FAIR-Eval: Reliable and Efficient Adaptive Benchmarking for Large Language Models

Anonymous ACL submission

Abstract

Evaluating large language models typically requires thousands of benchmark items, making the process expensive and increasingly impractical. Existing protocols rely on average accuracy over fixed item sets, treating all items as equally informative—yet many items are too easy or too hard to distinguish between models, making evaluation both wasteful and imprecise.Item Response Theory (IRT) with Computerized Adaptive Testing (CAT) offers a principled alternative—separately estimating model ability and item difficulty, then adaptively selecting the most informative items. However, practical deployment is bottlenecked by costly item calibration, and standard greedy selection causes severe over-exposure of frequently picked items while leaving the rest of the bank under-utilized, which risks data contamination over time. We propose **FAIR-Eval** with two components: **Feature-Driven Calibration (FDC)**, which jointly learns item parameters from text features and sparse responses, enabling low-cost bank expansion requiring only minimal model responses; and **Maximum Fisher Information Plus (MFI+)**, which counteracts greedy selection by incorporating exposure control, while simultaneously enforcing difficulty coverage and domain diversity to systematically activate under-utilized items. On a 6,796-item bank (193 calibration models, 14 held-out), FAIR-Eval achieves Pearson $r = 0.9922$ and Spearman $\rho = 0.9648$ against full evaluation using around 3% of items. Crucially, it broadens difficulty coverage from 44.6% to 60.4%, achieving a more comprehensive and robust assessment of model capabilities with minimal test length. Code and data are available at `https://anonymous.4open.science/r/FAIR-Eval-3E6C`.

1 Introduction

Evaluating large language models (LLMs) is essential for model development, deployment decisions, and public comparison. The prevailing protocol is simple: run a candidate model on a fixed benchmark, grade each response, and report an average score over all items (Zhuang et al., 2025). As benchmarks grow to thousands of questions and many models must be evaluated repeatedly, this static protocol becomes increasingly expensive.

The problem is not only cost. Average-score benchmarking treats every item as contributing the same amount of information, even though benchmark items vary widely in difficulty and discriminative value (Rodriguez et al., 2021; Embretson and Reise, 2025). Items that nearly all models solve, or nearly all models miss, provide little evidence for distinguishing nearby systems. As a result, average scores can be confounded by the difficulty distribution of the evaluated subset: a model tested on easier items may appear stronger than a comparable model tested on harder ones. This makes naive subset evaluation unreliable and weakens comparability across benchmarks or domains with different item mixtures (Truong et al., 2025; Polo et al., 2024).

Item Response Theory (IRT), originally developed for educational testing, addresses this by separating model ability from item properties such as difficulty and discrimination, producing ability estimates that are invariant to the particular item subset administered (Baker, 2001; Lord, 2012; Embretson and Reise, 2025). Computerized Adaptive Testing (CAT) further improves efficiency by selecting items according to the current ability estimate instead of administering the same fixed test to every model (Van der Linden et al., 2000; Wainer et al., 2000). Recent work has shown the promise of bringing these ideas to LLM evaluation (Rodriguez et al., 2021; Polo et al., 2024; Truong et al., 2025; Li et al., 2025; Hofmann et al., 2025).

However, deploying IRT-CAT for reusable LLM benchmarks faces two bottlenecks. First, *scalable calibration*: classical calibration requires dense model–item response data, and existing amortized

methods (Lalor et al., 2019; Truong et al., 2025) rely on a disjoint two-stage pipeline—first predicting item parameters from text features and then freezing them for ability estimation—which isolates training signals, leading to error propagation and scale misalignment. Second, *population-aware selection*: standard MFI greedily picks the most informative item for the current model, but in a shared-bank setting where many models are evaluated sequentially, this concentrates usage on a narrow item subset—degrading exposure balance, difficulty coverage, and domain diversity. Classical exposure-control methods target single-session settings with predefined blueprints and do not generalize to the heterogeneous domains of LLM benchmarks (Sympson and Hetter, 1985; Stocking and Lewis, 1998).

We propose **FAIR-Eval**, a benchmark-level adaptive evaluation framework for reliable and efficient LLM benchmarking. FAIR-Eval has two components that directly address the two bottlenecks above. **Feature-Driven Calibration (FDC)** jointly learns item parameters and model abilities from item features and sparse responses, reducing the need for dense calibration responses and keeping item and ability scales aligned. **Maximum Fisher Information Plus (MFI+)** extends standard MFI with exposure control, difficulty coverage, and domain diversity, making shared-bank maintenance part of the adaptive selection objective rather than a post-hoc correction. On a 6,796-item bank spanning five domains and calibrated with 193 anchor models, we evaluate FAIR-Eval on 14 held-out models. FAIR-Eval preserves agreement with full evaluation while using around 3% of items, achieving Pearson r=0.9922 and Spearman ρ=0.9648. It also improves difficulty coverage from 44.6% to 60.4%, showing that efficiency need not come at the cost of a narrower or less reusable item bank.

Our contributions are three-fold:

(*i*) **Problem formulation.** We identify large-scale LLM benchmarking as a shared-bank multi-model evaluation setting and argue that it should be treated as benchmark-level adaptive evaluation rather than test-by-test optimization.

(*ii*) **Method.** We propose FAIR-Eval, which combines FDC for scalable calibration with MFI+ for exposure-aware, coverage-aware adaptive selection.

(*iii*) **Evidence.** We show that FAIR-Eval preserves full-evaluation ranking fidelity with around 3% of items while maintaining a healthier shared item bank, and we analyze why both FDC and MFI+ are necessary to achieve this trade-off. Table 1 summarizes these distinctions.

Table 1: Comparison of evaluation paradigms. ✓= supported, × = not supported by the default formulation, and – = not applicable.

Capability	Static Bench.	IRT-CAT	**FAIR-Eval**
Adaptive item selection	×	✓	✓
Automated calibration	–	×	✓
Cross-dataset comparability	×	✓	✓
Domain diversity guarantee	×	×	✓
Difficulty coverage balance	×	×	✓
Real-time ability estimation	×	✓	✓

2 Related Work

Efficient LLM evaluation without IRT. A broad line of work reduces evaluation cost by selecting informative subsets of benchmark items, constructing compact proxy benchmarks, or predicting full-benchmark performance from partial observations (Vivek et al., 2024; Zhang et al., 2026). While improving efficiency, these subset-based approaches lack granular item-level measurement parameters and uncertainty-aware stopping rules, leaving their aggregated score estimates easily confounded by the difficulty and domain composition of the evaluated subset, especially when different models are tested on different item sets (Truong et al., 2025).

IRT-based evaluation and item calibration. Item Response Theory (IRT) separates examinee ability from item properties such as difficulty and discrimination (Lord, 2012; Embretson and Reise, 2025). Rodriguez et al. (2021) show that evaluation items are not equally informative for NLP model comparison; Polo et al. (2024) use IRT with CAT to preserve full-benchmark rankings with far fewer items; LEGO-IRT (Liao et al., 2025) broadens IRT to continuous outcomes; REEval (Truong et al., 2025) reduces calibration cost through amortized difficulty prediction; and Zhou et al. (2026) revisit LLM benchmarking from an IRT perspective. These works establish IRT as a strong basis for data-efficient evaluation, but calibration remains a bottleneck: amortized approaches typically follow a two-stage pipeline (Lalor et al., 2019; Sharpnack et al., 2024; Truong et al., 2025) that propagates first-stage errors and requires post-hoc scale alignment. FDC addresses this by jointly learning item

parameters and model abilities in a single training pass.

Adaptive testing and shared-bank selection. Computerized Adaptive Testing (CAT) selects items according to the current ability estimate rather than administering a fixed test (Van der Linden et al., 2000; Wainer et al., 2000). ATLAS (Li et al., 2025) and Fluid Benchmarking (Hofmann et al., 2025) apply this paradigm to LLM evaluation, achieving substantial cost reduction.

However, standard CAT optimizes single-examinee precision. In continuous LLM benchmarking, sequential models share one reusable bank, necessitating long-term bank protection against greedy over-exposure and subsequent data contamination. Unlike educational strategies (Sympson and Hetter, 1985; Kingsbury and Zara, 1989; van der Linden and Veldkamp, 2004) designed for independent human sessions, MFI+ embeds exposure control, difficulty coverage, and domain diversity directly into the shared-bank selection objective.

3 Method

3.1 FAIR-Eval Measurement Framework

FAIR-Eval facilitates benchmark-level adaptive evaluation through an offline-online measurement cycle centered on a shared item bank (Figure 1). The **offline stage** calibrates reusable items to provide a stable foundation for cross-model ability estimation, while the **online stage** adaptively selects items for each incoming model to maximize evaluation efficiency while preserving benchmark integrity. Within this framework, **FDC** (Feature-Driven Calibration) handles bank construction, and **MFI+** (Maximum Fisher Information) governs adaptive item selection.

Notation. Let $m \in \{1, \ldots, M\}$ index models, $i \in \{1, \ldots, N\}$ index items. The observed response is $y_{m,i} \in \{0, 1\}$, $\mathcal{S}_t = \{i_1, \ldots, i_t\}$ the set of items selected after t steps, and $\hat{\theta}_t$ the current ability estimate. We adopt the two-parameter logistic (2PL) model:

$$P(y_{m,i} = 1 \mid \theta_m, a_i, b_i) = \sigma(D \cdot a_i(\theta_m - b_i)), \tag{1}$$

where θ_m is model ability, $a_i > 0$ item discrimination, b_i item difficulty, $D = 1.702$[1] is the standard IRT scaling constant, and $\sigma(\cdot)$ the logistic function.

Offline-Online Cycle. In the *offline stage*, FDC (Section 3.2) takes tri-path item features and a sparse response matrix (≈34.91% observed) as input, jointly predicting item parameters $(\hat{a}_i, \hat{b}_i)$ and anchor model abilities $\{\hat{\theta}_m\}$; the calibrated bank is then filtered to 6,796 items. In the *online stage*, MFI+ (Section 3.3) adaptively selects items by balancing Fisher information with exposure control, difficulty coverage, and domain diversity. After each response ($y_{m,i} \in \{0, 1\}$), the ability estimate is updated via Weighted Likelihood Estimation (WLE) (Warm, 1989), yielding $\hat{\theta}_t$ and $\mathrm{SE}(\hat{\theta}_t) = 1/\sqrt{I(\hat{\theta}_t)}$ (Appendix E).

3.2 Feature-Driven Calibration (FDC)

Classical IRT calibration incurs $O(MN)$ data cost that must be repeated whenever the bank is expanded. FDC breaks this dependence by learning a mapping $g_\phi : \mathbf{x}_i \mapsto (\hat{a}_i, \hat{b}_i)$ from item features and sparse responses (34.91% observed), jointly training g_ϕ and model ability embeddings $\{\hat{\theta}_m\}$ under a single reconstruction objective. At inference, g_ϕ predicts parameters for new items from features alone, enabling scalable bank expansion. Corpus construction details are in Appendix A.

Problem formulation. Given an incomplete response matrix $\mathbf{Y} \in \{0, 1\}^{M \times N}$ and item features $\mathbf{x}_i$, FDC jointly learns g_ϕ and $\{\hat{\theta}_m\}_{m=1}^{M}$ by minimizing binary cross-entropy over observed entries. The abilities $\{\hat{\theta}_m\}$ serve as auxiliary training variables that enforce scale alignment; once trained, only g_ϕ is needed. The contrast with two-stage pipelines is in Figure 6 (Appendix D).

Three-Component Feature Representation. Each item is represented by three complementary features: $\mathbf{x}_i = [\mathbf{v}_i^{\text{stem}} \parallel \mathbf{v}_i^{\text{rationale}} \parallel \mathbf{f}_i^{\text{struct}}] \in \mathbb{R}^{8207}$, where $\mathbf{v}_i^{\text{stem}} \in \mathbb{R}^{4096}$ captures question semantics, $\mathbf{v}_i^{\text{rationale}} \in \mathbb{R}^{4096}$ captures solution complexity, and $\mathbf{f}_i^{\text{struct}} \in \mathbb{R}^{15}$ encodes surface-level signals (token length, negation, task type). This design reflects the intuition that difficulty depends not only on what an item asks, but also on the reasoning required to answer it. Detailed feature definitions and rationale generation are in Appendix A.

[1] The constant $D = 1.702$ is conventionally used in logistic IRT models to approximate the normal ogive (Camilli, 1994).

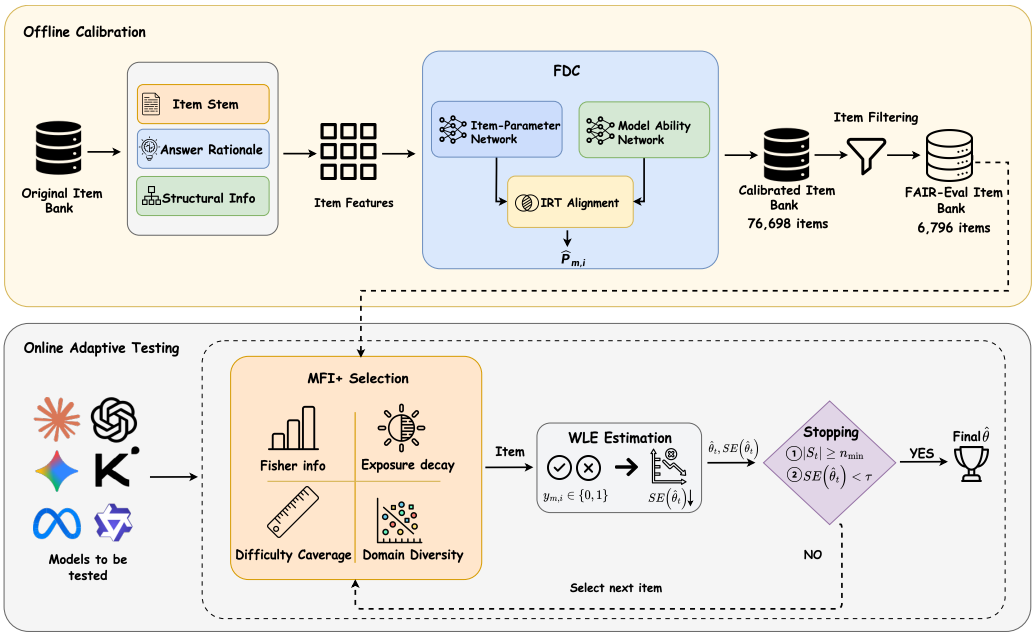


Figure 1: Overall framework of FAIR-Eval. *Top (offline bank construction)*: item features (stem embedding, rationale embedding, structural signals) are fed into FDC, which jointly predicts item parameters and *anchor* model abilities; the calibrated bank is then filtered to retain high-quality items. *Bottom (online bank administration)*: MFI+ selects items by combining Fisher information with exposure decay, difficulty coverage, and domain diversity; WLE updates the ability estimate after each response; the loop terminates when both precision and minimum-item criteria are met.

Multi-View Encoding Architecture. The FDC network g_ϕ (Appendix F.1) encodes $\mathbf{x}_i$ via two parallel pathways—a semantic pathway fusing stem and rationale embeddings, and a structural pathway encoding scalar features—then concatenates and maps the result to $\hat{a}_i$ (Softplus) and $\hat{b}_i$ (linear). During training, an auxiliary branch maps per-model embeddings to $\hat{\theta}_m$; at inference it is discarded. The predicted probability follows:

$$\hat{P}_{m,i} = \sigma\left(\frac{D \cdot \hat{a}_i(\hat{\theta}_m - \hat{b}_i)}{\tau}\right), \tag{2}$$

where $\tau = e^{\log \tau} > 0$ is a learnable temperature that scales the logit variance during early joint training to stabilize gradients. Full architectural details (layer dimensions, initialization, normalization) are in Appendix D.

Training strategy. Let $\mathcal{O}_i = \{m : O_{m,i} = 1\}$ denote models with observed responses for item i. The network g_ϕ and abilities $\{\hat{\theta}_m\}$ are trained end-to-end to minimize:

$$\begin{aligned}\mathcal{L} = -\frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}}\frac{1}{|\mathcal{O}_i|}\sum_{m\in\mathcal{O}_i}\Big[&y_{m,i}\log\hat{P}_{m,i} \\ &+ (1-y_{m,i})\log(1-\hat{P}_{m,i})\Big] \\ &+ \lambda_b\|\hat{\mathbf{b}}\|^2 + \lambda_\theta\|\hat{\boldsymbol{\theta}}\|^2,\end{aligned} \tag{3}$$

where $\mathcal{B}$ is the current batch, and $\hat{\mathbf{b}}_\mathcal{B}$, $\hat{\boldsymbol{\theta}}_\mathcal{B}$ denote the vectors of predicted difficulties and model abilities within $\mathcal{B}$. Regularization weights $\lambda_b, \lambda_\theta$ ensure scale zero-centering, while $\hat{\mathbf{a}}$ is constrained via Softplus without L_2 penalties. Full training details are in Appendix D.

3.3 Benchmark-Aware Item Selection: MFI+

MFI+ is the mechanism that optimizes benchmark-level bank health, not just per-model information. It extends standard Maximum Fisher Information selection with three mechanisms—exposure decay, difficulty coverage, and domain rotation—for the shared-bank setting in which items are reused across many sequential model evaluations.

Motivation: Three benchmark-level design goals. Standard Maximum Fisher Information (MFI) selection optimizes single-test precision by

greedily choosing the item with the highest information at the current ability estimate:

$$i_t^* = \arg\max_{i \notin \mathcal{S}_t} I_i(\hat{\theta}_t), \tag{4}$$

where $I_i(\theta)$ is the Fisher information for item i (defined in Section 3.1). This criterion is well suited to a single examinee, but it does not address three benchmark-level requirements in shared-bank LLM evaluation: (*i*) *bank longevity*—repeated selection of the same items accelerates overexposure; (*ii*) *population-level difficulty coverage*—MFI ignores cumulative coverage across models at different ability levels; and (*iii*) *diagnostic breadth*—greedy selection narrows semantic domain coverage.

MFI+ selection criterion. We define the augmented information score as:

$$\tilde{I}_i(\theta_t, \mathcal{S}_t) = I_i(\theta_t) \cdot \underbrace{\max\Big(0,\ 1 - \tfrac{e_i}{e_{\max}}\Big)^{\lambda}}_{\text{exposure decay}} \cdot \underbrace{\big(1 + \mu \cdot d(b_i, \mathcal{B}_{\mathcal{S}_t})\big)}_{\text{difficulty coverage}}, \tag{5}$$

where e_i is the cumulative exposure count of item i across all previously evaluated models, $e_{\max}$ is the exposure cap, $\lambda \geq 0$ controls exposure penalty strength, $d(b_i, \mathcal{B}_{\mathcal{S}_t}) = \min_{b \in \mathcal{B}_{\mathcal{S}_t}} |b_i - b|$ is the minimum difficulty distance to the already-selected set $\mathcal{B}_{\mathcal{S}_t} = \{b_j : j \in \mathcal{S}_t\}$, and $\mu \geq 0$ controls the coverage bonus strength.

Default values for $e_{\max}$, λ, and μ are in Appendix G; setting $\lambda = \mu = 0$ recovers standard MFI.

Domain rotation via cluster-based hierarchical selection. To ensure diagnostic breadth, MFI+ adds a cluster-based hierarchical selection procedure (Figure 11 in Appendix G): a cluster rotation policy first selects the target domain cluster (sampling uniformly from unvisited clusters, resetting when all have been visited), then samples from the top-c highest-$\tilde{I}_i$ items within that cluster to introduce controlled stochasticity (default $c = 5$). Domain clusters are pre-computed via embedding-based hierarchical clustering (Appendix G); full pseudocode is in Algorithm 1.

Stopping rule. Evaluation terminates when *either* the maximum budget is exhausted ($|\mathcal{S}_t| \geq n_{\max}$) *or* both $|\mathcal{S}_t| \geq n_{\min}$ (default 15) and $\mathrm{SE}(\hat{\theta}_t) < \tau$ (default 0.3) hold simultaneously.

4 Experiments

4.1 Experimental Settings

Item bank. We construct our evaluation bank from three sources (Truong et al., 2025; Du et al., 2026; Phan et al., 2025), yielding **6,796 items** across five domains (Mathematics: 1,335; Physics: 1,088; Chemistry: 509; Biology: 1,086; Other: 2,778) after filtering for psychometric quality ($a \in [0.2, 2.5]$, θ coverage $\in [-0.5, 2.5]$). Full construction details are in Appendix A.

Evaluation protocol. We evaluate 14 models (excluded from the 193-model FDC calibration set) that are stratified by ability range ($\hat{\theta} \in [0.027, 2.018]$), model family, and parameter scale (8B to frontier-tier); the full list is in Appendix I. Because full-static evaluation requires all 6,796 responses per model, we limit the full-reference subset to 14 models to control computational cost. `gpt-4-turbo-2024-04-09` is excluded to preserve independence from FDC rationale generation. As supplementary evidence, 54 models are evaluated adaptively and validated against Chatbot Arena rankings (Section 5.4).

Baselines. We compare against **Full-Static** (all 6,796 items; reference), **Budget-Static** (fixed subsets at 50/100/200 items), **Random-CAT** (adaptive with SE-based stopping, random selection), and **IRT-CAT** (standard MFI without exposure or coverage control). Subscripts (0.08, 0.1, 0.15, 0.2, 0.3) denote the SE threshold τ.

Metrics. We report θ MAE, Pearson r, Spearman ρ, mean rank shift, and average items used. For shared-bank quality, we report exposure and coverage statistics. Information Efficiency Score (IES) (Appendix E) jointly captures estimation quality and cost; lower is better.

4.2 Main Results: Efficiency and Ranking Fidelity

Ranking Fidelity and Efficiency Across Budget Tiers. Table 2 reports results across three budget tiers; all adaptive methods are averaged over 5 independent runs. In Tier 1 (≈200 items), FAIR-Eval$_{0.1}$ achieves Spearman ρ=0.9648, outperforming Budget-Static$_{200}$ (ρ=0.9279) and Random-CAT$_{0.15}$ (ρ=0.9204) at matched budget. In Tier 2 (100–120 items), FAIR-Eval$_{0.14}$ also strikes the best balance between fidelity and efficiency, achieving ρ=0.9428 with only 98 items and leading all

Table 2: Comparison of FAIR-Eval with baseline methods on ability estimation accuracy, ranking consistency, and evaluation efficiency over 14 models. **Bold** indicates the best method within each budget tier (excluding the reference upper bound). All metrics are reported after mean-sigma equating. IES is normalized against Budget-Static$_{200}$ (=1.000); lower is better.

Method	θ MAE ↓	Pearson r ↑	Spearman ρ ↑	Rank Shift ↓	Avg. Items ↓	IES ↓
Reference upper bound						
Full-Static	0.000	1.0000	1.0000	0.00	6796	—
Tier 1: Moderate budget (≈200 items)						
Budget-Static$_{200}$	0.100	0.9786	0.9279	0.97	200	1.000
Random-CAT$_{0.15}$	0.100	0.9780	0.9204	1.03	208	1.039
IRT-CAT$_{0.08}$	0.187	0.9189	0.9209	1.34	217	2.032
FAIR-Eval$_{0.1}$ (Ours)	**0.060**	**0.9922**	**0.9648**	**0.71**	205	**0.623**
Tier 2: Tight budget (≈100 items)						
Budget-Static$_{100}$	0.133	0.9590	0.8409	1.57	100	0.668
Random-CAT$_{0.2}$	0.129	0.9624	0.8422	1.50	118	0.764
IRT-CAT$_{0.1}$	0.174	0.9220	0.9018	1.42	128	1.116
FAIR-Eval$_{0.14}$ (Ours)	**0.068**	**0.9904**	**0.9428**	**1.00**	98	**0.336**
Tier 3: Ultra-low budget (≈50 items)						
Budget-Static$_{50}$	0.190	0.9175	0.7530	1.94	50	0.475
Random-CAT$_{0.3}$	0.213	0.8984	0.8163	1.73	53	0.564
IRT-CAT$_{0.15}$	0.200	0.9066	0.8716	1.48	49	0.494
FAIR-Eval$_{0.2}$ (Ours)	**0.125**	**0.9700**	**0.8840**	**1.20**	46	**0.285**

baselines in this budget tier, while also reducing scale distortion to θ MAE=0.068. In Tier 3 (≈40–50 items), FAIR-Eval$_{0.2}$ maintains the best ranking fidelity (ρ=0.8840) and lowest scale distortion (θ MAE=0.125) at 46 items, confirming that the advantage persists even under extreme budget constraints.

Why Greedy MFI Underperforms in Shared-Bank Evaluation. Across all tiers, standard IRT-CAT yields higher θ MAE than FAIR-Eval despite using comparable or more items; in Tier 1, it also fails to improve ranking fidelity over the non-adaptive baseline (Spearman ρ=0.9209 vs. Budget-Static's 0.9279). The mechanism is straightforward: greedy MFI selects items maximizing Fisher information for the current model's ability estimate, which concentrates usage on a narrow band of high-discrimination items near that estimate. In a shared bank where items vary widely in domain and difficulty, this concentration limits the coverage needed for reliable cross-model ranking, particularly when models cluster at similar ability levels. FAIR-Eval's MFI+ addresses this by integrating exposure decay, difficulty coverage, and domain diversity into the selection objective, reducing peak item exposure from 100% to 85.7% and raising difficulty coverage from 44.6% to 60.4% in Tier 1. Component-level ablation of each MFI+ element is in Section 5.2.

Scale Alignment and External Validity. All θ MAE values are reported after mean-sigma equating applied uniformly to every method; ranking metrics are unaffected by this affine correction. Beyond internal consistency, FAIR-Eval's ability ordering agrees with the full-benchmark accuracy ordering (Spearman ρ=0.99 across 14 models; Figure 9, Appendix D.3), confirming that θ-based rankings preserve the accuracy-based leaderboard while providing finer discrimination among similarly-performing models.

4.3 FDC Calibration Quality

The relevant question for FDC is not whether it perfectly reproduces classical calibration, but whether its predicted item parameters preserve enough psychometric structure for downstream adaptive measurement. We evaluate this through three layers of evidence: (*i*) held-out response reconstruction, (*ii*) agreement with EM-calibrated reference parameters, and (*iii*) downstream ranking fidelity (already established in Section 4.2).

Held-out response reconstruction. We evaluate FDC via 10-fold cross-validation (partitioned by item), reporting validation-fold AUC (Table 3). As an external baseline, we reproduce ADC (Truong et al., 2025), which predicts only $\hat{b}$ (a=1) from stem embeddings. Feature-level ablations are in Section 5.1.

FDC improves AUC by 11.6% over the baseline

Table 3: FDC calibration quality vs. the prior state of the art. AUC is computed on validation folds (mean ± std over 10 folds). Feature-level ablations are in Table 5.

Method	Target	AUC ↑
ADC (Truong et al., 2025)	$\hat{b}$ only (a=1)	0.7532 ± 0.0019
FDC (ours)	$(\hat{a}, \hat{b})$	$\mathbf{0.8405 \pm 0.0019}$

(0.7532 → 0.8405), confirming that joint training with tri-path features substantially outperforms the two-stage pipeline.

Agreement with EM-calibrated references. We directly compare FDC-predicted parameters with classical EM calibration on 2,658 high-density items. FDC achieves Pearson $r = 0.849$ (difficulty) and $r = 0.714$ (discrimination) against EM references under Haebara linking (Appendix Table 9 and Figure 7).

5 Analysis

5.1 FDC Ablation Study

To understand why FDC produces item parameters that support reliable downstream adaptive measurement, we ask two questions: (*i*) which design choices in the calibration mechanism matter most? and (*ii*) which item features carry the recoverable psychometric signal?

Mechanism ablation (Table 4). Using stem embeddings only, we progressively restore FDC's calibration ingredients starting from the two-stage ADC baseline (Truong et al., 2025) (V1). Joint training alone (V2, +0.023 AUC) already outperforms the two-stage baseline, confirming that end-to-end reconstruction addresses error propagation. Adding explicit discrimination prediction (V3, +0.018) further helps because MFI+ relies on Fisher information over both a and b. Replacing linear heads with the multi-view encoder (V4, +0.026) yields the largest gain, showing that the feature-to-parameter mapping requires non-linear capacity. The improvement is cumulative, not attributable to any single tweak.

Representation ablation (Table 5). Under the full FDC architecture, we isolate each input pathway. Semantic features dominate: both stem-only (0.8211) and rationale-only (0.8264) far exceed the baseline (0.7532). Rationale consistently outperforms stem across all configurations, consistent with the intuition that solution complexity is more predictive of difficulty than question semantics. Combining stem and rationale is additive (0.8372), capturing complementary aspects of *what is asked* vs. *what is required to answer*. Structural features add modest but stable gains (+0.003). Overall, item parameters are predictable from complementary semantic and reasoning-structure features, which makes feature-based bank extension feasible.

Table 4: FDC architectural ablation (stem embeddings only). †V1 is the external ADC baseline (Truong et al., 2025). AUC: mean ± std over 10 folds.

Ver.	Training	Params	AUC ↑
V1†	Two-stage, linear	$\hat{b}$ (a=1)	0.7532 ± 0.0019
V2	Joint, 1 linear head	$\hat{b}$ (a=1)	0.7764 ± 0.0016
V3	Joint, 2 linear heads	$(\hat{a}, \hat{b})$	0.7946 ± 0.0015
V4	**FDC encoder (ours)**	$(\hat{a}, \hat{b})$	$\mathbf{0.8211 \pm 0.0014}$

Table 5: FDC feature contribution ablation. All variants predict $(\hat{a}, \hat{b})$. AUC: mean ± std over 10 folds.

Method	Stem	Rat.	Struct	AUC ↑
Stem-only	✓	×	×	0.8211 ± 0.0014
Rationale-only	×	✓	×	0.8264 ± 0.0015
Stem + Struct	✓	×	✓	0.8236 ± 0.0018
Rat. + Struct	×	✓	✓	0.8278 ± 0.0014
Stem + Rat.	✓	✓	×	0.8372 ± 0.0023
FDC (ours)	✓	✓	✓	$\mathbf{0.8405 \pm 0.0019}$

5.2 MFI+ Component Ablation

Each MFI+ component targets a specific benchmark-level goal. To verify that these components are not interchangeable heuristics but address distinct shared-bank requirements, we conduct a reverse ablation: starting from the full MFI+ configuration, we remove one mechanism at a time (Table 6), evaluated under the same adaptive stopping settings across 14 models.

Each component maps to a distinct benchmark-level goal. **Cluster rotation** targets *diagnostic breadth*: removing it drops ρ to 0.8022 and collapses subject coverage to 0.279—the largest single effect. **Difficulty coverage** targets *population-level coverage*: its removal reduces ρ by 0.035 and increases rank shift to 1.00. **Exposure decay** regulates the *precision-sustainability trade-off*: removing it minorly improves θ MAE to 0.055 via unconstrained high-discrimination selection, but pushes peak exposure to 100.0%, causing rapid bank exhaustion. The three components thus correspond one-to-one to the goals in Section 3.3, and

Table 6: MFI+ component ablation (14 models). Starting from full MFI+, each row removes one mechanism to isolate its contribution. Peak Exp. = maximum per-item selection rate across all models; Diff. Cov. = percentage of difficulty bins covered by selected items; Subj. Cov. = effective coverage of subject clusters. IRT-CAT (MFI) is the no-mechanism baseline.

Method	Peak Exp. ↓	Diff. Cov. ↑	θ MAE ↓	Spearman ρ ↑	Rank Shift ↓	Subj. Cov. ↑
MFI+ (full)	**85.7%**	**60.4%**	0.060	**0.9648**	**0.71**	**0.892**
– Exposure decay	100.0%	56.2%	**0.055**	0.9560	**0.71**	**0.892**
– Difficulty cov.	**85.7%**	58.3%	0.073	0.9297	1.00	0.891
– Cluster rotation	92.8%	47.9%	0.320	0.8022	1.86	0.279
IRT-CAT (MFI)	100.0%	44.6%	0.187	0.9209	1.34	0.264

their combination is what enables FAIR-Eval to achieve ranking fidelity, diagnostic coverage, and bank sustainability simultaneously.

5.3 Bank Extensibility: Generalization to Out-of-Bank Items

We use **GPQA-Main** (Rein et al., 2023)—146 graduate-level science questions held out from the FDC training corpus—as a stress test. Item parameters are predicted from features alone (zero calibration responses); ability estimates are compared against a full-bank reference on a 2,683-item in-bank subset selected from the 6,796-item bank to match GPQA's subject coverage, after mean-sigma equating.

Table 7: Generalization to GPQA-main (out-of-bank items, 14 models). FDC-predicted parameters are used directly for ability estimation.

θ MAE ↓	Pearson r ↑	Spearman ρ ↑	Rank Shift ↓
0.2378	0.9027	0.8637	1.7143

FDC achieves Spearman $\rho = 0.8637$ and Pearson $r = 0.9027$ against the reference (Table 7) without any GPQA calibration responses, supporting that FDC can assign psychometrically useful parameters to unseen items from features alone—enabling bank expansion without dense response collection.

5.4 External Sanity Check: Agreement with Human Preference Rankings

As an external sanity check, we examine whether FAIR-Eval's ability rankings are broadly consistent with an independent human-preference ranking. Arena (Arena, 2026) and FAIR-Eval target different constructs — open-ended conversational preference vs. STEM reasoning depth — so we do not expect perfect agreement, but a reasonable correlation would indicate that our ability estimates are not disconnected from broader model-quality judgments. We identify 54 models common to both FAIR-Eval and Arena (a superset of the 14 held-out models used in the main evaluation) and compute Spearman rank correlation between the two orderings.

FAIR-Eval rankings achieve Spearman $\rho = 0.8176$ against Arena rankings across 54 models (Figure 8, Appendix D.2). Notable outliers are interpretable: reasoning-specialized models score higher on FAIR-Eval's STEM-focused bank than on open-ended Arena tasks, while models optimized for conversational fluency rank higher by human judges. This divergence is expected given the construct difference. The strong overall correlation suggests that FAIR-Eval ability estimates, while targeting a narrower construct, do not diverge from external human judgment at the population level.

6 Conclusion

We introduced FAIR-Eval, a framework that reframes large-scale LLM benchmarking as benchmark-level adaptive evaluation over a shared item bank. FDC enables scalable bank construction and expansion by jointly learning item parameters and model abilities from features and sparse responses. MFI+ sustains bank health by encoding exposure control, difficulty coverage, and domain diversity directly into the adaptive selection objective. On a 6,796-item bank evaluated on 14 held-out models, FAIR-Eval recovers full-evaluation ranking fidelity (Spearman $\rho = 0.9648$) with around 3% of items, while broadening difficulty and domain coverage relative to standard adaptive testing. The underlying principles apply to any assessment domain where large item banks must serve many examinees over extended periods.

Limitations

We scope four boundaries of the current work. Calibration assumes a representative model population; as architectures evolve, periodic re-calibration is needed, though FDC's feature-based design makes this a data update rather than a redesign. We estimate a single θ per model, prioritizing cross-model comparability over diagnostic profiling; multidimensional IRT is a natural extension when per-domain scores are required. Our experiments use fill-in-the-blank, knowledge-intensive QA; FDC and MFI+ operate on binary correctness and are format-agnostic in principle, but validation on open-ended generation remains future work. Full-static reference covers 14 held-out models due to inference cost, mitigated by stratified selection across ability, family, and scale, plus Arena validation on 54 models (Section 5.4).

Use of AI Assistants

We used AI assistants (e.g., ChatGPT, Claude) strictly for text polishing, grammatical corrections, and minor coding assistance during the implementation. The core research ideas, scientific methodology, and final manuscript writing were completely driven by the authors.

References

Arena. 2026. Chatbot arena leaderboard (text category). Accessed April 17, 2026.

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

Gregory Camilli. 1994. Teacher's corner: origin of the scaling constant d= 1.7 in item response theory. *Journal of educational and behavioral statistics*, 19(3):293–295.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, and 1 others. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Xeron Du, Yifan Yao, Kaijing Ma, Bingli Wang, Tianyu Zheng, Minghao Liu, Yiming Liang, Xiaolong Jin, Zhenlin Wei, Chujie Zheng, and 1 others. 2026. Supergpqa: Scaling llm evaluation across 285 graduate disciplines. *Advances in Neural Information Processing Systems*, 38.

Susan E Embretson and Steven P Reise. 2025. *Item response theory: Foundations for psychologists and social scientists*. Routledge.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

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

Valentin Hofmann, David Heineman, Ian Magnusson, Kyle Lo, Jesse Dodge, Maarten Sap, Pang Wei Koh, Chun Wang, Hannaneh Hajishirzi, and Noah A Smith. 2025. Fluid language model benchmarking. *arXiv preprint arXiv:2509.11106*.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, Adriane Boyd, and 1 others. 2020. spacy: Industrial-strength natural language processing in python.

Binyuan Hui, Jian Yang, Zeyu Cui, Jiaxi Yang, Dayiheng Liu, Lei Zhang, Tianyu Liu, Jiajun Zhang, Bowen Yu, Keming Lu, and 1 others. 2024. Qwen2. 5-coder technical report. *arXiv preprint arXiv:2409.12186*.

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

John P Lalor, Hao Wu, and Hong Yu. 2019. Learning latent parameters without human response patterns: Item response theory with artificial crowds. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4249–4259.

Peiyu Li, Xiuxiu Tang, Si Chen, Ying Cheng, Ronald Metoyer, Ting Hua, and Nitesh V Chawla. 2025. Adaptive testing for llm evaluation: A psychometric alternative to static benchmarks. *arXiv preprint arXiv:2511.04689*.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, and 1 others. 2022. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*.

Lele Liao, Qile Zhang, Ruofan Wu, and Guanhua Fang. 2025. Toward a unified framework for data-efficient evaluation of large language models. *arXiv preprint arXiv:2510.04051*.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 3214–3252.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, and 1 others. 2025. Deepseek-v3. 2: Pushing the frontier of open large language models. *arXiv preprint arXiv:2512.02556*.

Frederic M Lord. 2012. *Applications of item response theory to practical testing problems*. Routledge.

OpenAI, :, Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K. Arora, Yu Bai, Bowen Baker, Haiming Bao, Boaz Barak, Ally Bennett, Tyler Bertao, Nivedita Brett, Eugene Brevdo, Greg Brockman, Sebastien Bubeck, and 108 others. 2025. gpt-oss-120b gpt-oss-20b model card. *Preprint*, arXiv:2508.10925.

Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, Josephina Hu, Hugh Zhang, Chen Bo Calvin Zhang, Mohamed Shaaban, John Ling, Sean Shi, and 1 others. 2025. Humanity's last exam. *arXiv preprint arXiv:2501.14249*.

Felipe Maia Polo, Lucas Weber, Leshem Choshen, Yuekai Sun, Gongjun Xu, and Mikhail Yurochkin. 2024. tinybenchmarks: evaluating llms with fewer examples. *arXiv preprint arXiv:2402.14992*.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. 2023. Gpqa: A graduate-level google-proof q&a benchmark. *arXiv preprint arXiv:2311.12022*.

Pedro Rodriguez, Joe Barrow, Alexander Miserlis Hoyle, John P Lalor, Robin Jia, and Jordan Lee Boyd-Graber. 2021. Evaluation examples are not equally informative: How should that change nlp leaderboards? In *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)*, pages 4486–4503.

James Sharpnack, Phoebe Mulcaire, Klinton Bicknell, Geoff LaFlair, and Kevin Yancey. 2024. Autoirt: Calibrating item response theory models with automated machine learning. *arXiv preprint arXiv:2409.08823*.

Martha L Stocking and Charles Lewis. 1998. Controlling item exposure conditional on ability in computerized adaptive testing. *Journal of Educational and Behavioral Statistics*, 23(1):57–75.

James B Sympson and Robert D Hetter. 1985. Controlling item-exposure rates in computerized adaptive testing. In *Proceedings of the 27th annual meeting of the Military Testing Association*, pages 973–977.

Kimi Team, Tongtong Bai, Yifan Bai, Yiping Bao, SH Cai, Yuan Cao, Y Charles, HS Che, Cheng Chen, Guanduo Chen, and 1 others. 2026. Kimi k2. 5: Visual agentic intelligence. *arXiv preprint arXiv:2602.02276*.

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

Wim J Van der Linden, Cees AW Glas, and 1 others. 2000. *Computerized adaptive testing: Theory and practice*, volume 13. Springer.

Wim J van der Linden and Bernard P Veldkamp. 2004. Constraining item exposure in computerized adaptive testing with shadow tests. *Journal of Educational and Behavioral Statistics*, 29(3):273–291.

Rajan Vivek, Kawin Ethayarajh, Diyi Yang, and Douwe Kiela. 2024. Anchor points: Benchmarking models with much fewer examples. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1576–1601.

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

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

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Weiqi Zhai, Zhihai Wang, Jinghang Wang, Boyu Yang, Xiaogang Li, Xander Xu, Bohan Wang, Peng Wang, Xingzhe Wu, Anfeng Li, and 1 others. 2026. Hle-verified: A systematic verification and structured revision of humanity's last exam. *arXiv preprint arXiv:2602.13964*.

Guanhua Zhang, Florian E Dorner, and Moritz Hardt. 2026. How benchmark prediction from fewer data misses the mark. *Advances in Neural Information Processing Systems*, 38:135638–135678.

Hongli Zhou, Hui Huang, Ziqing Zhao, Lvyuan Han, Huicheng Wang, Kehai Chen, Muyun Yang, Wei Bao, Jian Dong, Bing Xu, and 1 others. 2026. Lost in benchmarks? rethinking large language model benchmarking with item response theory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 35085–35093.

Yan Zhuang, Qi Liu, Zachary Pardos, Patrick C Kyllonen, Jiyun Zu, Zhenya Huang, Shijin Wang, and Enhong Chen. 2025. Position: Ai evaluation should learn from how we test humans. In *Forty-second International Conference on Machine Learning Position Paper Track*.

A Item Bank Construction Details

Data sources. Our item bank integrates three data sources: (*i*) The amortized evaluation dataset from Truong et al. (2025), covering 22 benchmarks (e.g., MMLU by Hendrycks et al. (2020), GSM8K by Cobbe et al. (2021), and TruthfulQA by Lin et al. (2022)) and 172 model response records. (*ii*) SuperGPQA (Du et al., 2026), a graduate-level science QA dataset with 2,822 items and native rationales. (*iii*) HLE Gold subset (Phan et al., 2025; Zhai et al., 2026), containing 336 expert-verified high-difficulty items.

Cleaning pipeline

1. **Domain filtering**: Remove AirBench (AI safety), Dyck (formal languages), and Synthetic (artificial tasks) as they deviate from knowledge-intensive QA scenarios.

2. **Format conversion**: Convert MCQs to fill-in-the-blank when: (a) stem is semantically complete without options, (b) question intent is unambiguous, (c) answer is a closed-form short response (proper noun, number, formula). Remove items that fail these criteria.

3. **Deduplication**: Use MinHash LSH to detect near-duplicate items (Jaccard similarity > 0.85) and retain only one per cluster.

4. **Quality control**: Manually review 500 randomly sampled items to verify format conversion quality and rationale coherence. Items with ambiguous questions or incorrect rationales are removed.

HELM answer retrieval. Since the dataset constructed in this study requires answer features that are absent from the dataset provided by Truong et al. (2025), it is necessary to trace back to the HELM (Liang et al., 2022) dataset for retrieval and matching. Specifically, we first deduplicate the dataset at the question level by removing redundant responses generated by different models for identical questions, thereby extracting a set of unique question texts. We then locate the corresponding subsets within HELM based on dataset names. Subsequently, a three-stage matching scheme is applied to align the extracted questions with their counterparts in HELM: strict matching, normalized matching, and semantic matching are executed sequentially. Entries for which both the first two stages fail and whose semantic similarity falls below 90%

are deemed unmatched and excluded from further analysis. For successfully matched entries, their ground-truth answers are retrieved accordingly.

Rationale generation. Having obtained the aforementioned QA pairs, we observed that the answers provided in the dataset are typically in the form of short phrases, which do not capture the reasoning depth, knowledge density, or inferential complexity required for a comprehensive rationale. To address this limitation, we employed GPT-4-Turbo to generate detailed rationales for questions lacking such analytical content, thereby constructing the complete QAS pairs (Question-Answer-Solution).

We generate 3 rationales per item and select the one with median token length to avoid overly terse or verbose explanations. GPT-4-Turbo is not included in our evaluation set (Table 10), ensuring independence between feature generation and model evaluation.

The rationale generation prompts used for feature extraction are provided in Appendix B.

Feature extraction. Building upon the resulting QAS pairs, we perform feature extraction. The question text is concatenated with its corresponding dataset description, after which both the concatenated text and the rationale are separately fed into the qwen3-embedding:8b model to obtain semantic embeddings. Concurrently, structured features are extracted using spaCy 3.8.11 (Honnibal et al., 2020). The resulting embeddings and structured features are integrated to form the QASEF (Question-Answer-Solution-Embedding-Features) representation.

Final statistics. The FDC training corpus contains 76,698 items, 193 anchor models, and 5,167,011 observed responses (response density: 34.91%). The evaluation bank (6,796 items selected for psychometric quality) has the following domain distribution: Mathematics (1,335), Physics (1,088), Chemistry (509), Biology (1,086), Other (2,778). Item parameters are FDC-predicted (not EM-calibrated): difficulty $b \in [-1.2700, 4.6373]$ (mean$= 1.0713$, std$= 1.0226$); discrimination $a \in [0.2034, 2.4966]$ (mean$= 0.8095$, std$= 0.4836$).

Figure 2 shows the joint a–b distribution for each domain.

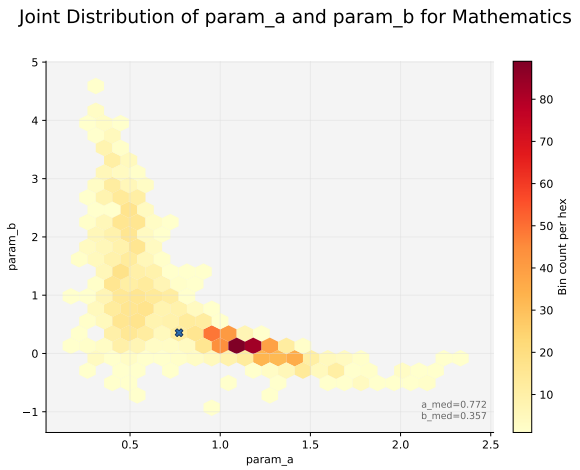
Licenses. The datasets used in this work are publicly available for research use, subject to their respective licenses and access conditions: MMLU (MIT), GSM8K (MIT), SuperGPQA (ODC-BY), HLE (MIT, gated access conditions), GPQA (MIT), and TruthfulQA (Apache 2.0). Our source code is released under the MIT License at https://anonymous.4open.science/r/FAIR-Eval-3E6C.

B Rationale Generation Prompts

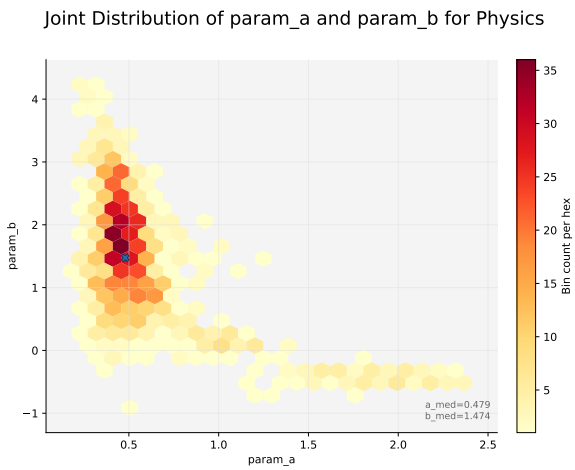
The following prompts were used to generate rationales via GPT-4-Turbo for all items in the training corpus.

System Prompt

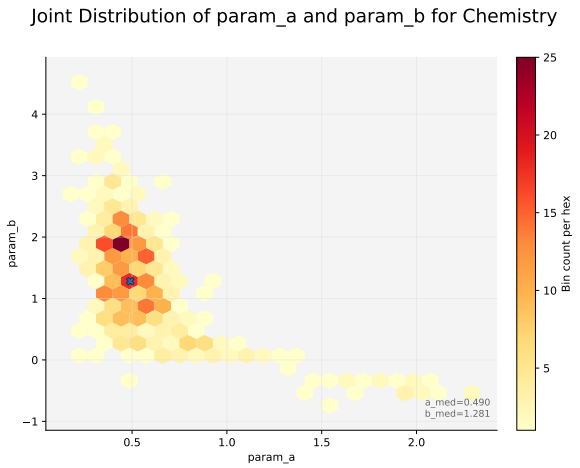
```
You are a careful assistant who writes
explanations for tasks whose acceptable final
answer(s) are already provided.
Your goal is not to discover a new answer.
Your goal is to choose one acceptable provided
answer and produce a clean explanation showing
why that answer is correct based on the task.
The provided acceptable answers may contain
multiple alternatives separated by '|'.
Treat '|' as "or", not "and".
You only need to justify one acceptable answer,
not all of them.
The explanation should vary naturally with
task difficulty: keep it short for simple
tasks, and expand only when multiple reasoning
steps are genuinely needed.
Do not produce raw scratchwork,
stream-of-consciousness, or conversational
filler.
You must output valid XML with exactly these
three tags and no other text:
<thinking>...</thinking>
<solution>...</solution>
<answer>...</answer>
Requirements:
1. In <thinking>, give a brief high-level plan
in 1 to 3 short sentences.
2. In <solution>, write a polished human-style
rationale with this general structure when
applicable:
evidence or key clue -> reasoning steps ->
conclusion.
3. Keep the structure consistent, but let the
length vary naturally with task complexity.
4. If the answer follows directly from one key
clue, keep <solution> short.
5. If the task requires multiple comparisons,
eliminations, or inference steps, include
those steps in order.
6. Do not include self-talk, meta commentary,
or filler such as "let me think", "I will solve
this", or "based on the question".
7. Do not add unnecessary background knowledge
or extra facts not needed to justify the chosen
answer.
8. Avoid hedging unless the task itself is
genuinely uncertain.
9. When the task provides explicit labels,
categories, or answer options, use them
exactly rather than paraphrasing.
```



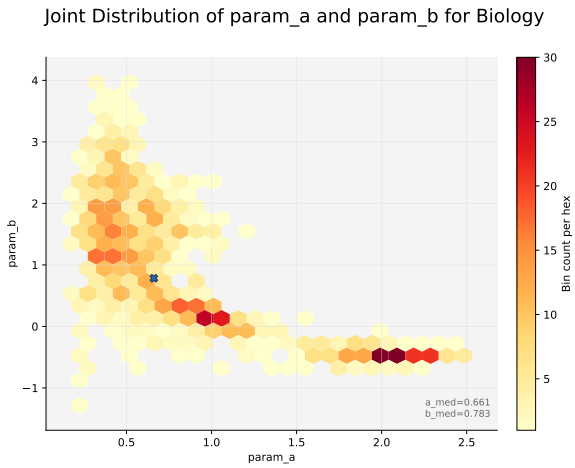
(a) Mathematics (1,335 items)



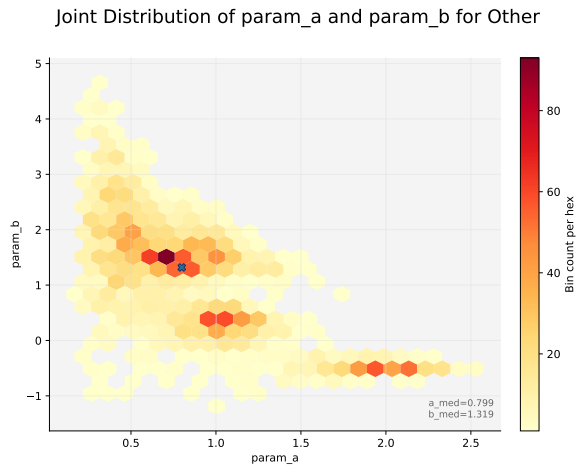
(b) Physics (1,088 items)



(c) Chemistry (509 items)



(d) Biology (1,086 items)



(e) Other (2,778 items)

Figure 2: Joint distribution of FDC-predicted discrimination a and difficulty b parameters for each domain in the 6,796-item evaluation bank.

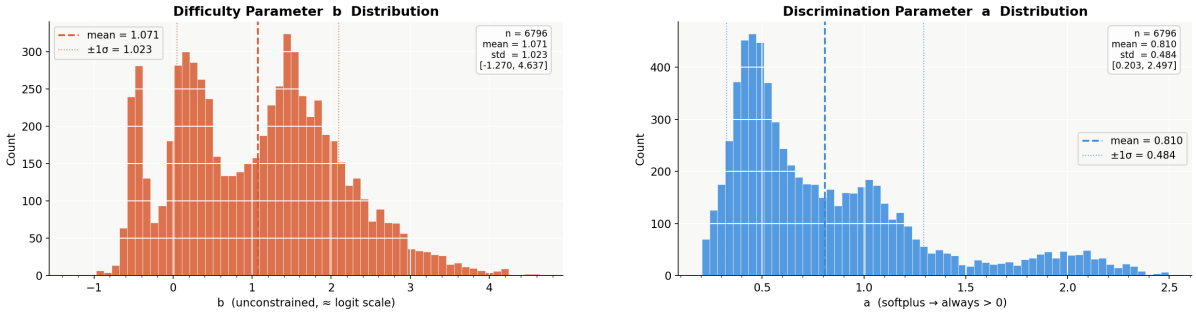


Figure 3: FDC-predicted parameter distributions for the 6,796-item evaluation bank. *Left*: difficulty b ($b \in [-1.27, 4.64]$, mean$= 1.071$, std$= 1.023$). *Right*: discrimination a ($a \in [0.20, 2.50]$, mean$= 0.810$, std$= 0.484$).

```
10. Choose exactly one answer from the
provided acceptable answers and justify only
that one.
11. Do not propose or invent any answer outside
the provided acceptable answers.
12. In <answer>, copy exactly one chosen
acceptable answer only.
13. If formulas, equations, symbolic
expressions, matrices, or other mathematical
notation are involved, write them clearly in
LaTeX in <solution>. Use LaTeX in <answer>
as well when the final answer itself is
mathematical.
14. Do not output markdown fences or any text
outside the XML tags.
15. If several provided answers are
equivalent, choose the one that is most
directly supported by the task.
```

User Prompt

```
The following task already has one or more
acceptable final answers.
Task:
{task}
Acceptable answers:
{correct_answer}
Important instructions:
- The acceptable answers are separated by '|'.
- These are alternative correct answers, not
answers that all need to be justified.
- Choose one acceptable answer that best fits
the task, and explain why it is correct.
- Do not search for a different answer outside
the provided set.
- If the task includes labeled options,
categories, or answer text, use them exactly.
If the task involves formulas, equations,
symbolic expressions, matrices, or other
mathematical notation, express them in LaTeX.
Write the response so that:
- <thinking> gives a brief plan
- <solution> explains the key evidence,
reasoning, and conclusion for one chosen
acceptable answer
- <answer> contains exactly one chosen
acceptable answer from the provided set
Return exactly:
<thinking>...</thinking>
<solution>...</solution>
<answer>...</answer>
```

C Anchor Model Statistics

Anchor model response collection. To align newly added items (SuperGPQA and HLE) with the existing ability scale, we collect additional responses from a separate set of 22 *bridge calibration models*. These bridge models are used only for calibration and scale-linking purposes, and are disjoint from all held-out models used for full-reference evaluation in Table 2 and Table 10. No responses from the 14 held-out full-reference models are used in FDC training, item filtering, anchor alignment, hyperparameter tuning, or model selection.

Together with the original response records, the final FDC calibration pool contains 193 calibration models. These calibration models span multiple families and capability levels:

- **Ability distribution**: Mean $\theta = 0.52$, std $= 0.89$, range $[-3.4, 2.1]$
- **Model families**: GPT (42 models), LLaMA (38), Mistral (24), Qwen (21), Claude (15), Gemini (12), others (38)
- **Parameter sizes**: 1B–7B (68 models), 8B–30B (54), 30B–70B (41), 70B+ (27)

This diversity helps FDC learn item difficulty patterns that are less tied to a single model family or capability band.

D FDC Network Training Details

Table 8 lists all hyperparameters for FDC training. We use 10-fold cross-validation partitioned by item: in each fold, 90% of items form the training split and the remaining 10% form the validation split. Early stopping monitors validation AUC; training halts if validation AUC does not improve for 15 consecutive epochs, subject to a minimum of 20 epochs.

Table 8: FDC network hyperparameters.

Hyperparameter	Value
Embedding dimension (stem, rationale)	4096
Projection dimension	128
Semantic encoder hidden dim	64
Structural encoder hidden dims	[64, 64]
Fusion encoder hidden dims	[64, 32]
Ability embedding dimension	64
Ability MLP hidden dims	[64, 32, 1]
Dropout rate	0.2
Optimizer	AdamW
Learning rate (semantic projection layer)	1×10^{-4}
Learning rate (all other parameters)	1×10^{-3}
Weight decay	1×10^{-3}
Batch size	256
Gradient clip norm	2.0
λ_b (difficulty regularization)	0.1
λ_θ (ability regularization)	0.1
Temperature τ (initial)	1.0
Early stopping patience	15 epochs
Minimum training epochs	20
Maximum training epochs	200

Computational resources. The FDC network contains 0.576M trainable parameters. Training (10-fold cross-validation) requires approximately 6 GPU-hours on a single NVIDIA GeForce RTX 3070.

The training set comprises items drawn from three data sources as described in Appendix A; the composition is summarized in Figure 4.

D.1 FDC Parameter Validation Against EM Calibration

Table 9: FDC parameter prediction accuracy vs. EM ground-truth on 2,658 high-density items, under four scale-linking methods. Lower MAE is better.

Scale-Linking Method	a MAE ↓	b MAE ↓
Raw EM (no linking)	**0.3214**	0.5768
Mean-Sigma	1.3285	**0.2927**
Stocking–Lord	0.3290	0.4374
Haebara	0.4593	0.3556

Figure 5 presents the training and validation loss curves, demonstrating stable convergence throughout the training process.

D.2 Arena Agreement Figure

Figure 8 visualizes the rank correspondence between FAIR-Eval ability estimates and Chatbot Arena human-preference rankings, as discussed in Section 5.4.

D.3 Ability–Accuracy Ranking Agreement

Figure 9 shows that FAIR-Eval's latent ability ordering aligns closely with raw accuracy rankings, confirming that θ estimates are faithful to observable performance while offering finer granularity among similarly-performing models.

E Supplementary Formulations

Weighted Likelihood Estimation (WLE). After each response, the ability estimate is updated by maximizing the weighted likelihood (Warm, 1989):

$$\hat{\theta}_t = \arg\max_{\theta} \ \ln L(\theta \mid \mathcal{S}_t) + \tfrac{1}{2} \ln I(\theta), \qquad (6)$$

where $\ln L(\theta \mid \mathcal{S}_t) = \sum_{i \in \mathcal{S}_t} \big[y_{m,i} \ln P_i(\theta) + (1 - y_{m,i}) \ln(1 - P_i(\theta)) \big]$ is the log-likelihood and $I(\theta) = \sum_{i \in \mathcal{S}_t} D^2 a_i^2 P_i(\theta)(1 - P_i(\theta))$ is the observed Fisher information. The bias-correction term $\frac{1}{2} \ln I(\theta)$ reduces the finite-sample bias of maximum likelihood estimation, which is particularly important in adaptive testing where few items are administered. The standard error is $\mathrm{SE}(\hat{\theta}_t) = 1/\sqrt{I(\hat{\theta}_t)}$.

Information Efficiency Score (IES). IES jointly captures estimation quality and evaluation cost relative to the Budget-Static$_{200}$ reference:

$$\mathrm{IES} = \left(\frac{\mathrm{MAE}_{\mathrm{method}}}{\mathrm{MAE}_{\mathrm{Static200}}} \right) \left(\frac{\mathrm{Items}_{\mathrm{method}}}{200} \right),$$

where $\mathrm{MAE}_{\mathrm{Static200}}$ is the MAE of random 200-item sampling. Lower IES is better; IES $= 1$ corresponds to Budget-Static$_{200}$.

F Architecture Illustrations

F.1 FDC Network Architecture

Figure 10 illustrates the FDC network architecture. The model maps heterogeneous item features into calibrated IRT parameters via two parallel encoding pathways: a semantic pathway that fuses stem and rationale embeddings, and a structural pathway that encodes scalar features and task-type information. The two representations are concatenated and passed through a fusion encoder, from which separate heads predict discrimination $\hat{a}_i$ and difficulty

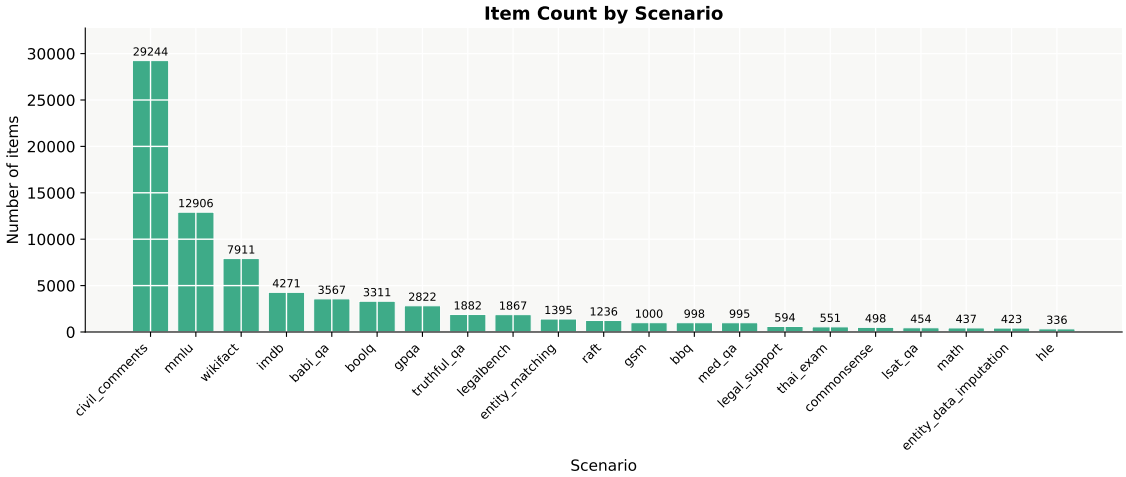


Figure 4: Item Count by Scenario in the Training Set.

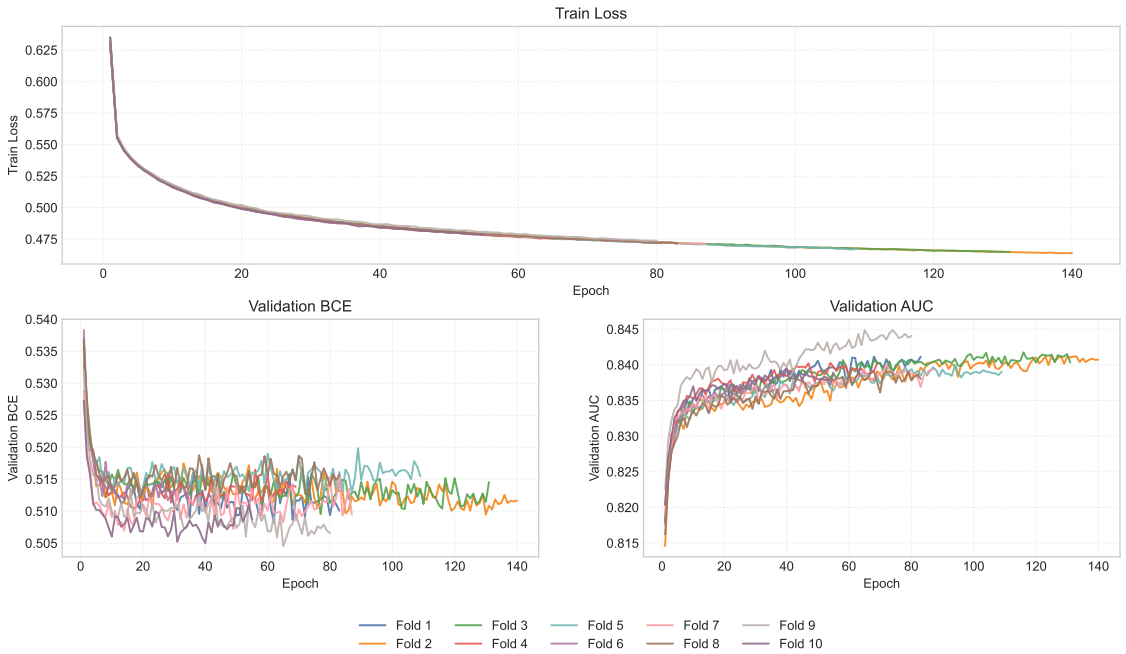


Figure 5: Training and Validation Curves Across 10 Folds.

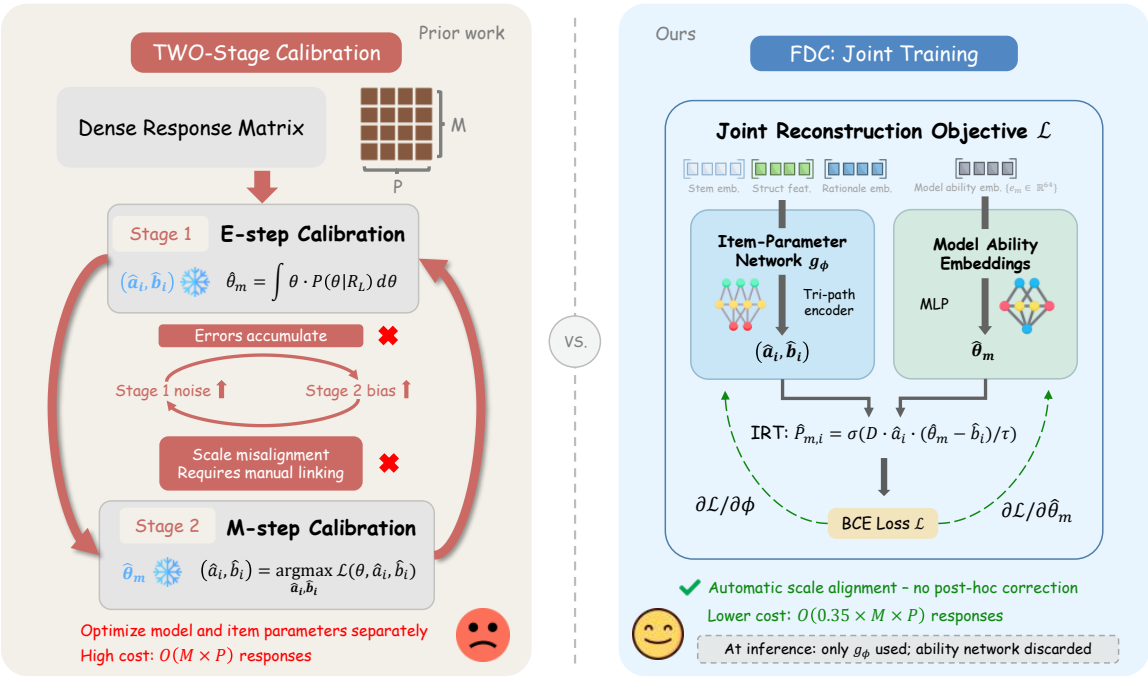


Figure 6: Two-stage calibration (left) vs. FDC joint training (right). Two-stage pipelines (Lalor et al., 2019; Truong et al., 2025) alternate between fixing items to update abilities and fixing abilities to update item parameters, then require post-hoc scale linking. FDC instead optimizes both under a single reconstruction objective, yielding automatically aligned item and ability scales.

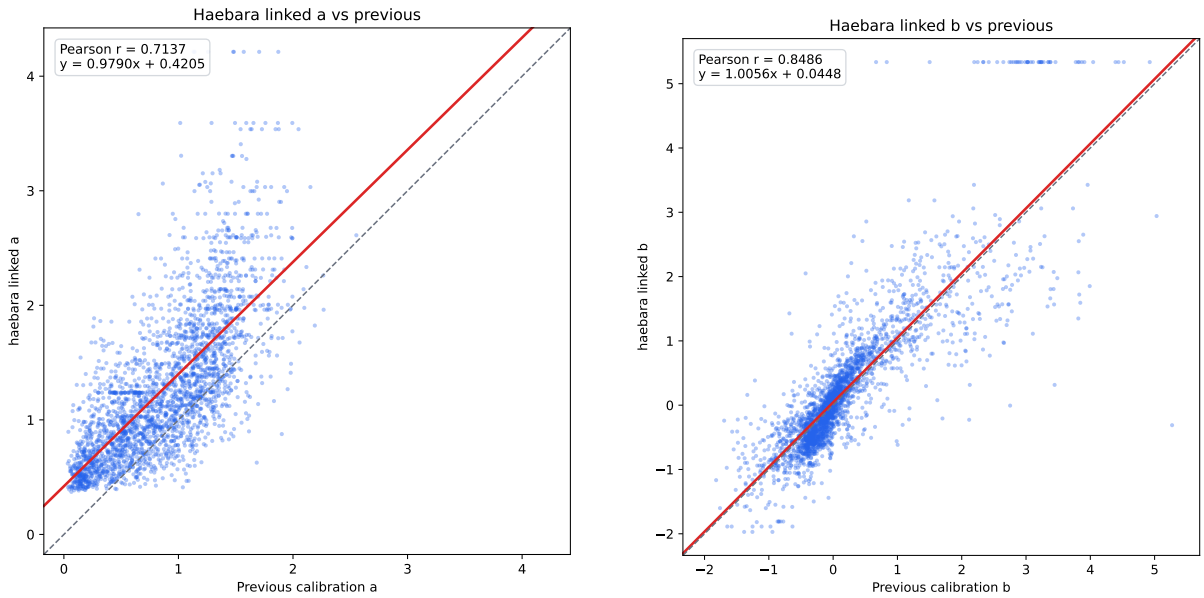


Figure 7: FDC parameter predictions (Haebara-linked) vs. EM ground-truth on 2,658 items. *Left*: difficulty b (Pearson $r = 0.849$, slope $= 1.006$). *Right*: discrimination a (Pearson $r = 0.714$, slope $= 0.979$). Red lines are OLS fits; diagonal would indicate perfect agreement.

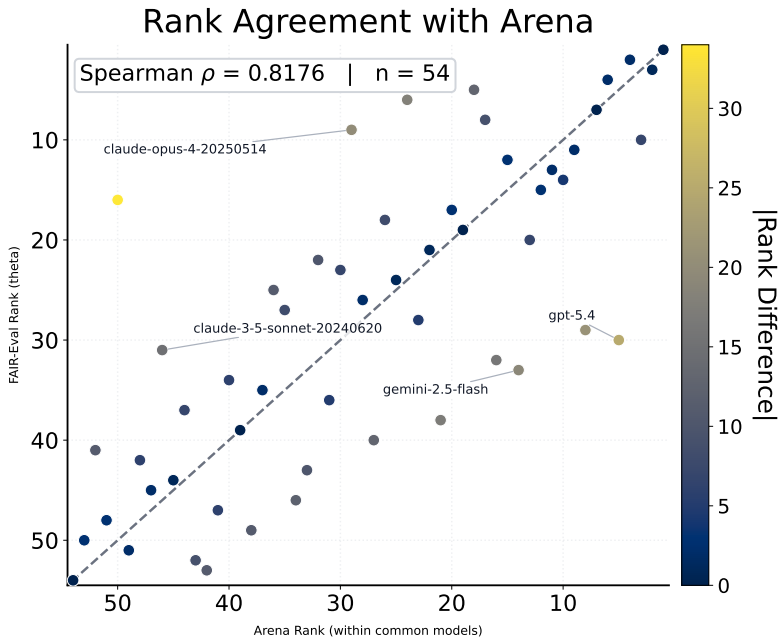


Figure 8: Rank agreement between FAIR-Eval ability estimates ($\hat{\theta}$) and Chatbot Arena rankings across 54 common models (Spearman $\rho = 0.8176$). Each point represents one model; color indicates absolute rank difference.

$\hat{b}_i$. During training, an auxiliary ability network (dashed branch) maps per-model embeddings to $\hat{\theta}_m$; this branch is discarded at inference.

G MFI+ Selection Algorithm

Figure 11 provides a visual overview of the two-stage MFI+ selection process: a cluster rotation policy first selects the target domain to ensure cross-domain coverage, then candidates within that cluster are scored using the augmented information criterion (Eq. 5) that combines Fisher information with exposure decay and difficulty coverage bonuses. Algorithm 1 gives the complete pseudocode.

Key design choices. (*i*) The exposure-decay multiplier $\max(0,\, 1 - e_i/e_{\max})^{\lambda}$ (with $e_{\max} = 100$, $\lambda = 0.1$) linearly down-weights frequently selected items and zeroes out any item that has been selected $e_{\max}$ times. (*ii*) The difficulty-coverage bonus $1 + \mu \cdot d(b_i, \mathcal{B}_{\mathcal{S}_t})$ (with $\mu = 0.1$) provides a proportional bonus to items whose difficulty is far from all already-selected difficulties, actively filling uncovered regions of the difficulty spectrum. (*iii*) The two-condition stopping rule requires minimum items ($|\mathcal{S}_t| \geq 15$) and sufficient precision ($\mathrm{SE}(\hat{\theta}_t) < 0.3$) before terminating, preventing premature stopping when estimation appears to have stabilized based on too few administered items.

H Supplementary Visualizations of Adaptive Selection Dynamics

This section presents supplementary visualizations of the adaptive selection dynamics of the two methods introduced in the main text, namely IRT-CAT and FAIR-Eval. For reference, IRT-CAT uses MFI as its item-selection rule, whereas FAIR-Eval uses MFI+. The figures below highlight how these different selection mechanisms shape the sequential evaluation process in practice, focusing on four aspects: item exposure concentration, estimation trajectories, subject composition of administered items, and cumulative difficulty-range coverage.

I Full Experimental Results

Table 10 reports per-model CAT results for all 54 evaluated models (excluding `gpt-4-turbo-2024-04-09` to preserve feature independence). Models marked with * constitute the 14-model full-reference subset used in Table 2; they were selected via stratified sampling across ability, family, and scale, and are excluded from the 193-model FDC calibration set.

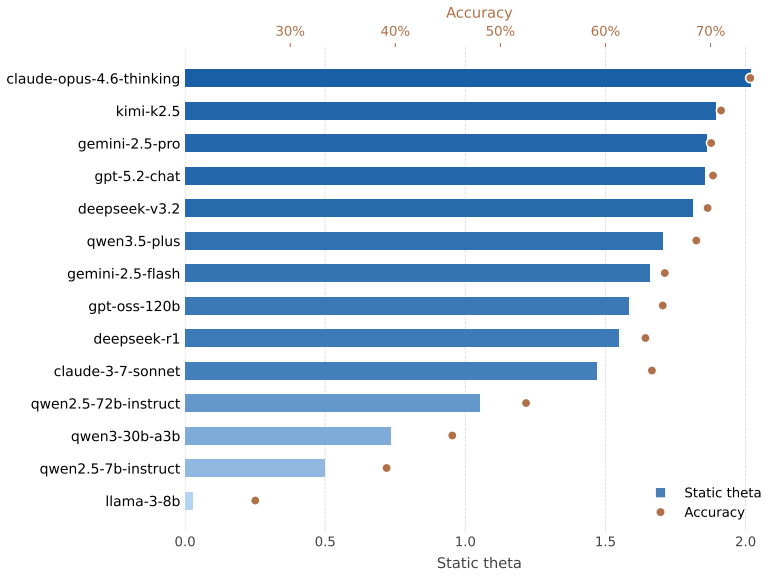


Figure 9: Rank agreement between FAIR-Eval ability estimates ($\hat{\theta}$) and full-benchmark raw accuracy across 14 held-out models (Spearman $\rho = 0.99$). The strong monotonic relationship confirms that θ-based rankings preserve accuracy-based orderings.

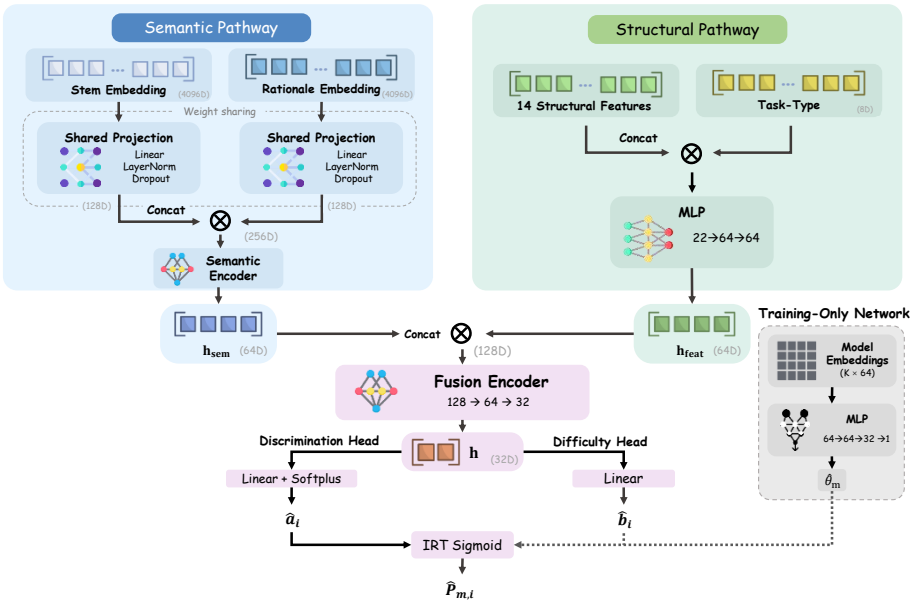


Figure 10: FDC network architecture. The main item network encodes stem and rationale embeddings through a shared semantic pathway, and 14 scalar features plus a task-type embedding through a structural pathway. Their outputs are fused into $\mathbf{h} \in \mathbb{R}^{32}$, from which separate heads predict discrimination $\hat{a}_i$ and difficulty $\hat{b}_i$, and the 2PL sigmoid computes the response probability $\hat{P}_{m,i}$. The dashed branch on the right is a training-only auxiliary ability network that maps model embeddings to $\hat{\theta}_m$; at inference, this branch is discarded.

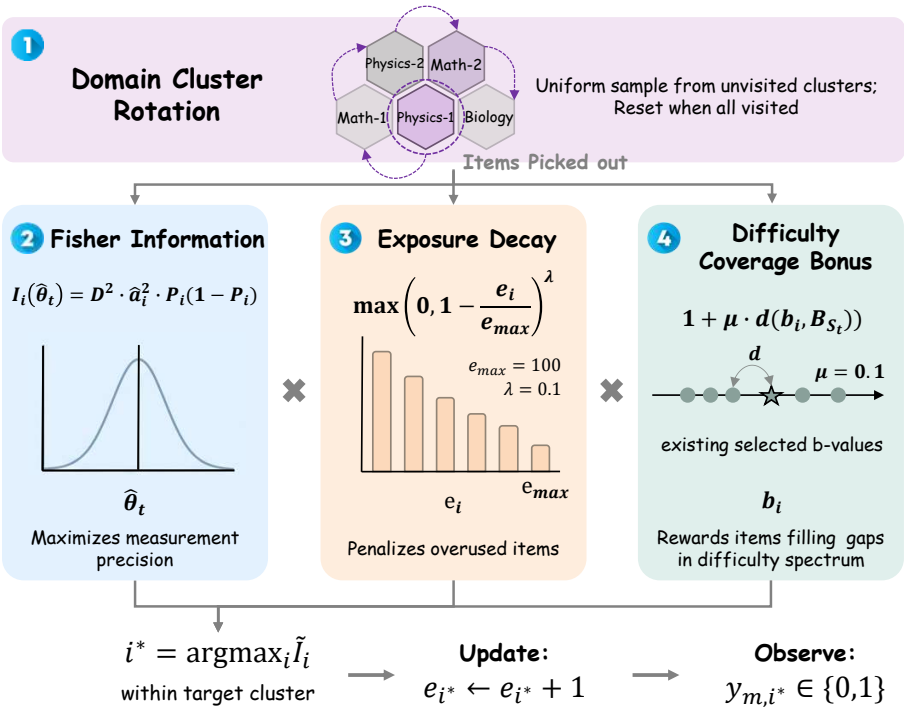


Figure 11: MFI+ item selection at step t. **Stage 1** (top): a domain cluster rotation policy samples uniformly from unvisited clusters and resets once all have been visited, ensuring cross-domain coverage. **Stage 2** (bottom): within the selected cluster, each candidate item is scored as the product of Fisher information, exposure decay, and difficulty coverage bonus (Eq. 5); one item i^* is sampled uniformly at random from the top-c highest-scoring candidates, its exposure count is incremented, and the response y_{m,i^*} is observed.

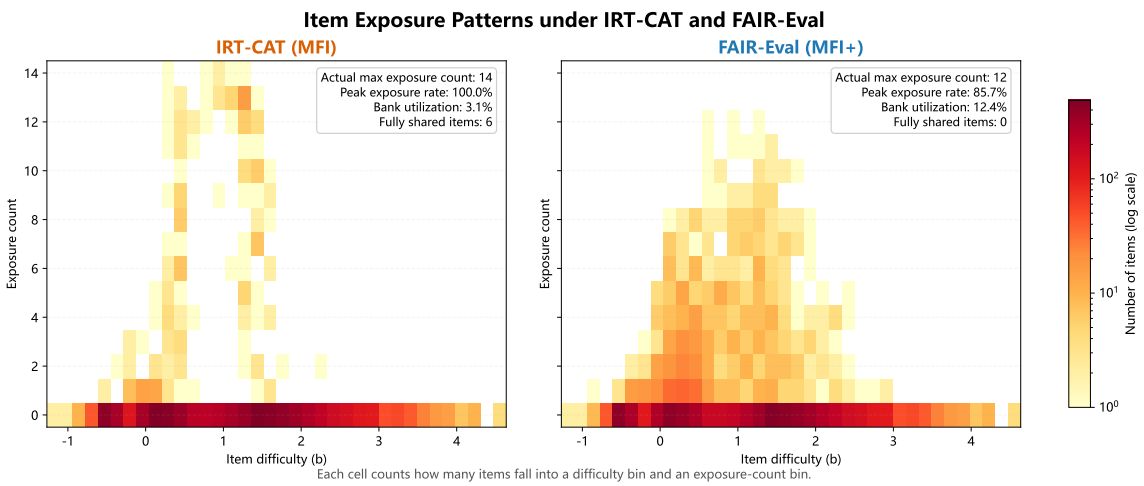


Figure 12: Item exposure patterns under IRT-CAT (MFI) and FAIR-Eval (MFI+). Each heatmap shows the number of items falling into each combination of item difficulty (x-axis) and exposure count (y-axis), where exposure count denotes how many evaluated models selected the item. Color indicates the number of items in each cell on a log scale, allowing direct comparison of how strongly item usage concentrates across the item bank under the two methods.

Algorithm 1 MFI+ Selection with Exposure Decay and Difficulty Coverage

Input: Item bank $\mathcal{I}$, current ability estimate $\hat{\theta}_t$, selected set $\mathcal{S}_t$, exposure counts $\{e_i\}$, difficulty values $\{b_i\}$, cluster assignments $\{\text{cluster}(i)\}$, used-cluster set $\mathcal{U}$
Output: Selected item i^*
Parameters: $\lambda = 0.1$ (exposure decay), $\mu = 0.1$ (coverage bonus), $e_{\max} = 100$, $D = 1.702$

// Stage 1: Select target cluster via rotation
if $\exists$ cluster k with $k \notin \mathcal{U}$ **then**
  $k^* \leftarrow$ random unused cluster
else
  $\mathcal{U} \leftarrow \emptyset$ // Reset rotation
  $k^* \leftarrow$ random cluster
end if

// Stage 2: Select best item within cluster k^*
for each item $i \in \mathcal{I}$ with $\text{cluster}(i) = k^*$ **do**
  // Compute Fisher information
  $P_i \leftarrow \sigma(D \cdot a_i(\hat{\theta}_t - b_i))$
  $I_i \leftarrow D^2 \cdot a_i^2 \cdot P_i \cdot (1 - P_i)$

  // Apply exposure decay
  $w_{\text{exp}} \leftarrow \max\big(0,\, 1 - e_i/e_{\max}\big)^{\lambda}$

  // Compute difficulty coverage bonus
  if $\mathcal{S}_t = \emptyset$ **then**
    $d_i \leftarrow 0$ // Initialization: no prior selected difficulties
  else
    $d_i \leftarrow \min_{j \in \mathcal{S}_t} |b_i - b_j|$ // Distance to nearest already-selected difficulty
  end if
  $w_{\text{cov}} \leftarrow 1 + \mu \cdot d_i$ // Bonus proportional to difficulty gap

  // Compute augmented information
  $I_i^+ \leftarrow I_i \cdot w_{\text{exp}} \cdot w_{\text{cov}}$
end for

$\mathcal{T} \leftarrow$ top-5 items in $\{i \in \mathcal{I} : \text{cluster}(i) = k^*\}$ by I_i^+
$i^* \leftarrow$ sample uniformly from $\mathcal{T}$ // Controlled stochasticity
$\mathcal{U} \leftarrow \mathcal{U} \cup \{k^*\}$ // Mark cluster as used
$e_{i^*} \leftarrow e_{i^*} + 1$ // Update exposure count

// Two-condition stopping rule
if $|\mathcal{S}_t| \geq 15$ **and** $\text{SE}(\hat{\theta}_t) < 0.3$ **then**
  return STOP
end if

return i^*

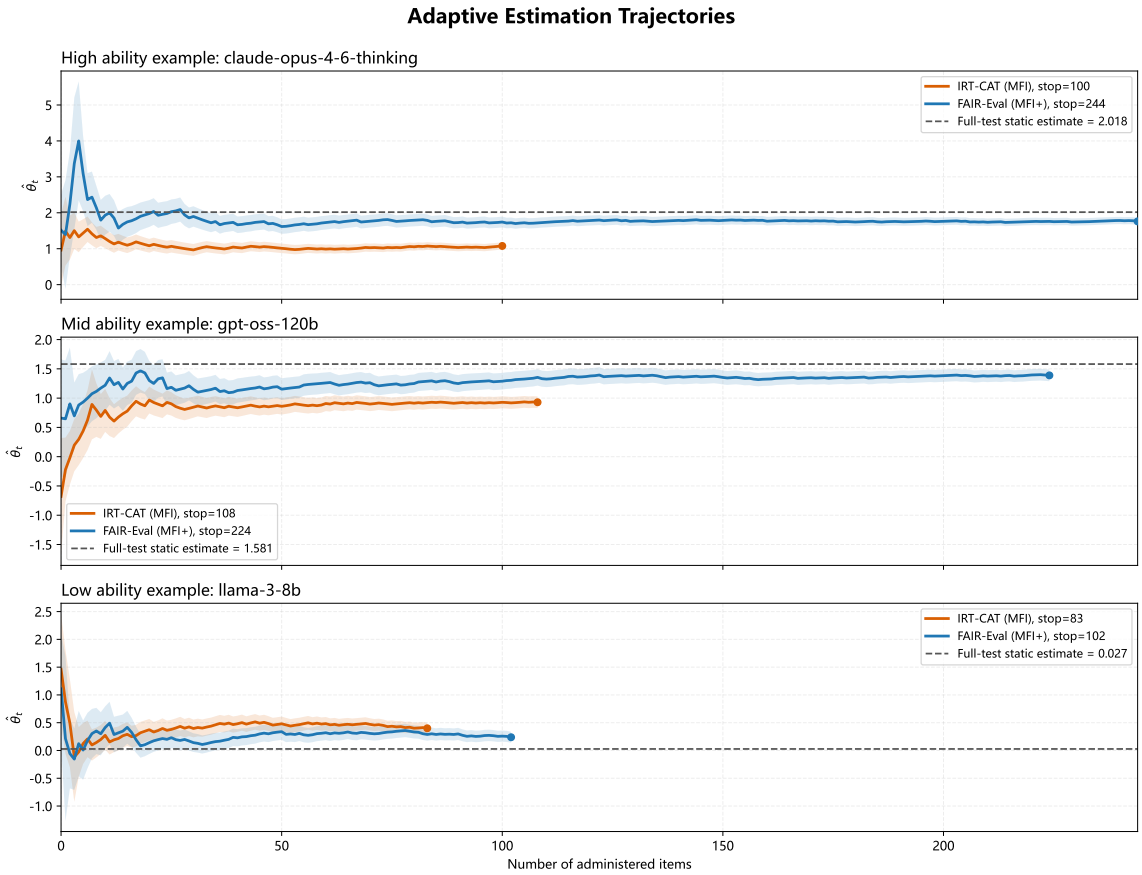


Figure 13: Ability estimation trajectories for representative high-, mid-, and low-ability examples under IRT-CAT (MFI) and FAIR-Eval (MFI+). In each panel, the solid lines show the sequential estimates as items are administered, the shaded bands indicate $\hat{\theta}_t \pm \mathrm{SE}$, and the horizontal dashed line marks the full-test static estimate for the same model. The figure highlights differences in convergence behavior, uncertainty, and stopping points induced by the two selection mechanisms.

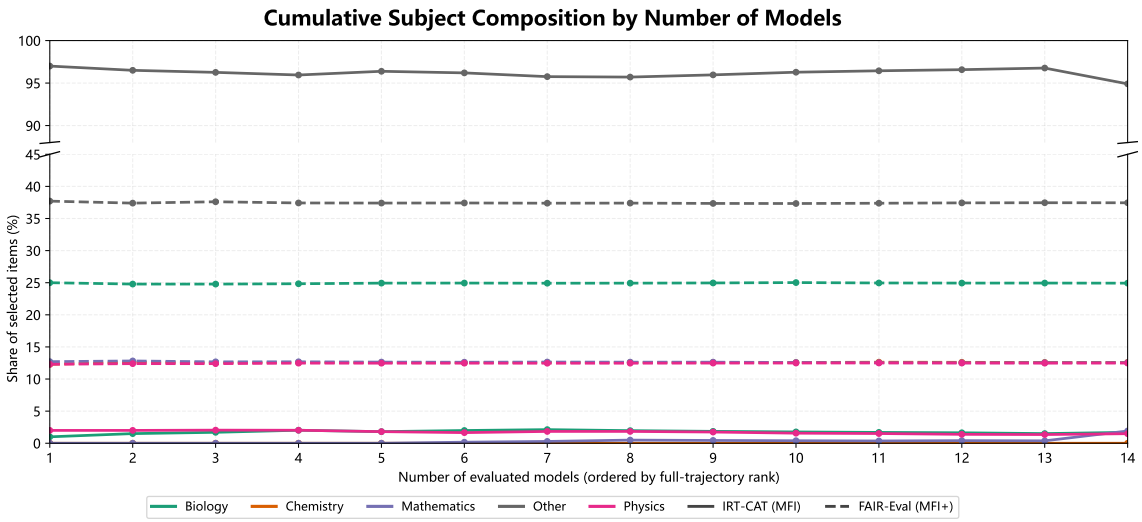


Figure 14: Cumulative subject composition of administered items under IRT-CAT (MFI) and FAIR-Eval (MFI+) as the number of evaluated models increases. The x-axis shows the number of models included, ordered by full-trajectory rank, and the y-axis gives the cumulative share of selected items assigned to each subject. Colors denote subjects and line styles distinguish the two methods, revealing how they differ in subject allocation over the growing set of models. A broken y-axis is used for readability.

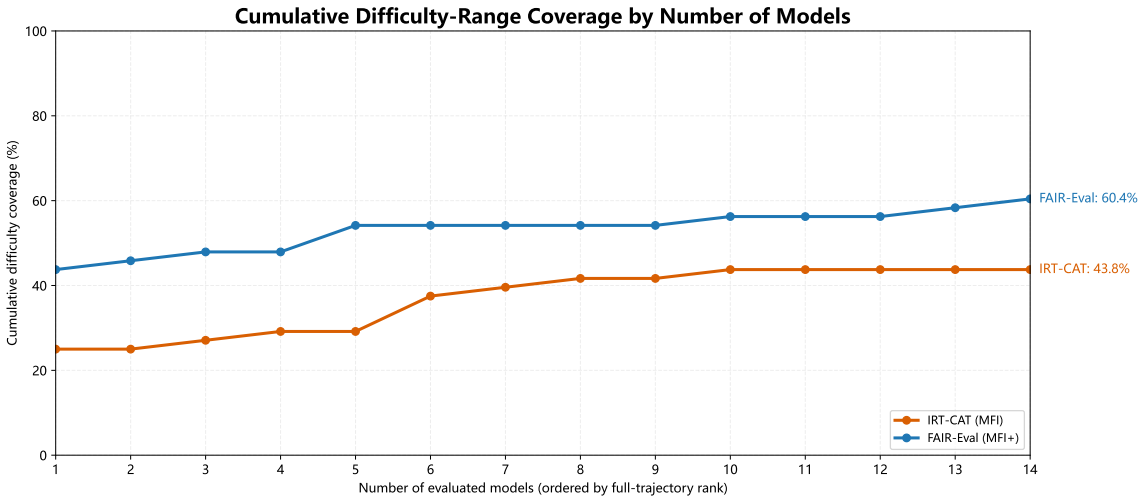


Figure 15: Cumulative difficulty-range coverage under IRT-CAT (MFI) and FAIR-Eval (MFI+) as the number of evaluated models increases. The x-axis shows the number of models included, ordered by full-trajectory rank, and the y-axis reports the percentage of difficulty bins covered by the union of administered items. The curves characterize how broadly the two methods explore the item bank's difficulty range across models.

Table 10: Per-model CAT results: ability estimates and the number of administered items.

No.	Model	$\hat{\theta}_{\text{FAIR-Eval}}$ ↑	Items Used	SE ↓
1	**claude-opus-4-6-thinking***	2.004	280	0.1
2	grok-4.20-beta-0309-reasoning	1.960	272	0.1
3	claude-opus-4-6	1.931	271	0.1
4	claude-opus-4-5-20251101	1.909	265	0.1
5	o3-2025-04-16	1.894	265	0.1
6	**kimi-k2.5***	1.871	260	0.1
7	qwen3-max-preview	1.871	260	0.1
8	qwen3-max-2025-09-23	1.867	259	0.1
9	claude-opus-4-20250514	1.863	255	0.1
10	**gemini-2.5-pro***	1.849	252	0.1
11	**deepseek-v3.2***	1.821	251	0.1
12	claude-opus-4-1-20250805	1.810	247	0.1
13	deepseek-v3.1	1.782	247	0.1
14	qwen3-vl-235b-a22b-instruct	1.782	243	0.1
15	qwen3-235b-a22b-instruct-2507	1.748	237	0.1
16	**qwen3.5-plus***	1.731	236	0.1
17	gpt-5.3-chat-latest	1.714	236	0.1
18	o1-2024-12-17	1.708	229	0.1
19	grok-4-fast-reasoning	1.663	231	0.1
20	qwen3-next-80b-a3b-instruct	1.640	226	0.1
21	gpt-4.1-2025-04-14	1.638	222	0.1
22	o4-mini-2025-04-16	1.637	223	0.1
23	grok-3-mini-beta	1.633	226	0.1
24	**deepseek-r1***	1.628	225	0.1
25	deepseek-v3	1.612	222	0.1
26	**gpt-oss-120b***	1.579	222	0.1
27	claude-sonnet-4-20250514	1.565	221	0.1
28	deepseek-v3-0324	1.559	216	0.1
29	claude-sonnet-4-5-20250929	1.526	212	0.1
30	gpt-5.4	1.525	212	0.1
31	claude-3-5-sonnet-20240620	1.525	216	0.1
32	**gpt-5.2-chat***	1.522	214	0.1
33	**gemini-2.5-flash***	1.453	207	0.1
34	**claude-3-7-sonnet-20250219***	1.448	206	0.1
35	o3-mini	1.430	204	0.1
36	qwen3-coder-480b-a35b-instruct	1.410	200	0.1
37	gpt-4o-2024-08-06	1.388	199	0.1
38	qwen3-30b-a3b-instruct-2507	1.367	197	0.1
39	gpt-4o-2024-05-13	1.366	200	0.1
40	qwen3-235b-a22b	1.344	197	0.1
41	claude-3-haiku-20240307	1.310	191	0.1
42	claude-3-5-haiku-20241022	1.288	190	0.1
43	gpt-4.1-mini-2025-04-14	1.276	190	0.1
44	qwen-max-0919	1.267	189	0.1
45	**qwen2.5-72b-instruct***	1.208	186	0.1
46	qwen3-32b	1.206	184	0.1
47	gpt-oss-20b	1.143	183	0.1
48	gpt-4-0314	1.135	179	0.1
49	**qwen3-30b-a3b***	1.015	169	0.1
50	gpt-4-0613	0.983	168	0.1
51	qwen2.5-coder-32b-instruct	0.806	151	0.1
52	gpt-4.1-nano-2025-04-14	0.775	154	0.1
53	**qwen2.5-7b-instruct***	0.505	150	0.1
54	**llama-3-8b-instruct***	0.445	120	0.1
	Mean	-	**215.7**	**0.1**

Note. Open-weight model series appearing in this table include DeepSeek (Liu et al., 2024; Guo et al., 2025; Liu et al., 2025), Qwen (Hui et al., 2024; Qwen et al., 2025; Yang et al., 2025), GPT-OSS (OpenAI et al., 2025), Llama (Grattafiori et al., 2024), and Kimi (Team et al., 2026). Closed-source API models are identified by provider version names and are not individually cited here.