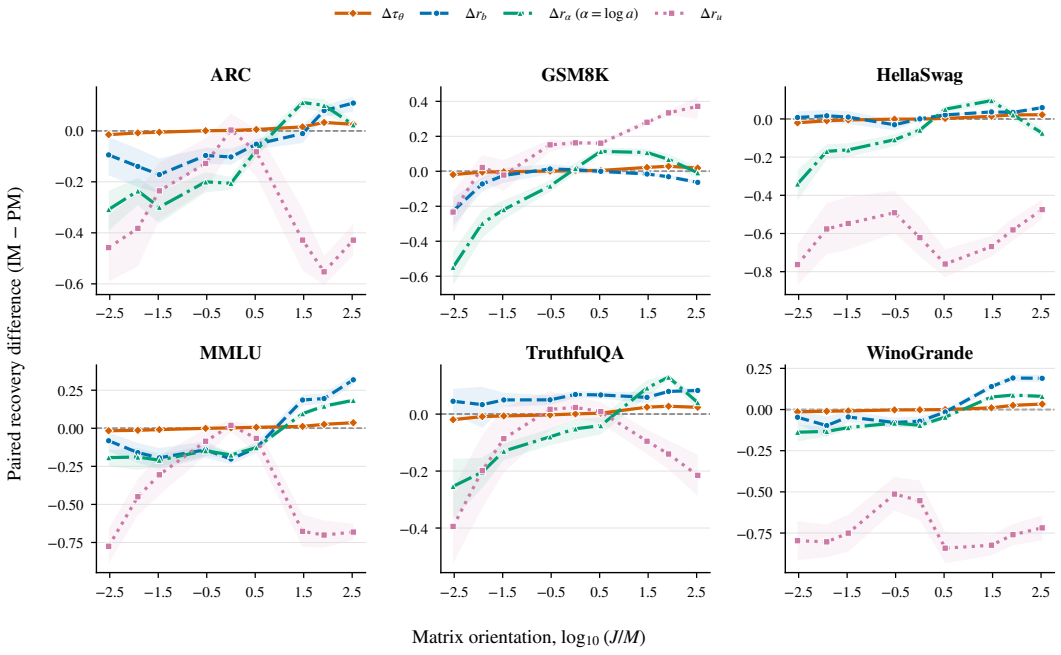


Figure 8: Paired IM–PM recovery differences under the U4PL, with the upper asymptote u replacing the 3PL lower asymptote c. Lines show paired means and bands show 95% bootstrap confidence intervals across 50 replications. Positive values favor IM.

the item-rich regime with $M \leq 40$, IM improves Kendall's τ by 0.019 on average under 3PL and by 0.023 under U4PL. The difference becomes small near the balanced geometry and reverses in the

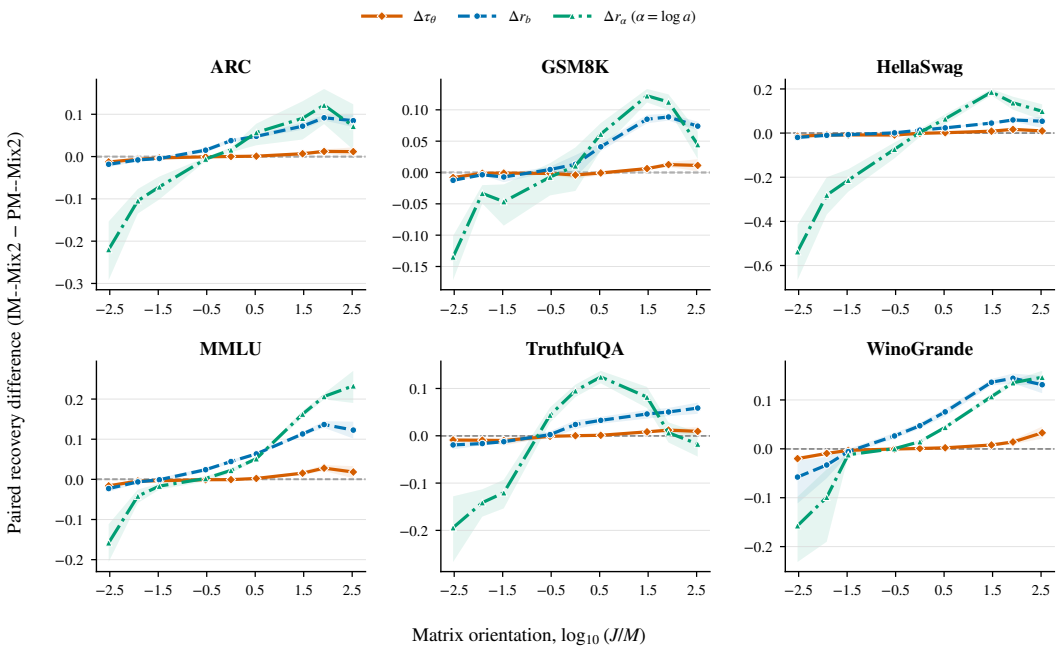


Figure 9: Paired IM–Mix2 minus PM–Mix2 recovery differences under two-component Gaussian mixing distributions. Curves report ability-ranking recovery and recovery of difficulty and log discrimination. Lines show paired means and bands show 95% bootstrap confidence intervals. Positive values favor IM–Mix2.

strongly model-rich regime. The same qualitative transition is also visible in recovery of the item parameters.

C Two-Component Gaussian Mixture Distributions

The main response-matrix geometry experiment uses a single Gaussian distribution on the marginalized side of the response matrix. We test whether the observed PM–IM crossover is driven by this unimodal assumption by replacing the Gaussian population with a two component mixture while leaving the response model and data-generating process unchanged. We evaluate both the relative PM–IM difference under two component Gaussian mixture distributions and the within orientation change from the single Gaussian to the mixture specification.

PM–Mix2 models the ability population as

$$F_\theta = \pi\,\mathcal{N}(\mu_1, s_1^2) + (1-\pi)\,\mathcal{N}(\mu_2, s_2^2), \tag{15}$$

with the global mixture normalized to mean zero and variance one for identification. IM–Mix2 instead models the joint item distribution as

$$\begin{pmatrix} b_j \\ \alpha_j \end{pmatrix} \sim \pi\,\mathcal{N}_2(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) + (1-\pi)\,\mathcal{N}_2(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2), \qquad \alpha_j = \log a_j, \tag{16}$$

with a separate full covariance matrix for each component. For identification, the weighted component mean is constrained to zero, $\pi\boldsymbol{\mu}_1 + (1-\pi)\boldsymbol{\mu}_2 = \mathbf{0}$, while the component ordering constraint removes label ambiguity.

We repeat the response-matrix geometry experiment on the same six MetaBench-calibrated populations, nine (M, J) geometries with $MJ \approx 48{,}000$, and 50 paired replications per condition. PM–Mix2 uses 21-point Gauss–Hermite quadrature per component, while IM–Mix2 uses a 15×15 product rule per component. The same latent-scale linking and recovery metrics are used as in the main experiment.

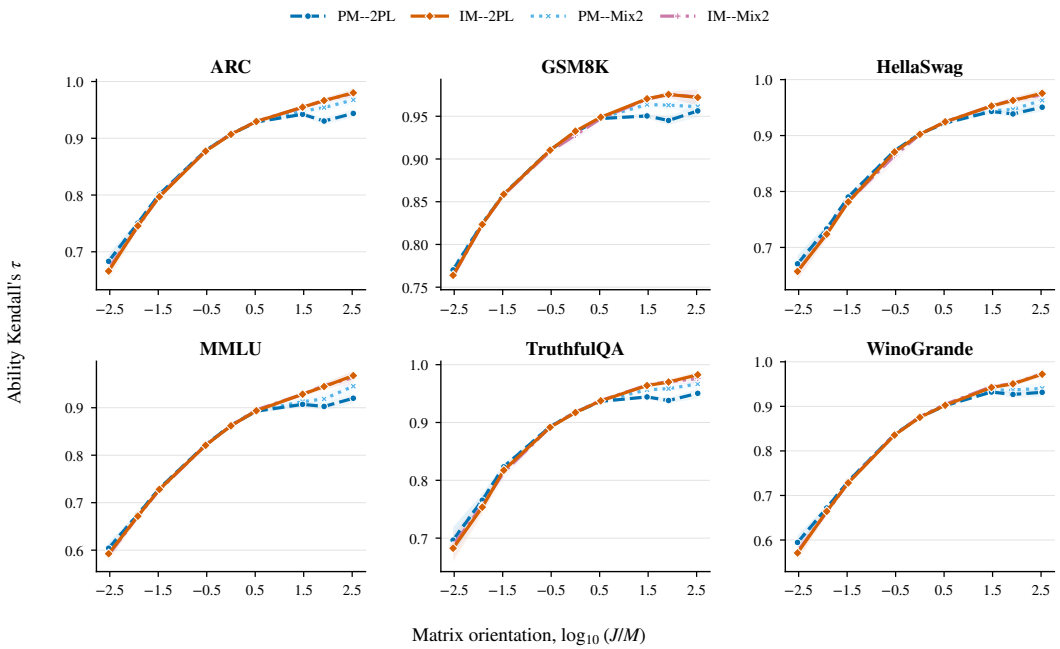


Figure 10: Ability-ranking recovery under the single-Gaussian and Mix2 estimators. All four methods are compared on identical response matrices for which every fit converged. Lines show mean Kendall's τ and bands show 95% bootstrap confidence intervals.

Figure 9 shows that allowing multimodality attenuates but does not eliminate the geometry-dependent crossover. In the item-rich regime with $M \leq 40$, IM–Mix2 improves ability Kendall's τ by 0.013 on average and has a positive mean effect in all 18 benchmark–geometry conditions. The corresponding mean improvements are 0.089 for difficulty recovery and 0.114 for log-discrimination recovery. Across the nine geometries, the benchmark-wise Spearman correlation between $\log_{10}(J/M)$ and the mean ranking advantage ranges from 0.830 to 1.000.

Figure 10 shows where the smaller ranking gap comes from. In the item-rich regime, mean Kendall's τ increases from 0.937 to 0.949 for PM when moving from a single Gaussian to Mix2, while remaining essentially unchanged at 0.962 for IM. The mean IM–PM ranking advantage therefore decreases from 0.026 to 0.013, corresponding to an approximately 49% reduction. A more flexible ability population can therefore explain a substantial part of PM's ranking deficit, but not the geometry effect itself.

Figures 11 and 12 show that the effect of mixture flexibility differs for item-parameter recovery. In the item-rich regime, Mix2 increases mean difficulty correlation from 0.836 to 0.841 for PM and from 0.915 to 0.928 for IM. For log discrimination, the corresponding changes are from 0.572 to 0.591 for PM and from 0.647 to 0.703 for IM. Mixture flexibility therefore primarily improves PM for ability ranking, but benefits IM more strongly for item recovery. The IM advantage in recovering item characteristics consequently persists, and in these experiments increases, under the more flexible population model.

Overall, the Mix2 experiment separates two sources of error. Distributional misspecification accounts for part of PM's item-rich ranking deficit, while the remaining geometry-dependent crossover persists after substantially increasing the flexibility of both marginalized populations.

D Matched Variational-Inference Comparison

The main experiments use marginal-likelihood estimators specialized to each marginalization direction. We test whether the geometry-dependent PM–IM difference depends on these specialized im-

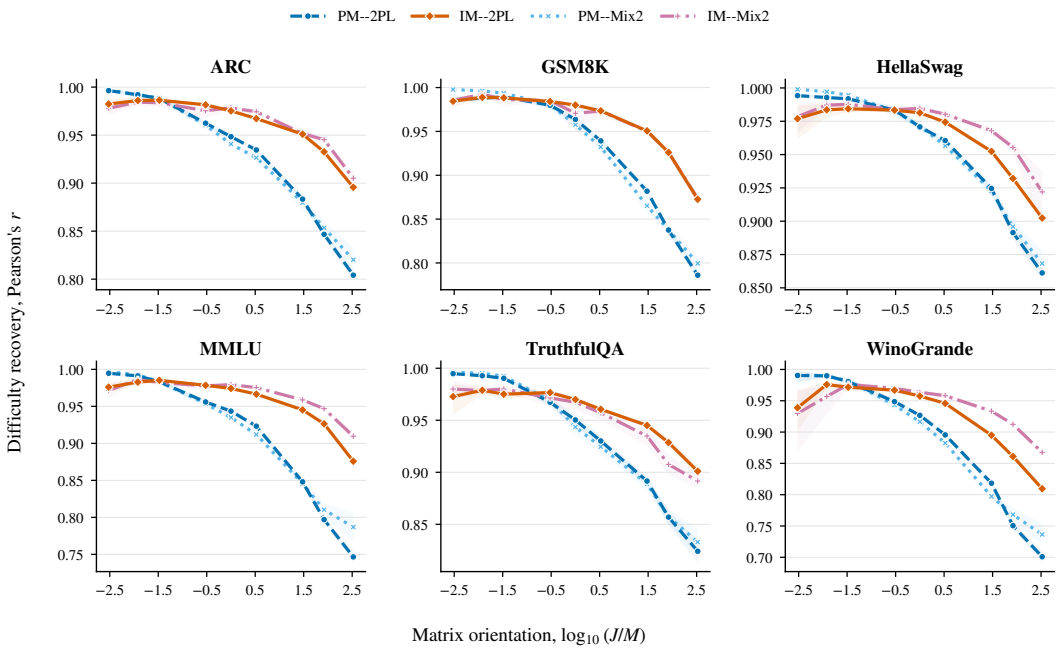


Figure 11: Difficulty recovery under the single-Gaussian and Mix2 estimators, evaluated on identical four-way complete replications.

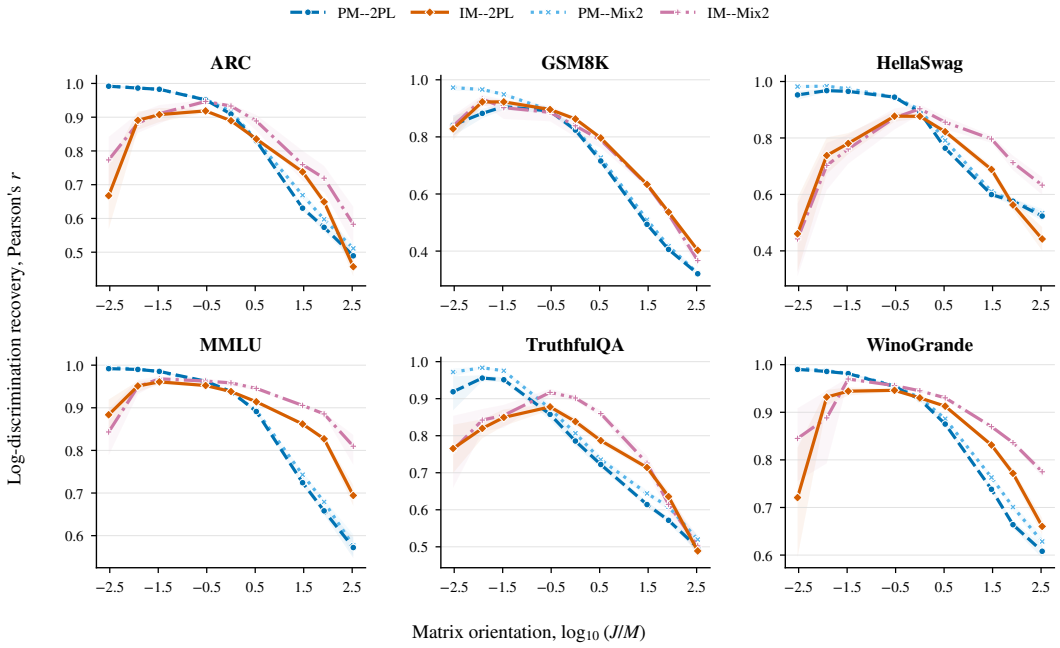


Figure 12: Log-discrimination recovery under the single-Gaussian and Mix2 estimators, evaluated on identical four-way complete replications.

plementations by constructing matched variational-inference estimators for the same 2PL response model.

Matched VI estimators. PM–VI assigns independent Gaussian variational factors to the random model abilities while treating (b_j, α_j) as item-specific point parameters under the standard-normal

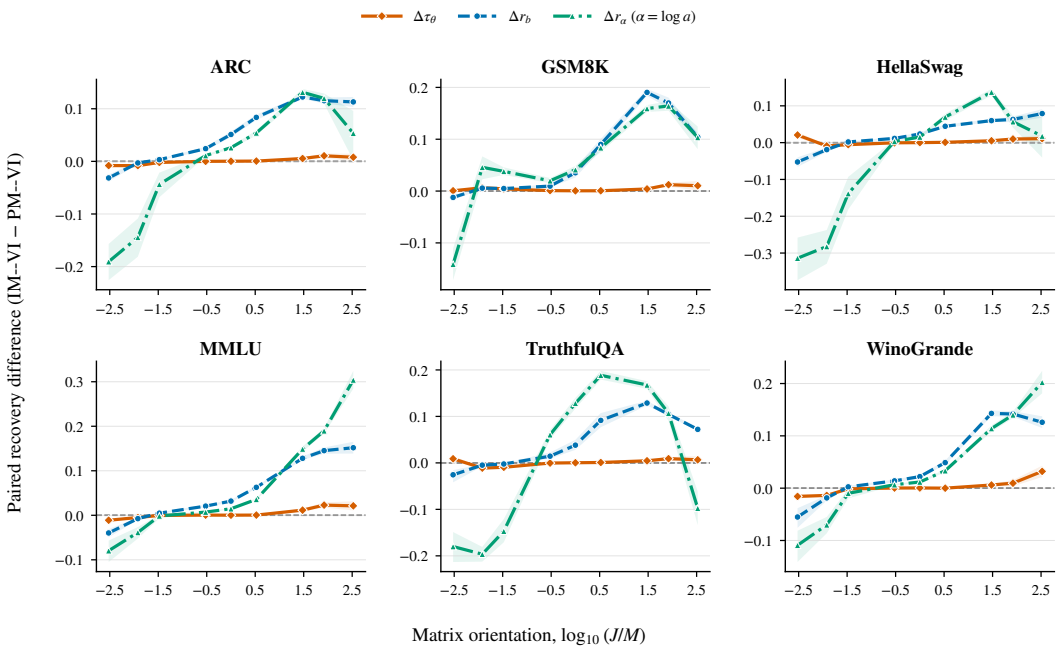


Figure 13: Matched-VI recovery differences across response-matrix geometries. Curves show paired IM–VI minus PM–VI differences in ability-ranking recovery, difficulty recovery, and log-discrimination recovery. Lines show means and bands show 95% bootstrap confidence intervals across 50 paired replications. Positive values favor IM–VI.

ability prior. IM–VI instead treats the evaluated abilities as fixed parameters and assigns each item a bivariate Gaussian variational factor over (b_j, α_j), retaining within-item posterior covariance while estimating the item-population covariance by empirical Bayes. The two estimators therefore retain the opposite marginalization directions of PM and IM while using the same stochastic optimization framework and variational-inference budget. Posterior means are used for recovery, and estimates are linked to the true latent scale using the same procedure as in the main response-matrix geometry experiment.

We repeat the experiment on the same six MetaBench-calibrated populations, nine (M, J) geometries with $MJ \approx 48{,}000$, and 50 paired replications per condition. PM–VI and IM–VI are fitted to the same generated response matrix in every replication.

Figure 13 shows that the geometry-dependent recovery pattern persists under matched variational inference. In the item-rich regime with $M \leq 40$, IM–VI improves ability Kendall's τ by 0.011 on average and outperforms PM–VI. The corresponding mean improvements are 0.120 for difficulty recovery and 0.123 for log-discrimination recovery. The ability-ranking advantage is smaller than under the specialized marginal-likelihood estimators, but remains positive on average throughout the item-rich conditions across all six benchmark-calibrated populations.

The matched-VI experiment therefore provides a complementary robustness check. Changing both estimators from specialized marginal-likelihood optimization to a common approximate-inference framework attenuates the ranking gap but does not remove the orientation-dependent pattern, while the advantage of IM in item-parameter recovery remains pronounced.

E Statistical Significance and Uncertainty

All primary method comparisons are paired at the response-matrix level. Within each benchmark and experimental condition, competing methods are evaluated on exactly the same sampled abilities, items, and responses. We assess paired differences in Kendall's τ using two-sided sign-flip randomization tests with 200,000 draws and report paired nonparametric 95% bootstrap confidence intervals

Table 2: Summary of the primary paired ranking comparisons. Mean effects are differences in Kendall's τ, with positive values favoring IM. "Positive" counts conditions with a positive mean paired effect.

Experiment	Comparison and scope	Mean effect	Positive	Adjusted $p < 0.05$
Response-matrix geometry	IM vs. PM, $M \leq 40$	$+0.027$	$18/18$	$18/18$
Efficient benchmark construction	IM vs. PM	$+0.012$	$35/36$	$28/36$
Efficient benchmark construction	IM vs. Random	$+0.034$	$36/36$	$36/36$
Cold-start sparse evaluation	IM vs. PM	$+0.029$	$30/30$	$24/30$
Cold-start sparse evaluation	IM vs. RAW	$+0.008$	$21/30$	$2/30$

based on 20,000 resamples. When multiple conditions are tested within the same comparison family, we control the family-wise error rate using Holm's procedure.

For the response-matrix geometry experiment, we compare IM and PM separately at each of the 6×9 benchmark–geometry conditions. We additionally test the geometry trend within each benchmark by computing Spearman's correlation between $\log_{10}(J/M)$ and the mean paired IM–PM difference across the nine geometries. Exact two-sided p-values are obtained by enumerating all $9!$ permutations and are Holm-adjusted across the six benchmarks. For efficient benchmark construction, we compare IM with PM and random item selection across the 6×6 benchmark–calibration conditions. For cold-start sparse model evaluation, we compare IM with PM and RAW scoring across the 6×5 conditions with nonzero additional budget.

Table 2 shows highly consistent evidence for the main IM–PM comparisons. In the item-rich geometry regime with $M \leq 40$, IM achieves higher Kendall's τ in all 18 benchmark–geometry conditions, and all 18 differences remain significant after Holm correction. The geometry trend is also significant for all six benchmark-calibrated populations, with Spearman's ρ ranging from 0.950 to 1.000 and Holm-adjusted $p \leq 3.53 \times 10^{-4}$.

For efficient benchmark construction, IM exceeds PM in 35 of 36 conditions, with 28 differences remaining significant after Holm correction, and exceeds random item selection in all 36 conditions. In cold-start sparse evaluation, IM exceeds PM in ranking recovery across all 30 nonzero-budget conditions, with 24 significant differences after correction. Its advantage over RAW scoring is less consistent, with positive effects in 21 of 30 conditions and only 2 remaining significant after correction.

As a sensitivity analysis, we also aggregate paired effects within each benchmark and treat the six benchmark-level effects as independent units. With only six such units, exact sign-flip inference has limited resolution, so we interpret the condition-level tests as evidence of consistency within the evaluated benchmark populations rather than strong evidence of generalization across benchmark populations.

F Extreme Latent Distributions

The MetaBench-calibrated experiments preserve realistic latent distributions but do not systematically vary their shape. We therefore test whether the geometry-dependent PM–IM crossover persists under deliberately non-Gaussian ability and item populations. The 2PL response model remains correctly specified throughout, while the latent populations are varied independently of the empirical MetaBench calibrations.

Data-generating processes. We use a Gaussian reference condition with $\theta_m \sim \mathcal{N}(0, 1)$ and

$$\begin{pmatrix} b_j \\ \alpha_j \end{pmatrix} \sim \mathcal{N}_2 \left(\begin{pmatrix} 0 \\ 0.8 \end{pmatrix}, \begin{pmatrix} 1 & -0.30 \\ -0.30 & 0.36 \end{pmatrix} \right), \qquad \alpha_j = \log a_j, \tag{17}$$

which gives $\mathrm{Corr}(b, \alpha) = -0.5$. We construct seven additional populations that introduce strong skewness, heavy tails, multimodality, or dependence.

- **Skewed ability.** A standardized skew-normal distribution with shape parameter 8.

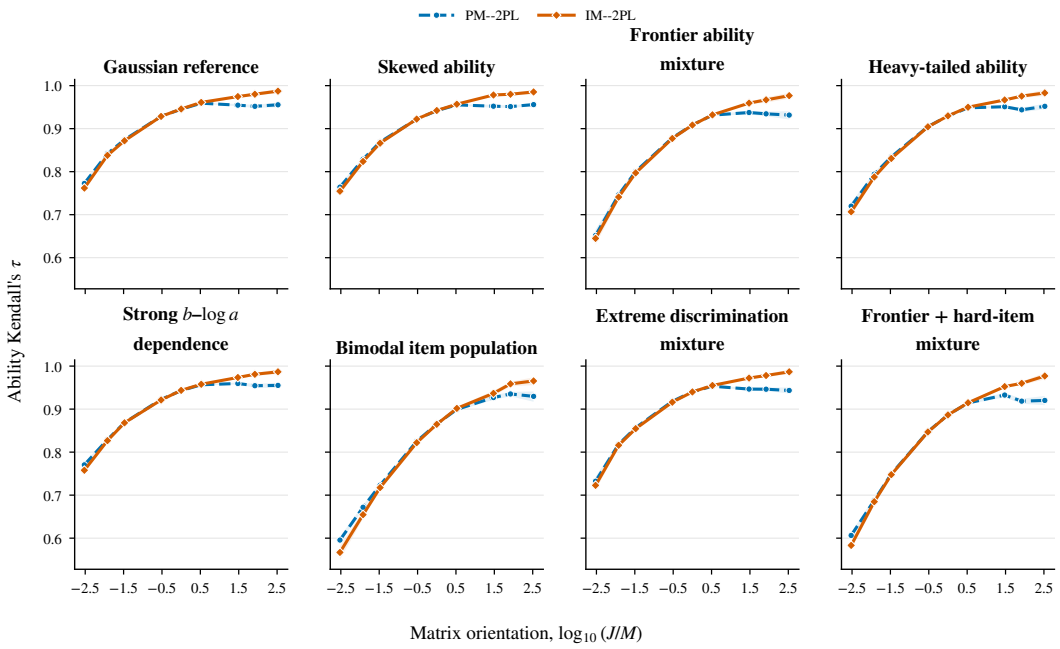


Figure 14: Ability-ranking recovery under extreme latent distributions. Each panel corresponds to one synthetic DGP. Curves report Kendall's τ between true and estimated abilities as the response matrix changes from model rich to item rich. Lines show paired means and bands show 95% bootstrap confidence intervals.

- **Frontier ability mixture.** The standardized mixture $0.9\,\mathcal{N}(-0.3, 0.45^2) + 0.1\,\mathcal{N}(2.7, 0.2^2)$.
- **Heavy-tailed ability.** A variance-standardized Student t_3 distribution.
- **Strong item dependence.** The reference item population with $\mathrm{Corr}(b, \alpha) = -0.9$.
- **Bimodal items.** An equally weighted mixture centered at $(b, \alpha) = (-2, 0.5)$ and $(2, 1.1)$, with within-component standard deviations $(0.35, 0.25)$ and correlation -0.3.
- **Extreme discrimination mixture.** $b \sim \mathcal{N}(0, 1)$ and α drawn with weights $(0.25, 0.70, 0.05)$ from Gaussian components centered at $(\log 0.1, \log 2, \log 12)$ with standard deviations $(0.20, 0.35, 0.15)$.
- **Joint frontier/hard-item mixture.** Frontier-mixture abilities combined with an item population containing 80% regular items centered at $(0, 0.6)$ and 20% hard, highly discriminating items centered at $(3, \log 8)$.

Generated discriminations are restricted to $a \in [0.03, 20]$ for numerical stability, and (b_j, α_j) are always sampled jointly from the specified item population.

Protocol. For each of the eight DGPs, we repeat the same nine response-matrix geometries and approximately fixed response budget $MJ \approx 48{,}000$ used in the main geometry experiment, with 50 paired replications per condition. PM and IM are fitted to the same generated response matrix within each replication. Importantly, both estimators retain their single-Gaussian working distributions even when the corresponding latent population is strongly non-Gaussian, making this a direct stress test of population misspecification. Recovery is evaluated using Kendall's τ for model abilities and Pearson correlation for item difficulty and log discrimination.

Results. The geometry-dependent crossover persists across all eight latent populations. In the item-rich regime with $M \leq 40$, mean ability Kendall's τ increases from 0.943 under PM to 0.973 under IM, corresponding to a mean improvement of 0.029. The improvement is positive for every DGP and ranges from 0.023 for bimodal items to 0.039 for the joint frontier/hard-item mixture.

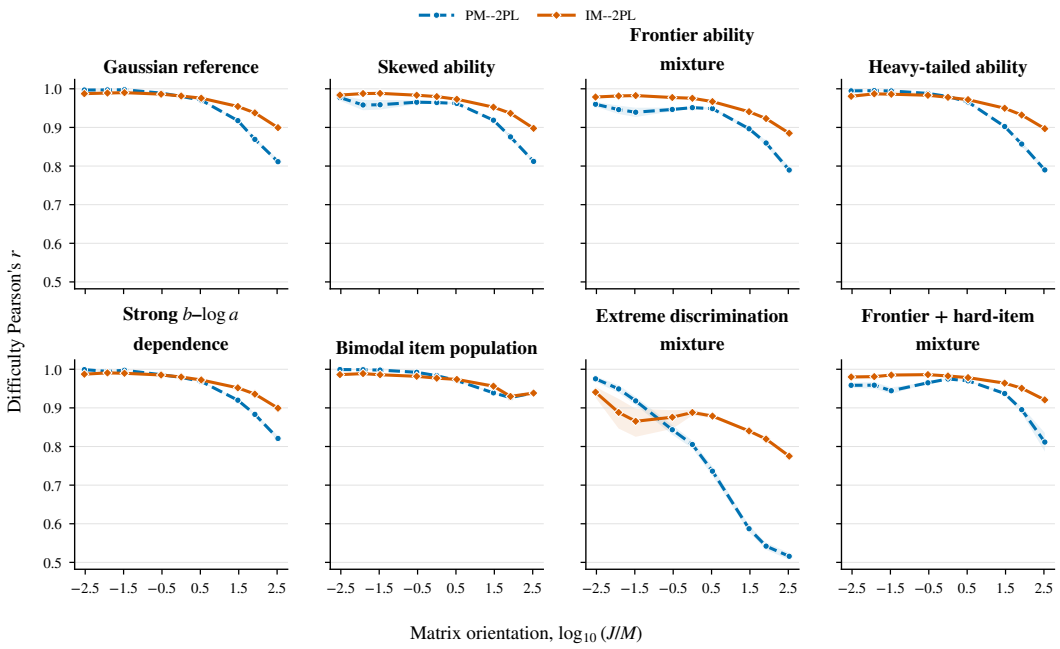


Figure 15: Difficulty recovery under the extreme-distribution stress tests. Curves report Pearson correlation between the true and linked estimated item difficulties. Experimental conditions match Figure 14.

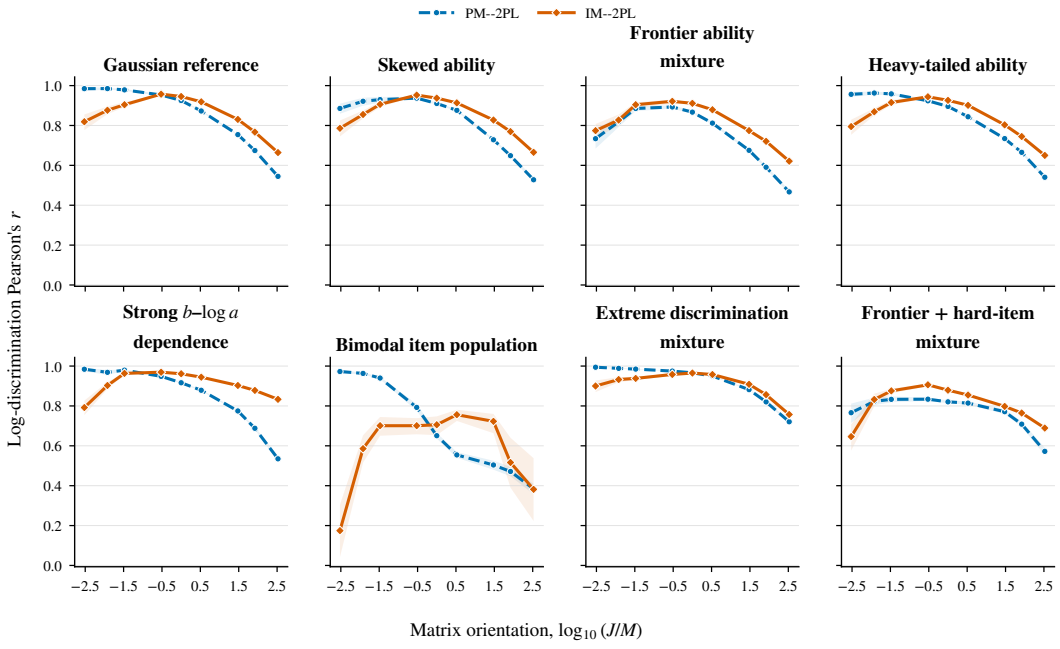


Figure 16: Log-discrimination recovery under the extreme-distribution stress tests. Curves report Pearson correlation between the true and linked estimated $\alpha_j = \log a_j$. Experimental conditions match Figure 14.

Over the same conditions, mean difficulty correlation increases from 0.834 to 0.916, while mean log-discrimination correlation increases from 0.641 to 0.744.

Table 3: Full-data 2PL calibration diagnostics. Score ρ compares observed and fitted model scores, while item ρ compares observed and fitted item accuracies. Both are Spearman correlations.

Benchmark	M	J	Response NLL	Score ρ	Item ρ
ARC	5,221	844	0.336	0.989	0.999
GSM8K	6,068	1,306	0.342	0.997	0.999
HellaSwag	5,221	5,711	0.236	0.997	0.998
MMLU	5,219	12,508	0.374	0.967	0.998
TruthfulQA	5,222	644	0.258	0.990	0.998
WinoGrande	6,550	1,045	0.382	0.977	0.996

The direction of the advantage again changes with response-matrix geometry. Averaged across DGPs, the IM–PM difference in ability Kendall's τ moves from -0.014 at $(M, J) = (4000, 12)$ to $+0.038$ at $(12, 4000)$ and is close to zero near the balanced geometry. Item-parameter recovery exhibits the same qualitative reversal. The persistence of this pattern under skewed, heavy-tailed, multimodal, strongly dependent, and extreme-discrimination populations indicates that the geometry effect is not specific to the empirical MetaBench calibrations or to approximately Gaussian latent populations.

G Benchmark Calibration

Data. We construct the simulation populations from the six complete response matrices released through the official MetaBench preprocessing pipeline (Kipnis et al., 2025). No additional item filtering or subsampling is applied. The resulting matrices contain between 5,219 and 6,550 models and between 644 and 12,508 items, with binary responses and no missing entries.

2PL calibration. Each benchmark is calibrated independently using a unidimensional 2PL model fitted to the complete response matrix with mean-field variational inference in `py-irt` 0.7.1 (Lalor & Rodriguez, 2023). We use the package's `vague` specification, corresponding to a standard-normal ability population together with Gaussian difficulty and log-normal discrimination priors. The fitted model abilities and the joint empirical item distribution $(b_j, \log a_j)$ define the benchmark-specific simulation population used in the first two experiments.

The exported parameters are converted to the canonical form

$$p_{mj} = \sigma[a_j(\theta_m - b_j)], \qquad a_j > 0. \tag{18}$$

We verify numerical equivalence between the native `py-irt` parameterization and the canonical representation, with a maximum absolute probability difference below 6×10^{-16} across the six calibrations.

The fitted response models closely reproduce both model-level scores and item-level accuracies, with score correlations between 0.967 and 0.997 and item correlations between 0.996 and 0.999. Figure 17 shows substantial variation across benchmarks, including visibly non-Gaussian and multimodal ability distributions. These empirical distributions are preserved when abilities and joint item-parameter rows are resampled in the benchmark-calibrated simulations.

Relation to prior calibration. Our calibration follows the same general strategy as Jiang et al. (2026), who fit IRT models to complete OpenLLM response matrices and use the resulting empirical parameter distributions for simulation. We independently recalibrate the MetaBench matrices because their fitted latent parameters are not publicly available. Raw IRT coordinates are identifiable only up to location and scale transformations, and estimates can additionally vary with prior specification, software implementation, and optimization. We therefore validate the calibration through response reconstruction and use the resulting benchmark-specific empirical distributions.

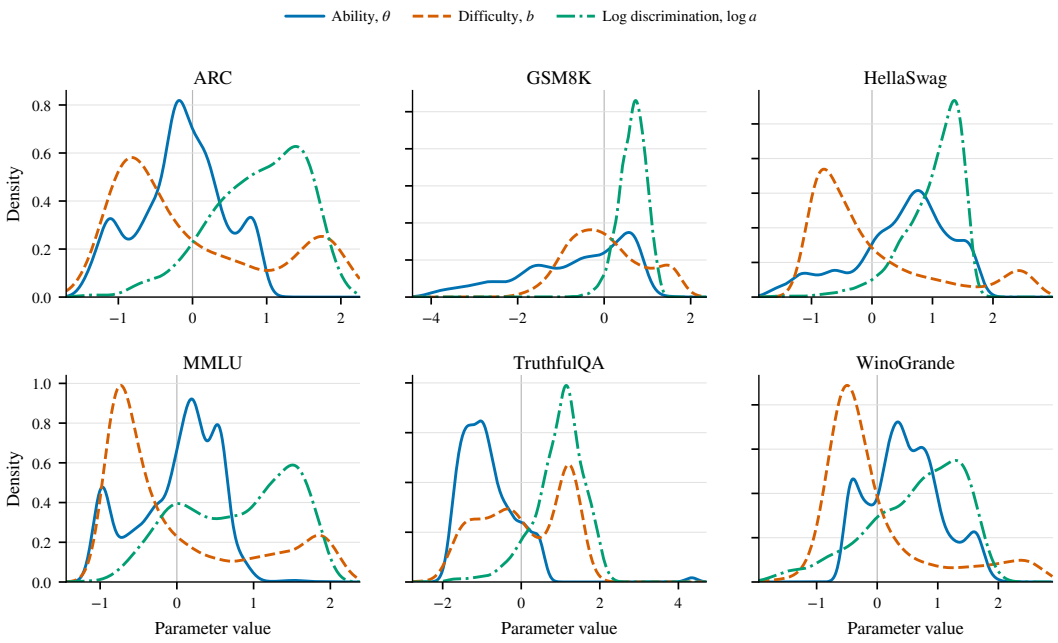


Figure 17: Empirical 2PL parameter distributions defining the benchmark-calibrated simulation populations. Each panel shows the fitted ability, difficulty, and log-discrimination distributions for one benchmark.

H Additional Evaluation Details

H.1 Response-Matrix Geometry Experiment

For each of the six benchmark-calibrated populations described in Appendix G, model abilities are sampled with replacement from the empirical ability distribution and item pairs $(b_j, \log a_j)$ are sampled jointly from the empirical item distribution. Binary responses are then generated from the corresponding 2PL probabilities.

We evaluate nine response-matrix geometries

$$(M, J) \in \{(4000, 12), (2000, 24), (1200, 40), (400, 120), (219, 219), (120, 400), (40, 1200), (24, 2000), (12, 4000)\}, \tag{19}$$

with an approximately fixed response budget of $MJ \approx 48{,}000$. Each benchmark–geometry condition is repeated 50 times, with PM and IM fitted to exactly the same simulated response matrix. Estimates are linked to the generating scale as described in Appendix A. We evaluate recovery of model abilities, item parameters, expected model scores, and model rankings.

H.2 Efficient Benchmark Construction

For each benchmark-calibrated population, we sample

$$M \in \{12, 24, 40, 120, 219, 400\} \tag{20}$$

calibration models and fit PM and IM to the same responses. Items are ranked by their estimated average Fisher information over the calibration models

$$\overline{\mathcal{I}}_j = \frac{1}{M} \sum_{m=1}^{M} a_j^2 p_j(\widehat{\theta}_m) \left[1 - p_j(\widehat{\theta}_m)\right]. \tag{21}$$

For IM, item-specific quantities are obtained from the empirical-Bayes estimates described in Appendix A. We retain the top 10% of items and evaluate the selected subset on 4,000 independently

sampled test models from the same benchmark population. Ranking preservation is measured by Kendall's τ between the selected-subset ranking and the ranking induced by the full item pool. Random subsets of the same size provide a non-IRT baseline. Each benchmark–calibration condition is repeated 30 times.

H.3 Cold-Start Sparse Model Evaluation

The cold-start experiment operates directly on the six observed MetaBench response matrices. In each replication, we sample $M = 24$ real models without replacement and use their full-benchmark score ranking only as the evaluation target. A shared anchor set of

$$A = \lceil 0.05J \rceil \tag{22}$$

randomly selected items is observed for all models.

Responses outside the anchor set follow a balanced sparse design with replication degree $r = 3$, so that every additional item is observed for exactly three models and the additional evaluation load is distributed as evenly as possible across models. The maximum feasible additional budget per model is

$$L_{\max} = \left\lfloor \frac{r(J - A)}{M} \right\rfloor . \tag{23}$$

We evaluate the anchor-only setting together with additional budgets corresponding to 10%, 25%, 50%, 75%, and 100% of $L_{\max}$. For a nonzero budget fraction p,

$$L = \left\lceil \frac{p}{100} L_{\max} \right\rceil . \tag{24}$$

Across the six benchmarks, $L_{\max}$ ranges from 76 additional responses per model for TruthfulQA to 1,485 for MMLU.

PM and IM are fitted to exactly the same observed sparse response matrix and do not use full-benchmark scores during estimation. We additionally compare against naive averaging over all responses observed for each model. Each benchmark–budget condition is repeated 30 times. The primary metric is Kendall's τ between the estimated model ranking and the full-benchmark score ranking, with complementary ranking and score-recovery metrics reported in the supplementary results.