DIAGNOSTIC FINE-GRAINED ABILITY PROFILING FOR LARGE LANGUAGE MODEL EVALUATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Evaluating large language models (LLMs) is a prerequisite for characterizing their ability boundaries and for guiding subsequent model improvement and real-world deployment. Current practices, however, rely predominantly on aggregate benchmark scores, which conceal the heterogeneity of underlying abilities and exhibit inconsistent rankings across evaluation settings. To address this limitation, we propose a diagnostic framework that estimates multidimensional ability profiles directly from test item level responses. This framework dynamically constructs a structured Q-matrix that explicitly represents the associations between test items and the fine-grained abilities they require. It then integrates this mapping with multidimensional Item Response Theory (mIRT) to jointly estimate an interpretable ability profile for each model, rather than collapsing distinct abilities into a single aggregate score. Instantiated with a 35-dimensional mathematical ability taxonomy, this approach demonstrates strong criterion validity (ability scores aligning with observed accuracy on associated items), high cross-benchmark reliability (consistent model rankings across benchmark pairs), and accurate prediction on held-out items. Furthermore, it generalizes to evaluation in physics, chemistry, and computer science, preserving high validity and reliability under varying item distributions. Overall, our framework replaces the single aggregate score with a multidimensional ability profile, providing a principled basis for more reliable, ability-aware evaluation in support of model improvement and deployment.

1 INTRODUCTION

Large language models (LLMs) have achieved remarkable progress across diverse domains, with performance on standardized benchmarks approaching or even surpassing human proficiency in specific tasks (Qin et al., 2025; Wu et al., 2024; Wang et al., 2025). Effective evaluation of these models characterizes their specific strengths and weaknesses, thereby informing targeted model refinement and task-specific model selection. However, current evaluation practices compress diverse item-level responses into a single aggregate score. The resulting simplification masks underlying fine-grained ability differences, and model rankings often inconsistent across different benchmarks.

Table 1: Ranking instability under different scenarios. (a) Across-benchmark variation: models show inconsistent rankings across mathematical benchmarks despite targeting the same domain. (b) Across-categories variation: models exhibit divergent rankings across categories in MMLU-Pro.

(a) Scores in different benchmarks of MATH

Models	MMLU-Pro-MATH	MMLU-MATH	MATH500
Gemma3-12b-it	0.7454	0.9078	0.8300
Qwen2.5-32b	0.6640	0.7928	0.7040
Qwen3-30b-a3b	0.7765	**0.9345**	**0.6620**
Qwen3-14b	0.7698	0.9078	0.7160
GPT-oss-20b	0.8734	0.9599	0.8700
Qwen3-32b	0.8024	0.9318	0.7120

(b) Scores in different categories of MMLU-Pro

Models	Physics	Computer Science	Chemistry	Math
Gemma3-12b-it	**0.6043**	0.5829	0.5866	**0.7454**
Qwen2.5-32b	**0.6089**	0.6390	0.5804	**0.6640**
Qwen3-30b-a3b	Q0.6390	0.7220	0.5989	0.7765
Qwen3-14b	0.6613	0.7122	0.6148	0.7698
GPT-oss-20b	0.7306	0.7171	0.7447	0.8734
Qwen3-32b	0.7367	0.7512	0.6890	0.8024

We conceptualize fine-grained abilities as the underlying attributes that drive model responses. Rather than viewing model performance as a single scalar score, we assume that solving a specific task requires a particular combination of these abilities. Because different benchmarks cover these abilities differently, aggregate scores tend to fluctuate across different scenarios. As shown in

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.

Table 1, rankings exhibit inconsistency across different benchmarks and domains. In part (a), which compares different benchmarks targeting mathematical, rankings fluctuate markedly (e.g., Qwen3-30b-a3b achieved the second-best ranking on MMLU-MATH but the worst ranking on MATH500.). In part (b), which examines subject categories within the same benchmark, relative ordering also varies (e.g., Gemma3-12b-it trails Qwen2.5-32b in Physics but leads in Math). These results confirm that aggregate scores obscure the fine-grained ability distribution that determine task-level performance, motivating the need for diagnostic evaluation beyond single scores.

To operationalize such diagnosis, recent studies have increasingly adapted psychometric frameworks from educational assessment to LLM evaluation, establishing them as promising methods for fine-grained ability assessment (Liu et al., 2023; Yao et al., 2026; Zhao et al., 2025; Zhou et al., 2026a). Theoretical foundations such as Classical Test Theory (CTT) (DeVellis, 2006), Item Response Theory (IRT) (Harvey & Hammer, 1999), and multidimensional IRT (mIRT) (Ackerman, 2014) underpin these approaches, as they were originally developed to infer examinees' underlying ability from item-level responses. Nevertheless, existing adaptations exhibit critical limitations in accurately modeling fine-grained abilities. Most current works treat benchmark items as monolithic units under unidimensional assumptions, lacking explicit mappings between item components and the specific fine-grained abilities they engage. Although a recent study (Zhou et al., 2026b) extends this line to multidimensional assessment, it still operates on a fixed, predefined ability taxonomy, treating model abilities as immutable dimensions rather than dynamic, context-dependent constructs jointly shaped by benchmark designs and task contexts. Consequently, current psychometric approaches either obscure variations across fine-grained abilities or rely on rigid taxonomies, failing to capture the ability disparities that drive the benchmark-to-real-world performance discrepancy.

In this paper, we propose a diagnostic framework that estimates a multidimensional fine-grained ability profile for LLMs. Grounded in established cognitive theories and domain expertise, the framework first constructs a domain-specific ability taxonomy. It then explicitly encodes item-ability associations via a structured Q-matrix, specifying which fine-grained abilities each benchmark item elicits. By integrating this explicit mapping with multidimensional Item Response Theory (mIRT), the framework enables ability-specific analysis of each model. This profile addresses the coarseness of aggregate scoring, enabling nuanced model comparison and interpretable diagnosis.

Experiments on 41 LLMs demonstrate the framework's validity, reliability, and predictive power. For 91% of the ability dimensions, the estimated ability score correlates with model accuracy on the associated items, indicating strong criterion validity. Furthermore, the framework maintains high cross-benchmark reliability, with 86% of the ability dimensions present in any two benchmarks exhibiting at least moderate consistency. It also predicts model performance more accurately than baselines in both within- and cross-benchmark settings (achieving an average AUC gain of 0.2). Finally, the framework extends to ability evaluation in physics, chemistry, and computer science.

The main contributions of this work are summarized as follows:

- We propose a diagnostic framework that shifts LLM evaluation from the single score to a multidimensional ability profile. By integrating item-level responses with multidimensional Item Response Theory (mIRT), the framework estimates interpretable, dimension-specific ability scores, enabling explicit diagnosis of fine-grained model abilities.
- We develop a benchmark-adaptive Q-matrix construction pipeline that explicitly associates test items with fine-grained ability dimensions. Rather than relying on fixed taxonomies, this pipeline adapts the diagnostic dimensions to each benchmark's structure and task context, so that the estimated ability scores reflect measurable, context-specific competencies.
- We empirically validate the framework across 41 LLMs and 4 scientific domains. The estimated profile aligns with accuracy on associated items, exhibits reliability across benchmarks, and predicts model performance on held-out items more accurately than baselines.

2 Diagnostic LLM Evaluation via Fine-Grained Abilities

2.1 Fine-Grained Ability Taxonomy

Our ability taxonomy is inspired by Bloom's Taxonomy (Forehand, 2010) and organizes model competence along two dimension categories: domain knowledge and cognitive processes. As il-

lustrated in Figure 1, instantiating the framework in the mathematical domain, we construct a 35-dimensional fine-grained ability taxonomy. Each dimension is further decomposed into fine-grained, interpretable subcategories that serve as the basis for diagnostic evaluation.

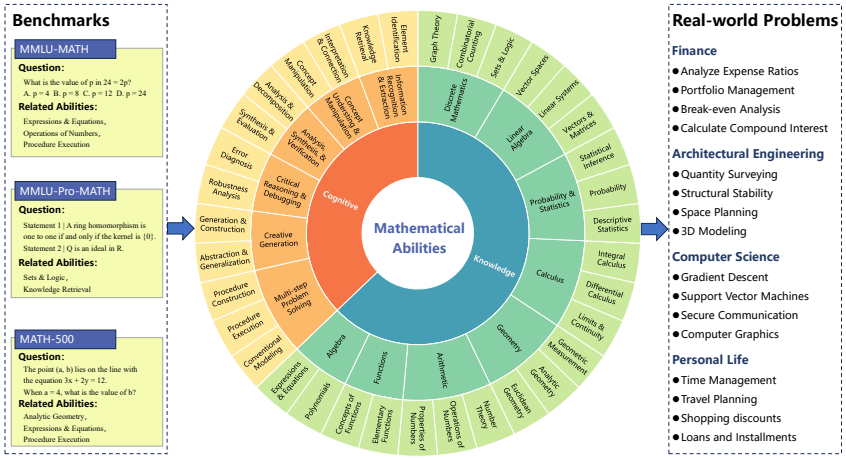

Figure 1: Hierarchical structure of mathematical abilities based on Bloom's Taxonomy. The central diagram organizes ability profiles into Knowledge and Cognitive dimensions, comprising fine-grained abilities. The figure illustrates the relationship between evaluation and application: benchmark questions (left) are mapped to these fine-grained abilities, which correspond to real-world problem domains (right). This structure connects standardized assessment metrics with practical task requirements, enabling diagnostic evaluation beyond aggregate scores.

Domain knowledge refers to the model's mastery of static, factual, or procedural content within a specific discipline. In the context of mathematical reasoning, we define knowledge subcategories, such as Algebra, Functions, Arithmetic, and Geometry, by aligning with standard mathematical curricula and the scope of the benchmarks used. **Cognitive processes** capture the model's capacity to apply knowledge in task-solving contexts. Rather than directly adopting Bloom's original cognitive levels, which were designed for human learners, we adapt them to the LLM setting by focusing on observable reasoning behaviors. As shown in Appendix D.2, this yields six operationalized cognitive dimensions. These dimensions are defined based on the types of reasoning required to solve benchmark items, ensuring alignment between ability definitions and empirical evaluation.

2.2 Overview of Evaluation Framework

We construct a diagnostic evaluation framework that decomposes model performance on existing benchmarks into fine-grained abilities. These abilities encompass domain knowledge (e.g., algebraic rules, geometric facts) and cognitive processes (e.g., multi-step reasoning, problem decomposition). As shown in Figure 2, the framework comprises three phases. (1) A model-item response matrix (R-matrix) derived from existing benchmark evaluations without modifying the original assessment protocol, where entries indicate model correctness on individual items; (2) An item-ability association matrix (Q-matrix) that explicitly maps each benchmark item to the fine-grained abilities it engages; (3) Multidimensional Item Response Theory (mIRT) that estimates latent ability levels for each model based on the Q-matrix and R-matrix.

Model-Item Response Matrix Construction. We evaluate each target model on all items in a given benchmark and record its responses in binary form, where 1 indicates a correct answer and 0 indicates an incorrect or incomplete response. This yields an $m \times n$ response matrix $\mathbf{R}$, with m denoting the number of models and n the number of benchmark items. This matrix serves as the empirical basis for subsequent diagnostic inference.

Item-Ability Association Matrix Construction. The association matrix, denoted $\mathbf{Q} \in \{0,1\}^{n \times k}$, specifies which abilities are required to solve each item of benchmarks. Here, k is the total number of associated abilities, for example, "recalling trigonometric identities" or "constructing a proof by contradiction". An entity $q_{ij} = 1$ indicates that item i depends on ability j; otherwise, $q_{ij} = 0$.

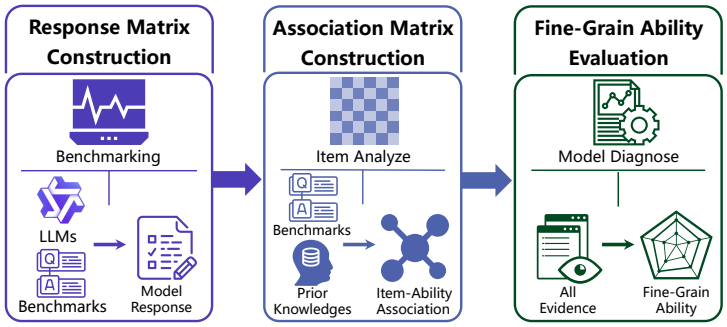

Figure 2: Overview of the fine-grained ability evaluation pipeline. The pipeline comprises three sequential phases: (1) Response Matrix Construction, where LLMs are evaluated on benchmarks to compile item-level response records; (2) Association Matrix Construction, which automatically associates benchmark items with fine-grained abilities using prior knowledge and benchmark analysis to form a structured Q-matrix; and (3) Fine-Grained Ability Evaluation, where multidimensional IRT jointly analyzes the response and association matrices to estimate underlying fine-grained ability scores. Expert validation is optionally applied to ensure Q-matrix reliability.

Fine-Grained Ability Estimation. Given the response matrix $\mathbf{R}$ and the Q-matrix $\mathbf{Q}$, we estimate each model's ability profile using a multidimensional diagnostic model grounded in mIRT. Specifically, the probability that model m answers item i correctly is modeled as a function of its proficiency across the k abilities, along with the item's discrimination and difficulty parameters. These profiles enable interpretable diagnosis beyond aggregate scores.

2.3 Constructing the Q-Matrix via Expert-Supervised LLM Annotation

To enable diagnostic evaluation on existing benchmarks, we construct the Q-matrix, an item-ability association matrix that maps each benchmark item to fine-grained abilities, using a human-in-the-loop annotation pipeline. This pipeline leverages LLMs for scalable ability labeling while incorporating domain experts at critical stages to ensure conceptual fidelity and alignment with educational standards. The process proceeds in three stages. Throughout the process, stage (b) undergoes full expert review, while stages (a) and (c) are spot-checked by experts at a 10% random sampling rate, with corrections applied to the entire dataset when systematic errors are detected.

(a) Ability Pool Generation. For each benchmark item, we prompt an LLM to identify the fine-grained abilities it assesses, covering both domain knowledge (e.g., Algebra, Geometry) and cognitive processes (e.g., Multi-step Problem Solving, Logical Inference). The union of all identified abilities across items forms an initial ability pool.

(b) Hierarchical Ability Dimensions Refinement. Guided by standard mathematical curricula and our adapted Bloom's Taxonomy, we design structured prompts to organize the ability pool into a hierarchical dimensions spanning multiple granularities. This system is then fully reviewed and revised by domain experts to ensure conceptual coherence and alignment with educational standards.

(c) Q-Matrix Construction. From the refined hierarchy, we select a target granularity level as the final set of evaluation dimensions. An LLM is then prompted to map each benchmark item to the relevant dimensions in this set, producing a preliminary Q-matrix.

2.4 Estimating Fine-Grained Abilities via mIRT

Latent trait measurement is a well-established paradigm in psychometrics, with mature methodologies in educational assessment and cognitive diagnosis. To estimate fine-grained ability profiles of LLMs, we adapt mIRT to model the relationship between model responses and underlying competencies spanning both domain knowledge and cognitive processes.

Item Response Theory (IRT) posits a probabilistic relationship between examinees' latent traits and their responses. Based on IRT, our framework enables two key functionalities: (1) estimat-

ing model's fine-grained ability profile from its responses on benchmark items, and (2) predicting model's performance on unseen items by jointly modeling ability profiles and item characteristics.

In practice, we adapt the NeuralCD framework (Wang et al., 2023), a neural extension of mIRT, to estimate LLM ability profiles based on our ability taxonomy (Section 2.1). The NeuralCD framework achieves joint modeling of the model ability profile and item characteristics (i.e., discrimination and difficulty) through an interaction layer and three non-negative fully connected layers. Specifically, the interaction layer is inspired by mIRT and formalized as follows:

$$\boldsymbol{x} = \boldsymbol{q} \odot (\boldsymbol{a} - \boldsymbol{I}_{\text{diff}}) \odot \boldsymbol{I}_{\text{disc}} \tag{1}$$

Here, $\boldsymbol{q} \in \{0,1\}^K$ is the item-ability association vector, $\boldsymbol{a} \in [0,1]^K$ denotes the model's ability profile, $\boldsymbol{I}_{\text{diff}} \in [0,1]^K$ represents item difficulty parameter, $\boldsymbol{I}_{\text{disc}} \in [0,1]^K$ is the item discrimination parameter, To satisfy the bounded range of these parameters, $\boldsymbol{I}_{\text{diff}}$ and $\boldsymbol{I}_{\text{disc}}$ are obtained by applying a sigmoid function to unbounded latent IRT parameters. $\odot$ denotes element-wise multiplication, and $\boldsymbol{x} \in \mathbb{R}^K$ is the input to the subsequent MLP layer. The model's ability profile $\boldsymbol{a}$ is computed via:

$$\boldsymbol{a} = \text{sigmoid}(\boldsymbol{M}_m \times \mathbf{A}) \tag{2}$$

where $\boldsymbol{M}_m \in \{0,1\}^{1\times M}$ is a one-hot vector indicating the m-th model, and $\mathbf{A} \in \mathbb{R}^{M\times K}$ is a trainable matrix whose rows represent the fine-grained ability profiles of all M evaluated models. This formulation allows us to jointly learn item parameters $(\boldsymbol{I}_{\mathbf{diff}}, \boldsymbol{I}_{\mathbf{disc}})$ and model ability profiles $\mathbf{A}$ yielding an interpretable, multi-dimensional profile beyond a signle score.

3 Experiment and Result Analysis

3.1 Experiment Setup

We evaluate our fine-grained diagnostic framework on a diverse set of benchmarks and models. The effectiveness of the framework is validated from three perspectives: (1) diagnostic utility of the fine-grained ability estimates, (2) cross-benchmark consistency of ability estimates, and (3) predictive performance on held-out items. We further assess cross-domain generalization by extending the evaluation to three subjects.

For the mathematical domain, we use three widely adopted benchmarks: MMLU-MATH, MMLU-Pro-MATH, and MATH500. As analysis in Appendix E.1, the three mathematical benchmarks exhibit distinct distributions of item coverage across ability dimensions. Our evaluation encompasses 35 open-source LLMs and 6 online LLMs. All models are evaluated through the OpenCompass benchmarking suite (OpenCompass Contributors, 2023), which provides consistent prompt templates, decoding strategies, and scoring rules across benchmarks. For each model, a binary response matrix is generated by evaluating it on all benchmark items, with correct answers encoded as 1 and incorrect responses as 0. Further details of the experimental setup are provided in Appendix E.

3.2 Criterion Validity of Fine-Grained Ability Profiles

Validity refers to the extent to which a measurement tool accurately reflects the construct it is intended to assess. We focus on criterion-related validity, which evaluates whether estimated ability scores correlate with observed performance on relevant tasks. To this end, we compute Spearman's rank correlation coefficient (ρ) between estimated ability scores for each dimension and model's accuracies on the set of benchmark items linked to that dimension. For each dimension, accuracy is defined as the mean binary accuracy across all items associated with that dimension in the Q-matrix. High correlation indicates that the estimated ability levels reliably capture the model's actual performance on relevant evaluation items.

Results across the three mathematical benchmarks are visualized in Figure 3. We define correlation strength as follows: strong ($\rho > 0.7$), moderate ($0.5 < \rho \leq 0.7$), weak ($0.3 < \rho \leq 0.5$), and non-significant ($\rho \leq 0.3$ or $p > 0.05$).

To assess the robustness of our diagnostic framework, we examine the consistency of criterion validity for each of the 35 ability dimensions across all three mathematical benchmarks. This analysis

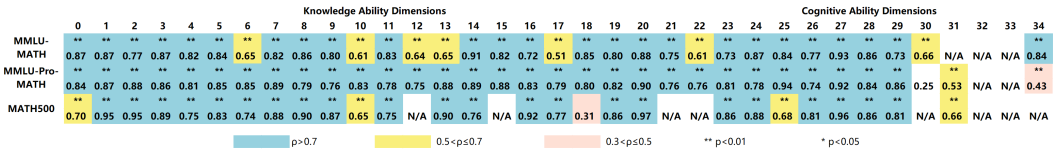

Figure 3: **Correlation between estimated fine-grained abilities and model performance on associated benchmark items.** Spearman's rank correlation coefficient (ρ) measures the association between each ability dimension (columns) and model accuracy on its linked items across three benchmarks (rows). **Color coding**: Blue indicates $\rho > 0.7$ (strong), yellow indicates $0.5 < \rho \leq 0.7$ (moderate), and red indicates $0.3 < \rho \leq 0.5$ (weak). **Significance markers**: ** denotes $p < 0.01$, * denotes $p < 0.05$. **N/A** indicates that no items in the benchmark are associated with that dimension (i.e., invalid for evaluation). Dimensions with consistent strong correlation across benchmarks reflect stable diagnostic signals; weak correlations often arise from low item coverage (< 10 items).

reveals two key observations. **First**, 27 dimensions exhibit strong criterion correlation in at least two benchmarks, demonstrating stable diagnostic power. Notably, 19 dimensions maintain strong correlation across all three benchmarks. The consistency of these signals across diverse item sets confirms that our framework captures generalizable model competencies. **Second**, dimensions with weak or non-significant correlations are consistently associated with sparse item coverage. As shown in Figure 6, these dimensions correspond to subsets containing fewer than 10 questions. For such low-sample dimensions, correlation estimates are inherently unstable, leading to reduced validity. Crucially, no dimension shows high correlation in one benchmark but low in another, indicating that our Q-matrix and ability taxonomy produce coherent, cross-dataset interpretable profiles. All reported correlations with moderate or stronger strength have p-values < 0.01, confirming statistical significance. Together, these results validate that our method provides reliable, fine-grained diagnostics, provided sufficient items per dimension are available.

3.3 Cross-Benchmark Consistency of Fine-Grained Ability Estimates

Building upon the within-benchmark validity analysis, we further evaluate the generalization of our framework across benchmarks with differing distributions of ability dimensions. To ensure reliable comparison, we retain only those ability dimensions for which both benchmarks contain more than 10 associated items. This threshold is chosen to mitigate instability from sparse item coverage.

For each ability dimension shared by a pair of benchmarks, we compute Spearman's rank correlation coefficient (ρ) between the estimated scores across the evaluated models, quantifying the ranking consistency of the two benchmarks on that dimension. The results are visualized in Figure 4, with the same correlation theshold as in the within-benchmark scenario. All reported correlations with $\rho > 0.3$ are statistically significant at $p < 0.05$; those with $\rho > 0.5$ satisfy $p < 0.01$. Overall, the fine-grained ability estimates demonstrate strong cross-benchmark consistency. Among the 19 ability dimensions that are valid across all three benchmarks (i.e., each has more than 10 associated items in every benchmark), 14 dimensions exhibit moderate or stronger correlation ($\rho > 0.5$) across all three benchmark pairs, indicating robust generalization. Across all pairwise correlations, the median is $\rho = 0.66$, and 86% reach at least moderate strength ($\rho > 0.5$).

These results confirm that our method produces stable, generalizable ability profiles across diverse mathematical benchmarks, reinforcing its suitability for evaluation.

3.4 Predicting Model Performance on Unseen Items

Grounded in mIRT, our framework enables prediction of correct or incorrect responses to unseen items based on the association between each item and fine-grained abilities.

We evaluate this predictive capability on three mathematical benchmarks, and use the area under the ROC curve (AUC) as the evaluation metric. For each scenario, we compute AUC separately for each of the 41 models, then report the distribution of these per-model AUC scores. Experiments are structured into two settings, with results visualized in Figure 5.

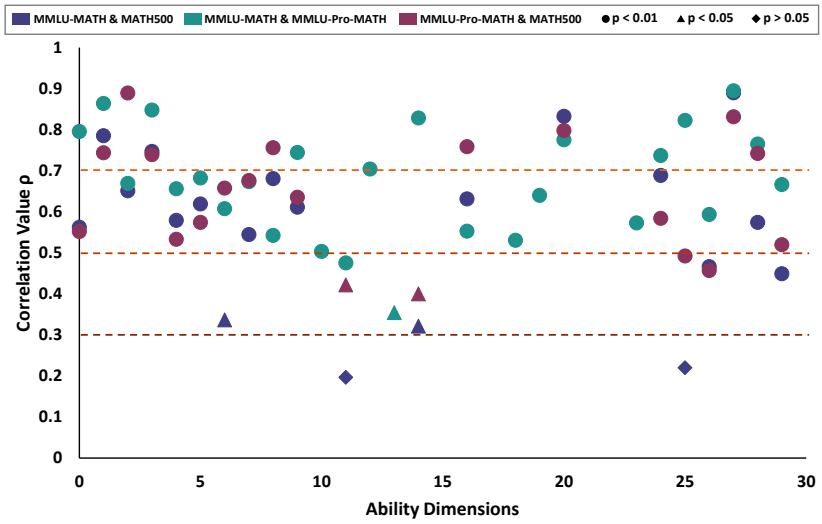

Figure 4: Cross-benchmark correlation of estimated fine-grained ability levels. Spearman's ρ between paired benchmarks, computed over 19 of 35 ability dimensions with $>$ 10 items in all benchmarks. Color: purple = MMLU-MATH vs MATH500, blue = MMLU-MATH vs MMLU-Pro-MATH, red = MMLU-Pro-MATH vs MATH500. Shape: squares ($p < 0.01$), triangles ($p < 0.05$), diamonds ($p > 0.05$). Dashed lines mark thresholds: $\rho = 0.3$ (weak), 0.5 (moderate), 0.7 (strong). 14 of these 19 dimensions show $\rho > 0.5$ across all pairwise comparisons.

(a) Within-Benchmark Prediction. We randomly split each benchmark into an evaluation set (90%) and a prediction set (10%), repeating this procedure 10 times to account for partition variance. We estimate model ability profile using the evaluation set, then predict responses on the prediction set. For unseen items, since no response data is available, we follow common IRT conventions and set the item difficulty to 0 and discrimination to 1. After the same sigmoid transformation used in our model, these become $I_{\text{diff}} = 0.5$, $I_{\text{disc}} \approx 0.73$.

(b) Cross-Benchmark Prediction. We construct six ordered benchmark pairs (e.g., MMLU-MATH $\rightarrow$ MATH500), using the source benchmark for ability estimation and the target for prediction. Target items use the same fixed parameters ($I_{\text{diff}} = 0.5$, $I_{\text{disc}} \approx 0.73$).

As shown in Figure 5, we compare against three baselines using an identical procedure for predicting on unseen items: (1) **Adele** (Zhou et al., 2026b): a recent approach that extracts model ability profiles using general scales and automated rubric-based annotation to predict performance on novel tasks, achieving mean AUCs of 0.60 ± 0.09 (MMLU-MATH), 0.64 ± 0.09 (MMLU-Pro-MATH), and 0.67 ± 0.15 (MATH500); (2) **Standard unidimensional IRT**: models a single latent trait (AUC = 0.500); (3) **Random prediction**: samples probabilities from $\mathcal{N}(0.5, 0.2^2)$ truncated to $[0, 1]$, yielding AUCs of 0.4968 ± 0.1157 (MMLU-MATH), 0.4708 ± 0.1326 (MMLU-Pro-MATH), and 0.4992 ± 0.0586 (MATH500). Our framework achieves a mean AUC greater than 0.85 for within-benchmark predictions, while its cross-benchmark results range from 0.77 to 0.86. Our framework consistently outperforms all three baselines across these settings, reliably predicting model behavior on unseen items in different scenarios.

3.5 Cross-Domain Generalization of the Diagnostic Framework

To evaluate the generalizability of our evaluation framework beyond mathematical reasoning, we extend the analysis to three additional domains: physics, chemistry, and computer science, using the corresponding subsets from the MMLU-Pro benchmark. Following the pipeline in Sections 2.3–2.4, we construct domain-specific Q-matrices via expert-supervised LLM annotation, aligning ability dimensions with each domain's knowledge structure.

For criterion validity, we analyze only those ability dimensions with sufficient item coverage (> 10 associated items). The proportion of dimensions achieving moderate or stronger correlation ($\rho >$

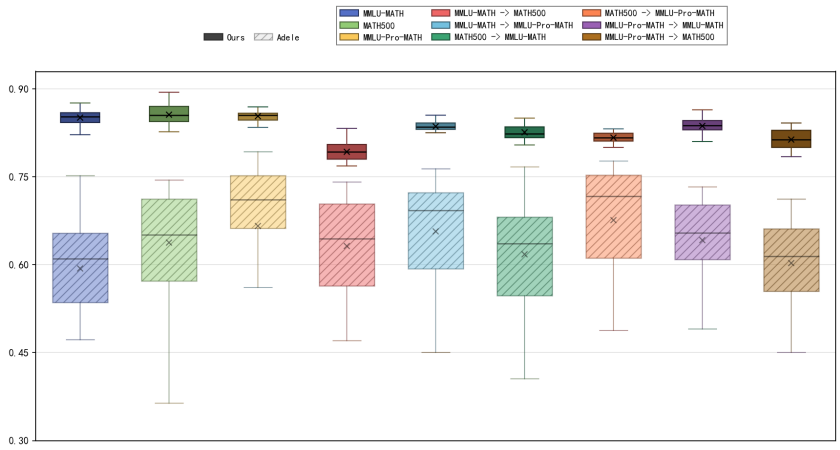

Figure 5: Distribution of AUC scores for unseen item prediction across 41 LLMs. The first three groups represent within-benchmark evaluation, while the remaining six correspond to cross-benchmark generalization. Solid bars denote our proposed framework, and hatched bars indicate the baseline method (Adele). Boxplots display the median, interquartile range (IQR), and outliers (whiskers = 1.5×IQR). Our approach consistently outperforms the baseline across all settings.

0.5) is as follows: **Chemistry**: 48 out of 58 dimensions (82.8%); **Physics**: 27 out of 27 dimensions (100%); **Computer Science**: 10 out of 12 dimensions (83.3%). For prediction performance, the AUC values are 0.7664 in chemistry, 0.7925 in physics, and 0.7824 in computer science. These results are comparable to those observed in mathematical benchmarks, indicating that the framework maintains consistent predictive power when transferred to new scientific and technical domains.

These findings confirm that our method can be effectively extended beyond mathematics without loss of diagnostic or predictive fidelity.

4 Related Work

4.1 Large Language Model Evaluation

Current evaluation of LLMs relies primarily on standardized benchmarks that report aggregate performance metrics. The MMLU benchmark (Hendrycks et al., 2020) evaluates cross-domain knowledge through multiple-choice questions spanning 57 subjects, while MATH and GSM8K (Cobbe et al., 2021) assess mathematical reasoning using specialized problem sets. More recent efforts such as Humanity's Last Exam (Phan et al., 2025) compile expert-level challenges to probe advanced deductive capabilities. Despite their widespread adoption, these benchmarks treat each task as a monolithic unit, collapsing diverse cognitive and knowledge-based processes into a single scalar score. This coarse-grained approach yields inconsistent model rankings across benchmarks and provides limited insight into the underlying sources of success or failure. To address this limitation, several works have introduced more structured evaluation schemes. AtomThink (Xiang et al., 2024) models complex reasoning as compositions of atomic operations, illustrating how higher-level capabilities emerge from basic modules. FineMath (Liu et al., 2025) decomposes elementary mathematics into 17 problem types to enable fine-grained analysis of reasoning pipelines. SciEval (Sun et al., 2024) proposes a four-dimensional taxonomy, including knowledge, application, analysis, and evaluation, to measure 24 scientific competencies. These approaches, however, rely on hand-crafted taxonomies that reflect designer-specific assumptions about ability structure. They also tend to assess abilities in isolation or focus on generative modeling rather than diagnostic evaluation. None of them establishes a shared, psychometrically grounded ability space that supports cross-benchmark comparison under heterogeneous distributions, interpretable diagnosis of model ability profiles, or prediction of model behavior on unseen items.

4.2 Psychometrics and Cognitive Diagnosis

Psychometrics provides a formal framework for measuring latent traits such as cognitive abilities and domain knowledge through observable responses. Classical Test Theory (CTT) (DeVellis, 2006) models an observed score as the sum of a true ability and random error, but treats ability as a single global construct. Item Response Theory (IRT) (Harvey & Hammer, 1999) models the probability of a correct response as a function of both examinee ability and item properties, such as difficulty and discrimination, enabling more precise individual assessment. mIRT (Ackerman, 2014) extends this paradigm by representing ability as a vector of fine-grained traits, allowing items to depend on multiple underlying dimensions simultaneously. This advancement is formalized in Cognitive Diagnosis Models (CDMs) (De La Torre, 2009), which explicitly define the relationship between test items and required skills via a Q-matrix. The Q-matrix is a binary mapping that indicates which attributes each item measures. Recent work integrates neural networks into this framework. NeuralCD (Wang et al., 2023) replaces linear link functions with deep architectures to model complex interactions between skills and items, while preserving the interpretability of the Q-matrix. Collectively, these approaches establish a principled foundation for decomposing performance into interpretable ability components, a perspective increasingly relevant to the evaluation of LLMs.

4.3 Diagnosis Model-based LLM Evaluation

Recent work has begun to adapt psychometric models, particularly Item Response Theory (IRT), to improve algorithmic and LLM evaluation. Castleman et al. (2025) demonstrate that datasets like GSM8K and MathOdyssey fail to reliably assess state-of-the-art models, advocating for IRT-based analysis to enhance evaluation validity. Liu et al. (2023) apply IRT to recommendation algorithms, showing that high rankings do not necessarily reflect progress in user preference modeling. Within the LLM domain, Zhou et al. (2026a) integrate IRT with neural networks to mitigate the confounding effects of benchmark difficulty, while Rodriguez et al. (2021) quantify how item characteristics influence model rankings. However, these foundational approaches predominantly model a single latent trait and do not estimate multidimensional ability profiles. To capture multidimensional capabilities, recent methods have extended classical psychometric frameworks. JE-IRT (Yao et al., 2026) jointly embeds LLMs and items into a shared space to decompose aggregate scores into multiple ability dimensions, and Zhao et al. (2025) adopt a Bayesian framework to predict performance in vision-language tasks. However, these embedding-based methods typically rely on matrix completion techniques that require partial response records for target items, limiting their zero-shot predictive power on entirely unseen tasks. Alternatively, Adele (Zhou et al., 2026b) analyzes model capabilities and task requirements through 19 predefined dimensions and predicts model performance on tasks. However, this approach still relies on a fixed taxonomy, which limits its dimension coverage across different domains or tasks.

5 Conclusion and Future Directions

We propose a cognitive diagnostic framework for fine-grained LLM evaluation grounded in multidimensional Item Response Theory (mIRT). The framework constructs domain-specific ability taxonomies by integrating cognitive theory with domain knowledge structures; for mathematics, this yields a 35-dimensional taxonomy spanning domain knowledge and cognitive processes. Evaluated on 41 models, the framework demonstrates strong criterion validity, consistent ability estimates across benchmarks, and accurate prediction of unseen items. The framework generalizes across scientific domains, producing consistent diagnostic performance in physics, chemistry, and computer science. By estimating interpretable fine-grained ability profiles, our approach provides a principled foundation for LLM assessment with potential to ability-guided model selection, ability-aware benchmark design, and alignment between benchmark evaluations and real-world task demands.

Limitations include the manual effort required for taxonomy construction and Q-matrix annotation, potential subjectivity in mapping items to fine-grained abilities, and uneven ability coverage in existing benchmarks. Future work will extend the framework to additional domains, develop ability-aware benchmarks with balanced coverage, and explore ability-guided model training based on diagnostic profiles.

AI DISCLOSURE

In this work, we used generative AI tools for generating synthetic datasets and assisting with translation. We have not used generative AI tools for designing or providing feedback on research methodology or experiments, implementing methods, cleaning or reformatting datasets, supporting qualitative or thematic data analysis, or interpreting results, and developing theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, or proposing or refining hypotheses are not applicable to this work. We have reviewed all AI-assisted work.

"Generate synthetic datasets" refers to our Q-matrix construction pipeline via expert-supervised LLM annotation, as detailed in Section 2.3. This pipeline is organized into three stages: (a), (b), and (c). To ensure annotation reliability, stage (b) undergoes full expert review, while stages (a) and (c) are spot-checked by domain experts at a 10% random sampling rate, with corrections propagated to the entire dataset whenever systematic errors are identified. Additionally, we provide a supplementary analysis of the framework's validity under complex LLM-generated annotations in Appendix A. As for "assist with translation" and "draft parts of a research paper", all LLM-generated text was verified for correctness by two authors.

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 have made significant efforts to ensure the reproducibility of our work. The source code and the raw LLM response data on the benchmarks are publicly available in our anonymous repository: anonymous.4open.science/r/FineGrainedEvaluation-7B48AWF/. The prompts used for LLM-assisted Q-matrix construction are fully documented in Appendix C. Furthermore, detailed experimental setups and parameters are provided in Appendix E, and all benchmarks utilized in this study are publicly accessible. We believe these resources provide sufficient information for reproducing our results.

REFERENCES

Terry A Ackerman. Multidimensional item response theory models. *Wiley StatsRef: Statistics Reference Online*, 2014.

Jane Castleman, Nimra Nadeem, Tanvi Namjoshi, and Lydia T Liu. Rethinking math benchmarks for llms using IRT. *Proceedings of Machine Learning Research*, 273:66–82, 2025.

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

Jimmy De La Torre. A cognitive diagnosis model for cognitively based multiple-choice options. *Applied Psychological Measurement*, 33(3):163–183, 2009.

Robert F DeVellis. Classical test theory. *Medical care*, 44(11):S50–S59, 2006.

Mary Forehand. Bloom's taxonomy. *Emerging perspectives on learning, teaching, and technology*, 41(4):47–56, 2010.

Robert J Harvey and Allen L Hammer. Item response theory. *The Counseling Psychologist*, 27(3): 353–383, 1999.

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

Yan Liu, Renren Jin, Ling Shi, Zheng Yao, and Deyi Xiong. FineMath: A Fine-Grained Mathematical Evaluation Benchmark for Chinese Large Language Models. *ACM Transactions on*

Asian and Low-Resource Language Information Processing, 24(12):1–15, December 2025. ISSN 2375-4699. doi: 10.1145/3773083.

Yang Liu, Alan Medlar, and Dorota Glowacka. What we evaluate when we evaluate recommender systems: Understanding recommender systems' performance using item response theory. In *Proceedings of the 17th ACM Conference on Recommender Systems*, RecSys '23, pp. 658–670, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 979-8-4007-0241-9. doi: 10.1145/3604915.3608809.

OpenCompass Contributors. OpenCompass: A universal evaluation platform for foundation models, 2023.

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

Yihao Qin, Shangwen Wang, Yiling Lou, Jinhao Dong, Kaixin Wang, Xiaoling Li, and Xiaoguang Mao. SoapFL: A standard operating procedure for LLM-based method-level fault localization. *IEEE Transactions on Software Engineering*, 51(4):1173–1187, 2025. doi: 10.1109/TSE.2025.3543187.

Pedro Rodriguez, Joe Barrow, Alexander Miserlis Hoyle, John P. Lalor, Robin Jia, and Jordan Boyd-Graber. Evaluation examples are not equally informative: How should that change NLP leaderboards? 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. 4486–4503, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.346.

Liangtai Sun, Yang Han, Zihan Zhao, Da Ma, Zhennan Shen, Baocai Chen, Lu Chen, and Kai Yu. SciEval: A Multi-Level Large Language Model Evaluation Benchmark for Scientific Research. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(17):19053–19061, March 2024. ISSN 2374-3468, 2159-5399. doi: 10.1609/aaai.v38i17.29872.

Fei Wang, Qi Liu, Enhong Chen, Zhenya Huang, Yu Yin, Shijin Wang, and Yu Su. NeuralCD: A General Framework for Cognitive Diagnosis. *IEEE Transactions on Knowledge and Data Engineering*, 35(8):8312–8327, August 2023. ISSN 1041-4347, 1558-2191, 2326-3865. doi: 10.1109/TKDE.2022.3201037.

Xu Wang, Qisheng Xu, Kele Xu, Ting Yu, Bo Ding, Dawei Feng, and Yong Dou. Large pretrained foundation model for key performance indicator multivariate time series anomaly detection. *IEEE Open J. Comput. Soc.*, 6:176–187, 2025. doi: 10.1109/OJCS.2024.3521217.

Xun Wu, Shaohan Huang, Guolong Wang, Jing Xiong, and Furu Wei. Boosting text-to-video generative model with mllms feedback. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 139444–139469. Curran Associates, Inc., 2024. doi: 10.52202/079017-4426.

Kun Xiang, Zhili Liu, Zihao Jiang, Yunshuang Nie, Runhui Huang, Haoxiang Fan, Hanhui Li, Weiran Huang, Yihan Zeng, Jianhua Han, et al. Atomthink: A slow thinking framework for multimodal mathematical reasoning. *arXiv preprint arXiv:2411.11930*, 2024.

Louie Hong Yao, Nicholas Jarvis, Tiffany Zhan, Saptarshi Ghosh, Linfeng Liu, and Tianyu Jiang. A geometric lens on LLM abilities through joint embedding item response theory. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856.

Qinyu Zhao, Ming Xu, Kartik Gupta, Akshay Asthana, Liang Zheng, and Stephen Gould. Can We Predict Performance of Large Models across Vision-Language Tasks? *arXiv preprint arXiv:2410.10112*, May 2025.

Hongli Zhou, Hui Huang, Ziqing Zhao, Lvyuan Han, Huicheng Wang, Kehai Chen, Muyun Yang, Wei Bao, Jian Dong, Bing Xu, Conghui Zhu, Hailong Cao, and Tiejun Zhao. Lost in Benchmarks? Rethinking Large Language Model Benchmarking with Item Response Theory. *Proceedings of*

the AAAI Conference on Artificial Intelligence, 40(41):35085–35093, March 2026a. ISSN 2374-3468, 2159-5399. doi: 10.1609/aaai.v40i41.40814.

Lexin Zhou, Lorenzo Pacchiardi, Fernando Martínez-Plumed, Katherine M. Collins, Yael Moros-Daval, Seraphina Zhang, Qinlin Zhao, Yitian Huang, Luning Sun, Jonathan E. Prunty, Zongqian Li, Pablo Sánchez-García, Kexin Jiang-Chen, Pablo A. M. Casares, Jiyun Zu, John Burden, Behzad Mehrbakhsh, David Stillwell, Manuel Cebrian, Jindong Wang, Peter Henderson, Sherry Tongshuang Wu, Patrick C. Kyllonen, Lucy Cheke, Xing Xie, and José Hernández-Orallo. General scales unlock AI evaluation with explanatory and predictive power. *Nature*, 652(8108): 58–67, April 2026b. ISSN 0028-0836, 1476-4687. doi: 10.1038/s41586-026-10303-2.

A Component Validation Analysis

We validate the component reliability of our diagnostic pipeline on the MMLU-Pro-MATH benchmark. Evaluation focuses on two key metrics: *within-benchmark AUC* (predictive accuracy) and *#Strong corr. dims* (discriminative validity against external criteria, $r > 0.5$). These experiments verify that diagnostic fidelity stems from robust methodological design rather than configuration sensitivity, confirming stable performance under annotation uncertainty and structural perturbations.

A.1 Impact of Granularity and Dimension Composition

We first examine the sensitivity of our diagnostic framework to taxonomy configuration, focusing on granularity and dimension composition. As shown in Table 2, varying the number of dimensions from 14 (*coarse*) to 135 (*fine-grained*) yields statistically comparable within-benchmark AUC scores (0.8570 ± 0.0088, 0.8565 ± 0.0087, and 0.8498 ± 0.0079, respectively). The marginal performance variance across granularity levels indicates that our joint estimation mechanism is robust to dimensional resolution, allowing practitioners to select the granularity that best aligns with computational constraints or diagnostic resolution requirements without compromising predictive validity.

Table 2: Impact of Taxonomy Configuration on Diagnostic Performance. We evaluate five taxonomy variants varying in granularity and dimension composition. Results show that diagnostic fidelity remains stable across granularity levels but degrades when decoupling knowledge and cognitive constructs, validating the necessity of integrated modeling.

Taxonomy	#Dims	#Strong corr. dims	Within benchmark AUC
original(Ours)	35	30/33(0.91)	0.8565±0.0087
coarse	14	12/12(1.00)	0.8570±0.0088
fine-grained	135	120/129(0.93)	0.8498±0.0079
knowledge-only	22	22/22(1.00)	0.8257±0.0090
cognitive-only	13	8/11(0.73)	0.7657±0.0107

In contrast, decoupling knowledge and cognitive constructs significantly degrades diagnostic fidelity. The *knowledge-only* configuration (22 dims) reduces AUC to 0.8257 ± 0.0090, while the *cognitive-only* setup (13 dims) suffers a pronounced drop to 0.7657 ± 0.0107, accompanied by a lower proportion of criterion-aligned dimensions (0.73 vs. 0.91 in our integrated taxonomy). This confirms that mathematical proficiency emerges from the synergistic interplay of domain-specific knowledge and higher-order cognitive operations. Isolating either component obscures latent ability signals and weakens discriminative power, validating our design choice to jointly model knowledge-cognitive dimensions. Collectively, these results demonstrate that while granularity serves as a flexible hyperparameter, the theoretical integration of knowledge and cognitive dimensions is indispensable for maintaining robust diagnostic performance.

A.2 Robustness to Q-Matrix Structural Noise

Table 3: Impact of Q-matrix noise at varying proportions on diagnostic performance. We flip Q-matrix entries at three different proportions to evaluate the robustness of our framework against Q-matrix perturbations. Results show that noise at all three levels causes only a slight degradation in diagnostic performance, demonstrating that our proposed framework is robust to Q-matrix noise.

Ratio	#Strong corr. dims	Within benchmark AUC
original	30/33	0.8565±0.0087
0.05	30/33	0.8475±0.0090
0.10	29/33	0.8454±0.0092
0.15	29/33	0.8443±0.0079

To assess the framework's resilience to annotation errors, we simulate structural noise by randomly flipping Q-matrix entries at ratios $\epsilon \in \{0.05, 0.10, 0.15\}$, mimicking realistic mis-specifications in item-ability mappings. As shown in Table 3, diagnostic performance degrades gracefully under increasing perturbation: within-benchmark AUC decreases marginally from 0.8565 ± 0.0087 (original) to 0.8443 ± 0.0079, while the proportion of criterion-aligned dimensions remains stable

($30/33 \rightarrow 29/33$). This graceful degradation indicates that our joint estimation mechanism inherently compensates for sparse structural misspecifications through probabilistic inference over latent ability profiles. Notably, even at 15% annotation noise, the framework maintains strong discriminative validity, confirming that LLM-assisted or crowdsourced Q-matrix construction—despite imperfect precision—is sufficient for reliable diagnostic purposes. These results validate the practical deployability of our method in scenarios where expert-validated item-ability mappings are unavailable or cost-prohibitive.

A.3 Reliability of LLM-Assisted Annotation

Table 4: Consistency of Q-Matrix Annotations. Agreement metrics (Cohen's κ, raw agreement) with LLM baselines (full benchmark) and human validation (20% random sample). Consistent moderate-to-substantial agreement ($\kappa \in [0.62, 0.73]$) confirms annotation reliability.

Q matrix	Cohen Kappa([-1,1])	Raw Agreement([0,1])
Deepseek 3.2	0.62	0.91
Gemini3 Pro	0.73	0.93
Human	0.61	0.91

To further validate annotation reliability, we conduct dual-source consistency analysis: LLM-based evaluation on the full benchmark and human expert validation on a random 20% sample from MMLU-Pro-MATH. Three human raters independently annotated these items following the methodology of Zhou et al. Zhou et al. (2026b). As shown in Table 4, our Q-matrix annotations demonstrate moderate-to-substantial agreement across both validation sources (Cohen Kappa $\kappa \in [0.61, 0.73]$, raw agreement $\in [0.91, 0.93]$). This consistent cross-source agreement confirms that our annotation pipeline yields reliable item-ability mappings without dependence on any single annotator, enabling scalable, cost-effective diagnostic deployment.

B Case Study: Heterogeneity Behind Aggregate Scores

To demonstrate the practical utility of fine-grained diagnostic profiles, we conduct an in-depth analysis focusing on **cross-model heterogeneity**, comparing models with comparable aggregate performance but diverse architectural backgrounds. Based on the criterion validity analysis in Figure 3, we select MMLU-Pro-MATH as our primary benchmark due to its highest coverage of psychometrically sound dimensions. Specifically, we focus on the 30 dimensions that exhibit strong criterion validity ($\rho > 0.7$, $p < 0.01$) across three independent benchmarks, ensuring that all diagnostic claims are grounded in empirically verified signals rather than measurement artifacts.

To better reveal the fine-grained, multi-dimensional capability profiles obscured by overall scores, we selected five models with comparable performance on the MMLU-Pro-MATH dataset for a detailed fine-grained ability evaluation. Table 5 illustrates the variations of ability profile across representative dimensions.

Table 5: Fine-grained capability heterogeneity behind similar aggregate scores. Five models with comparable aggregate performance on MMLU-Pro-MATH (0.73–0.78) exhibit dramatic variations across ability dimensions, with inter-model score ranges from 0.41 to 0.71. Dimensions are sorted by score range (descending). **Bold** values indicate the highest score per dimension.

	Gemma3-12b-it	**Qwen3-14b**	**Qwen3-30b-a3b**	**Qwen3-4b**	**QwQ-32b**	**Range**
Score of MMLU-Pro-MATH	0.7454	0.7698	0.7765	0.7313	0.7365	0.0452
Element Identification	0.1684	0.6588	0.6182	0.8165	**0.8826**	0.7142
Linear Systems	0.7893	0.6610	0.7285	0.3994	**0.9011**	0.5016
Knowledge Retrieval	0.3824	0.6509	0.7790	**0.8787**	0.8661	0.4963
Graph Theory	0.3769	0.6388	0.6623	0.7758	**0.8640**	0.4870
Sets and Logic	0.4069	0.6681	0.8243	0.8887	**0.8916**	0.4847
Concept Manipulation	0.4313	0.5665	0.8151	**0.8871**	0.8624	0.4559
Vector Spaces	0.4633	0.6390	0.7519	0.7542	**0.8922**	0.4289
Integral Calculus	0.3685	**0.6675**	0.5647	0.3314	0.2590	0.4085

Despite their similar aggregate scores (ranging from 0.73 to 0.78) on MMLU-Pro-MATH, the models exhibit extreme polarization in fine-grained capabilities. The most striking disparity occurs in the *Element Identification* dimension, where scores span from 0.17 (Gemma3-12B) to 0.88 (QwQ-32B)—a gap of 0.71 points. Similarly, the *Linear Systems* dimension shows a 0.50 point disparity (0.40 vs. 0.90), and *Sets and Logic* exhibits a 0.48 point gap (0.41 vs. 0.89). These patterns confirm that **aggregate metrics systematically mask fine-grained capability differences across specific dimensions**. Under such conditions, single-dimensional scores fail to capture extreme performance variations and may even lead to misleading conclusions.

C METHODOLOGY FOR FINE-GRAINED ABILITY TAXONOMY AND Q-MATRIX CONSTRUCTION

We employ a human-in-the-loop framework that leverages large language models (LLMs) under expert supervision to construct a fine-grained ability taxonomy grounded in domain knowledge and Bloom's taxonomy, and to subsequently derive the Q-matrix. The process consists of three sequential stages. First, in item analysis, we map benchmark items to relevant cognitive attributes using established educational frameworks (e.g., K-16 mathematics curricula) and Bloom's taxonomy, yielding an initial attribute pool; human experts sample LLM-generated mappings to assess validity and iteratively refine the prompting strategy to improve alignment with pedagogical standards. Second, during taxonomy construction, we organize this pool into a hierarchical structure: the LLM proposes groupings and provides concise definitions for abilities at each level, which domain experts then fully review to ensure coherence, non-redundancy, and fidelity to educational theory. Third, in Q-matrix construction, we select an appropriate granularity level from the taxonomy as the target ability dimensions for evaluation, and under expert supervision, the LLM annotates each benchmark item with binary indicators denoting required abilities, producing the final Q-matrix. The prompts used in each stage are provided below.

C.1 PROMPT OF ITEM ANALYSIS

In the item analysis stage, we prompt the LLM to map each benchmark item to relevant cognitive attributes based on domain knowledge and Bloom's taxonomy. The prompt used is as follows.

System Prompt:

```
You are a professional mathematics education assessment expert with 15
    ↪ years of experience in mathematics curriculum design and test
    ↪ question analysis. You excel at precisely deconstructing the
    ↪ ability structure required for test questions from the dual
    ↪ perspectives of cognitive psychology and mathematics pedagogy.
Please analyze the following mathematics test question and output **both
    ↪ the specific abilities covered by this question and the reasons why
    ↪  these abilities are required**.
For each type of ability, please include:
1. **What specific abilities are covered** - List the exact abilities
    ↪ needed
2. **Why these abilities are covered** - Explain how each ability is
    ↪ applied in solving this specific question

**Ability Categories:**
- Knowledge-based abilities: Identify specific, concrete mathematical
    ↪ facts, concepts, formulas, theorems, or procedures that the user
    ↪ must recall or recognize. Output must be extremely concise, but not
    ↪  too closely related to the specific formulas or data in the
    ↪ question. Using only standard domain-specific keywords or very
    ↪ short phrases (e.g., 'Pythagorean theorem', 'quadratic formula', '
    ↪ derivative rules'). **Avoid complete sentences or explanatory
    ↪ clauses.**
- Cognitive abilities: Identify the higher-order mental processes and
    ↪ strategic abilities required to apply the knowledge. Output must be
    ↪  concise, using generic, transferable skill names expressed as
    ↪ short noun phrases or gerunds (e.g., 'algebraic manipulation', '
    ↪ logical reasoning', 'modeling', 'interpretation'). **Avoid broad
    ↪ verbs like 'solve' or 'use'; strive for precise, compact terms.**

**Strict Output Requirements:**
- Each ability must be directly relevant to answering the test question
- Knowledge-based abilities should be specific to mathematical concepts
- Cognitive abilities should use professional ability names (e.g., "
    ↪ logical reasoning," "spatial imagination")
- For each ability, provide a clear explanation of why it's needed for
    ↪ this specific question
```

```
- Only analyze abilities and their justifications; do not evaluate
    ↪ difficulty or provide suggestions
- Output strictly in JSON format, without any additional text
- All skills must be in English, using precise and commonly accepted
    ↪ terminology

**Output Format:**
{
  "knowledge_abilities": [
    {
      "ability": "mathematical knowledge point 1",
      "explanation": "why this knowledge is required for solving this
    ↪ question"
    },
    {
      "ability": "specific mathematical knowledge point 2",
      "explanation": "why this knowledge is required for solving this
    ↪ question"
    }
  ],
  "cognitive_abilities": [
    {
      "ability": "cognitive ability name 1",
      "explanation": "why this cognitive ability is applied in solving
    ↪ this question"
    },
    {
      "ability": "cognitive ability name 2",
      "explanation": "why this cognitive ability is applied in solving
    ↪ this question"
    }
  ]
}
```

User Prompt:

```
Analyze the following problem:
[problem text]
```

C.2 Prompt of Taxonomy Construction

In the taxonomy construction stage, human experts provide an initial set of fine-grained knowledge points derived from item analysis. The LLM is then tasked with organizing these into a hierarchical competency framework grounded in educational theory. The prompts used are as follows.

System Prompt:

```
You are an expert in mathematics education and curriculum design. Your
    ↪ task is to construct a comprehensive and well-structured
    ↪ mathematical competency framework based on a set of fine-grained
    ↪ knowledge points provided by the user.

**1. Core Design Principles**
- **Theoretical Anchoring**: The framework design must be primarily based
    ↪  on the  **Revised Bloom's Taxonomy**'s cognitive process dimension
    ↪  (Remember, Understand, Apply, Analyze, Evaluate, Create) and
    ↪ knowledge dimension (Factual, Conceptual, Procedural, Metacognitive
    ↪ ). Additionally, references can be made to classifications in
    ↪ mathematics education, such as  **Declarative Knowledge**  (related
    ↪  to abstraction and logical reasoning) and  **Procedural Knowledge
    ↪ **  (related to computation and modeling).
```

```
- **Granularity Consistency**: This is a key requirement. Each branch of the framework must maintain balanced hierarchical depth and subdivision granularity.
    - **Horizontal Consistency**: All nodes at the same level (e.g., all secondary nodes) should represent competency domains of roughly the same scale.
    - **Vertical Progression**: The division of levels should have clear logic (e.g., "Knowledge Domain -> Core Concept -> Specific Component" or "Cognitive Process -> Specific Practice -> Performance Description"), with a depth typically not exceeding 4-5 layers to avoid some paths being too deep while others are too shallow.

**2. Competency Framework Structure Requirements**
The framework must consist of the following two parts:
- **Knowledge Abilities**: Reflects familiarity with and mastery of mathematical knowledge itself. The structure should embody the internal logic of mathematical content (e.g., Number and Quantity, Geometry, etc.).
- **Cognitive Abilities**: Reflects the application and thinking processes applied to mastered knowledge. The structure should reflect a spectrum of cognitive skills from lower-order to higher-order, integrating mathematical practices such as problem-solving, reasoning, modeling, and communication.

**3. Level Definition and Description Specifications**
You must provide a clear description for **every custom non-leaf node** (i.e., all category and subcategory nodes) in the framework to ensure the intent of the hierarchical construction is explicit and aligns with research consensus. Descriptions should follow this format:
- `"level_definition"`: Define the core connotation of the abilities at this level in 1-2 sentences.
- `"differentiation_criteria"`: Explain the principles (e.g., different knowledge domains, cognitive behaviors, practical contexts) used to differentiate between sibling nodes under the same parent node.
- `"example_verbs_or_actions"`: Provide 2-3 typical action verbs or task examples that align with the abilities at this level (e.g., for an "Apply" level, these could be "execute a computation", "use a formula to solve a routine problem").

**4. Knowledge Point Mapping and Output Format**
- You must ensure that all **fine-grained knowledge points** provided by the user can be accurately categorized as **leaf nodes** under the appropriate parent node at the most granular level of the framework.
- The output must be in strict, directly parsable JSON format, containing no non-ASCII characters. The JSON structure must strictly adhere to the following template, with your "description" for each non-leaf node embedded in the corresponding node's `"description"` field.

**Output Format:**
{
  "knowledge_abilities": {
    "children": {
      "node1": {
        "name": "node1",
        "description": "description of node1",
        "children": {
          "node1-1": {
            "name": "node11",
            "description": "description of node1-1",
```

```
            "children": {
              "node1-1-1": {
                "name": "node111",
                "description": "description of node1-1-1",
                "children": {
                  "node1-1-1-1": "knowledge point 11",
                  "node1-1-1-2": "knowledge point 12"
                }
              }
            }
          }
        }
      }
    }
  },
  "cognitive_abilities": {
    "children": {
      "node2": {
        "name": "node2",
        "description": "description of node2",
        "children": {
          "node2-1": {
            "name": "node21",
            "description": "description of node2-1",
            "children": {
              "node2-1-1": {
                "name": "node211",
                "description": "description of node2-1-1",
                "children": {
                  "node2-1-1-1": "knowledge point 21",
                  "node2-1-1-2": "knowledge point 22"
                }
              }
            }
          }
        }
      }
    }
  }
}

Note: Ensure all provided knowledge points can be found with
    ↪ corresponding parent nodes in the tree structure, and place them at
    ↪  appropriate levels. The output must be valid JSON that can be
    ↪ directly parsed.
```

User Prompt:

```
Please generate the competency framework based on the following knowledge
    ↪  points:
[selected fine-grained abilities]
```

C.3 Prompt of Q-Matrix Construction

In the Q-matrix construction stage, the LLM annotates each benchmark item by mapping it to relevant abilities from the finalized taxonomy. The system and user prompts used are as follows.

System Prompt:

```
You are a professional math test item analysis assistant. Please strictly
    ↪  adhere to the following requirements for the analysis task:

**Task Requirements:**
```

```
1. For each input test question, analyze the mathematical abilities it
    ↪ assesses.
2. **Abilities** must and can only come from the user-provided ability
    ↪ list. Do not add points outside the list.
3. The output must be in **strict JSON format**, with the specific
    ↪ structure detailed below.
4. For each related ability, provide a brief explanatory note justifying
    ↪ why and how the question assesses this ability.

**Analysis Principles:**
1. Read each question carefully to identify the core concepts, formulas,
    ↪ theorems, or methods being tested.
2. Strictly match the identified core concepts against the provided
    ↪ abilities list.
3. When matching, the ability name should be an exact or core synonymous
    ↪ match (as per the given list).
4. Explanations should be specific, pointing to conditions, problems, or
    ↪ solution steps within the question.
5. **All special characters must be expressed in escaped form** to ensure
    ↪  valid JSON format. For example:
   - Double quotes should be escaped as \"
   - Backslashes should be escaped as \\
   - Other JSON special characters should be escaped accordingly
6. **JSON key names must strictly adhere to the following specifications
    ↪ **:
   - Key names **must not contain any extra spaces**
   - Key names must exactly match those in the example.

**JSON Key Examples:** "question_content", "related_abilities", "ability
    ↪ ", "explanation"

**Output JSON Format:**
{
  "question_content": "Given a right triangle with legs of lengths 3 and
    ↪ 4, find the length of the hypotenuse.",
  "related_abilities": [
    {
      "ability": "Pythagorean Theorem",
      "explanation": "The question explicitly gives the two legs of a
    ↪ right triangle and asks for the hypotenuse, requiring direct
    ↪ application of the Pythagorean Theorem \"a^2 + b^2 = c^2\"."
    },
    {
      "ability": "Square Root Computation",
      "explanation": "After applying the Pythagorean Theorem, the solver
    ↪ must compute the square root of 25 to obtain the final answer."
    }
  ]
}

**Output Format Specification:**
- Use the above JSON structure. Place the analysis results for each
    ↪ question in a list if multiple questions are processed (though
    ↪ typically one at a time in our pipeline).
```

User Prompt:

```
Analyze the following problem:
[problem text]
```

D MATHEMATICAL ABILITY TAXONOMY

Following our methodology for constructing fine-grained ability taxonomies grounded in cognitive theory, we analyze three mathematical benchmarks (MMLU-MATH, MMLU-Pro-MATH, and MATH500) through the lens of K-16 mathematics curricula and Bloom's taxonomy. This yields a hierarchical taxonomy for mathematical problem solving.

The final taxonomy comprises 35 fine-grained abilities, refined from an initial set of 45 dimensions through iterative expert review. These abilities are organized into two primary dimensions: *Knowledge* (abilities 0–21) and *Cognitive* (abilities 22–34). The structure is visualized in Figure 1, with the complete numbered list presented below.

D.1 KNOWLEDGE DIMENSION

- **Arithmetic**
  - *Properties of Numbers (ability_0)*: Recognize characteristics such as parity, divisibility, and magnitude.
  - *Operations of Numbers (ability_1)*: Perform computations on integers, fractions, and decimals.
  - *Number Theory (ability_2)*: Apply concepts of primes, divisibility, and modular arithmetic.
- **Algebra**
  - *Expressions and Equations (ability_3)*: Simplify symbolic expressions and solve algebraic equations.
  - *Polynomials (ability_4)*: Manipulate polynomials through addition, multiplication, and factorization.
- **Functions**
  - *Concepts of Functions (ability_5)*: Understand input-output relationships and function properties.
  - *Elementary Functions (ability_6)*: Work with linear, quadratic, exponential, and logarithmic functions.
- **Geometry**
  - *Euclidean Geometry (ability_7)*: Apply axioms and theorems to plane geometric figures.
  - *Analytic Geometry (ability_8)*: Use coordinate systems to represent and solve geometric problems.
  - *Geometric Measurement (ability_9)*: Compute lengths, areas, volumes, and angles.
- **Calculus**
  - *Limits and Continuity (ability_10)*: Analyze function behavior near points or at infinity.
  - *Differential Calculus (ability_11)*: Compute derivatives and apply them to rates of change and optimization.
  - *Integral Calculus (ability_12)*: Evaluate integrals and interpret them as accumulated quantities.
- **Probability and Statistics**
  - *Descriptive Statistics (ability_13)*: Summarize data using measures of center and spread.
  - *Probability (ability_14)*: Calculate likelihoods using classical or theoretical models.
  - *Statistical Inference (ability_15)*: Draw conclusions about populations from sample data.
- **Linear Algebra**
  - *Vectors and Matrices (ability_16)*: Perform operations including addition and multiplication.

  - *Linear Systems (ability_17)*: Solve systems of linear equations using algebraic or matrix methods.
  - *Vector Spaces (ability_18)*: Analyze subspaces, bases, and dimension.
- **Discrete Mathematics**
  - *Sets and Logic (ability_19)*: Apply set operations and logical reasoning.
  - *Combinatorial Counting (ability_20)*: Count arrangements using permutations and combinations.
  - *Graph Theory (ability_21)*: Analyze connectivity and structures in networks.

D.2 Cognitive Dimension

- **Information Recognition and Extraction**
  - *Element Identification (ability_22)*: Recognize and classify mathematical objects such as numbers or symbols.
  - *Knowledge Retrieval (ability_23)*: Recall definitions, formulas, or procedures from memory.
- **Concept Understanding and Manipulation**
  - *Interpretation and Connection (ability_24)*: Link mathematical representations to meanings or prior knowledge.
  - *Concept Manipulation (ability_25)*: Reorganize or transform concepts for new problem contexts.
- **Analysis, Synthesis, and Verification**
  - *Analysis and Decomposition (ability_26)*: Break complex problems into simpler components.
  - *Synthesis and Evaluation (ability_27)*: Combine ideas and judge solution validity or efficiency.
- **Critical Reasoning and Debugging**
  - *Error Diagnosis (ability_28)*: Detect and explain mistakes in reasoning or computation.
  - *Robustness Analysis (ability_29)*: Test solutions under varying conditions or edge cases.
- **Creative Generation**
  - *Generation and Construction (ability_30)*: Create new mathematical objects, models, or representations.
  - *Abstraction and Generalization (ability_31)*: Derive general principles from specific examples.
- **Multi-step Problem Solving**
  - *Procedure Construction (ability_32)*: Design step-by-step methods for novel problems.
  - *Procedure Execution (ability_33)*: Carry out algorithms or computational steps accurately.
  - *Conventional Modeling (ability_34)*: Translate real-world scenarios into mathematical forms.

E EXPERIMENT SETUP DETAILS

We evaluate 41 language models on three mathematical benchmarks (MMLU-MATH, MMLU-Pro-MATH, and MATH500). For each model, we derive a 35-dimensional ability profile using multi-dimensional item response theory (mIRT) with the Q-matrix defined in Appendix B. The validity of these profiles is assessed along three dimensions: (1) construct validity through internal consistency, (2) cross-benchmark reliability, and (3) predictive utility for held-out item performance. Additionally, we evaluate cross-domain generalization by applying the same profiling pipeline to three non-mathematical subsets of MMLU-Pro (Chemistry, Physics, and Computer Science).

For ability estimation, we implement the mIRT-based evaluation model described in Section 2.4, a three-layer fully connected network with hidden dimensions (256, 1024, 128). The model employs CrossEntropyLoss as the loss function and is optimized using Adam with an initial learning rate of 0.001 (decayed by 0.8) and a batch size of 256. All experiments use an NVIDIA RTX 3090 GPU.

E.1 EVALUATED BENCHMARKS

We evaluate models on six subsets across three benchmark suites: MMLU, MMLU-Pro, and MATH500. The full list is presented in Table 6.

Table 6: List of evaluated benchmarks with detailed metadata.

ID	Subject	Name	# Items	Source	Answer Format
1	Math	MMLU-MATH	748	cais/mmlu	Multiple-choice
2		MMLU-Pro-MATH	1351	TIGER-Lab/MMLU-Pro	Multiple-choice
3		MATH500	500	HuggingFaceH4/MATH-500	Free-response
4	Chemistry	MMLU-Pro-Chemistry	1132	TIGER-Lab/MMLU-Pro	Multiple-choice
5	Physics	MMLU-Pro-Physics	1299	TIGER-Lab/MMLU-Pro	Multiple-choice
6	Computer science	MMLU-Pro-CS	410	TIGER-Lab/MMLU-Pro	Multiple-choice

As shown in Figure 6, we further quantify the coverage of each of the 35 fine-grained abilities across the three mathematical benchmarks (MMLU-MATH, MMLU-Pro-MATH, MATH500) by counting the number of items associated with each ability in the Q-matrix.

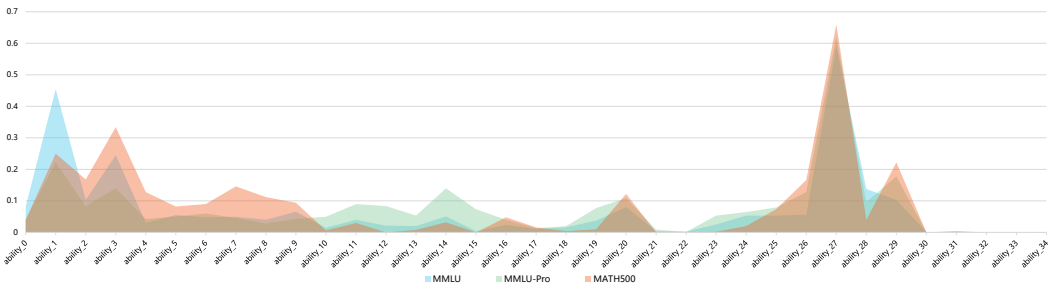

Figure 6: **Item coverage across the 35 fine-grained abilities in three mathematical benchmarks.** Each curve shows the proportion of a benchmark's items associated with each ability. The distinct shapes of the three curves indicate that the benchmarks cover and weight these abilities differently.

E.2 EVALUATED MODELS

We evaluate a total of 41 language models across two categories: online API models and open-source models. The full list is presented in Table 7.

E.3 CONSISTENCY ANALYSIS

We adopt the Spearman rank correlation coefficient (ρ) as the primary measure of association for three reasons. First, ability estimates from multidimensional item response theory (mIRT) are

Table 7: List of evaluated language models with unique identifiers.

ID	Category	Series	Model	Parameter Size
1	Closed Source Model	DeepSeek	Deepseek-v3.2	N/A
2			Deepseek-v3	N/A
3		GPT	GPT-4o	N/A
4			GPT-5.1	N/A
5			GPT-o1	N/A
6			GPT-o3	N/A
7	Open Source Model	Gemma3	Gemma3-12b-it	12B
8			Gemma3-1b-it	1B
9			Gemma3-4b-it	4B
10		GPT	GPT-oss-20b	20B
11		Llama3	Llama3-8B-Instruct	8B
12			Llama3-8B	8B
13		Llama3.1	Llama3.1-8B-Instruct	8B
14			Llama3.1-8B	8B
15		Llama3.2	Llama3.2-1B-Instruct	1B
16			Llama3.2-1B	1B
17			Llama3.2-3B-Instruct	3B
18			Llama3.2-3B	3B
19		Mistral	Mistral-7B-Instruct-v0_1	7B
20			Mistral-7B-Instruct-v0_2	7B
21			Mistral-7B-Instruct-v0_3	7B
22		Qwen2.5	Qwen2.5-0.5b-instruct	0.5B
23			Qwen2.5-0.5b	0.5B
24			Qwen2.5-1.5b-instruct	1.5B
25			Qwen2.5-1.5b	1.5B
26			Qwen2.5-14b-instruct	14B
27			Qwen2.5-14b	14B
28			Qwen2.5-32b-instruct	32B
29			Qwen2.5-32b	32B
30			Qwen2.5-3b-instruct	3B
31			Qwen2.5-3b	3B
32			Qwen2.5-7b-instruct	7B
33			Qwen2.5-7b	7B
34		Qwen3	qwen3-0.6b	0.6B
35			Qwen3-1.7b	1.7B
36			Qwen3-14b	14B
37			Qwen3-30b-a3b	30B
38			Qwen3-32b	32B
39			Qwen3-4b	4B
40			Qwen3-8b	8B
41		QwQ	QwQ-32b	32B

interval-scale but may violate the linearity and normality assumptions required by Pearson correlation. Second, our focus is on *rank-order consistency* of model capabilities (e.g., whether a model consistently ranks higher across abilities), which Spearman directly captures. Third, Spearman correlation is robust to outliers and monotonic nonlinearities, critical given the heterogeneous performance distributions across 41 models and 35 ability dimensions.

To assess the validity and consistency of fine-grained ability profiles, we compute pairwise Spearman rank correlations between model ability vectors across the 35 mathematical dimensions.

Ability-Performance Association. For each ability dimension j, we compute the Spearman correlation between models' estimated abilities on dimension j and their observed accuracy on items associated with that dimension. The observed accuracy of m-th model for dimension j on a benchmark is calculated as:

$$\mathrm{ACC}_{mj} = \frac{\sum_{i=1}^{N} q_{ij} \cdot r_{im}}{\sum_{i=1}^{N} q_{ij}}, \tag{3}$$

where:

- $q_{ij} \in \{0, 1\}$ is the Q-matrix entry indicating whether item i requires ability j,
- $r_{im} \in \{0, 1\}$ is the correctness of m-th model's binary response on item i,
- N is the total number of items in the benchmark.

This metric represents the proportion of correctly answered items among those that associated with ability i.

Cross-Benchmark Consistency. For each model, we compute the Spearman correlation between its ability profiles estimated from different mathematical benchmarks. High consistency indicates stable rank-order ability estimation across benchmark pairs.