

AI Evaluation Should Require Standardized Item-Level Data Releases

Han Jiang¹ Susu Zhang² Dongyao Zhu⁵ Yuzhuo Bai⁶ Sang T. Truong⁴
Xiaoyuan Yi³ Sanmi Koyejo⁴ Xing Xie³ Ziang Xiao¹

¹Johns Hopkins University ²University of Illinois Urbana-Champaign
³Microsoft Research Asia ⁴Stanford University ⁵North Carolina State University
⁶Tsinghua University
hjiang66@jh.edu, ziang.xiao@jhu.edu

Abstract

This position paper argues that standardized item-level benchmark data should become the default infrastructure for AI evaluation. Current evaluations suffer from underspecified item selection, construct misalignment, and poor generalization; a disproportionate focus on aggregate model scores compounds these failures by keeping them undetected and unfixable. Without item-level evidence, validity claims cannot be assessed, resulting in inflated capability claims, misdirected research, and unwarranted trust in deployed systems. Our position is that valid evaluation requires empirical evidence from item-level benchmark data, including item content, test conditions, model responses, and item-level score statistics, and that the standardized release of such data should be treated as core AI evaluation infrastructure. Such releases also enable transparency, replicability, and auditability of evaluation results. To show that the norm is both feasible and consequential, we construct OPENEVAL, an item-level repository of 10.7M+ responses across 160K+ items from widely-used benchmarks, under a unified schema that the AI evaluation community can build upon. We demonstrate how item-level data can identify low-quality items, document construct misalignment, and recover validity evidence about benchmarks’ internal structure. We address objections concerning contamination and coordination burden, and show that both are tractable relative to the costs of decisions made on inadequately validated evaluative claims.

1 Introduction

Generative AI is moving rapidly into high-stakes deployments, while AI evaluation, dominated by benchmarking practice [20], has become the primary instrument for understanding model capabilities, informing AI policy, and guiding responsible deployment. However, the empirical foundation of such instruments is thinner than their influence suggests. Benchmarks are released, leaderboards are populated, and deployment decisions are made, while the underlying response data that would let the community audit whether benchmarks measure what they claim to measure and evaluate benchmark reliability are rarely shared and analyzed.

This gap entails consequences. Critical design choices, including capability definitions, content curation, and metric selection, often lack transparency or formal justification [48]. This opacity undermines the validity evidence [9, 48] needed to support the interpretations of results, making it unclear whether benchmarks genuinely measure their intended constructs, despite explicit metadata or task descriptions [1]. Compounding these design limitations, the speed of AI development manifests as benchmark saturation [55], rapidly outdated content [36], and widespread data contamination [76], rendering aggregate scores uninformative or misleading for deployment decisions [26]. New bench-

marks proliferate to address these concerns, but often recycle existing practices without adding evaluative insight [8], widening the socio-technical gap between technical solutions and real-world requirements [67, 44].

Crucially, many validity issues are not diagnosable from benchmark-level aggregate scores alone. Foundational questions—including whether items effectively differentiate model capabilities, how construct-irrelevant nuisance factors drive performance, or whether gains reflect genuine improvement rather than artifacts—are inherently item-level inquiries. Without item-level benchmark data, our field lacks the empirical evidence required to evaluate and curate effective benchmarks. We argue that **standardized item-level benchmark data should become the default infrastructure for AI evaluation, released alongside aggregate scores whenever legally and ethically feasible**: designing valid evaluations requires empirical evidence from item-level model responses, and the standardized release of such data should be treated as core evaluation infrastructure on par with the benchmarks themselves.

Shifting the focus from aggregate model scores to item-level model outputs has deep roots in measurement science, where item-level assessment data have long been significant to test development and validation across education, psychology, and other social science disciplines. Bringing this granularity to AI evaluation would transform fragmented leaderboard results into cumulative empirical evidence: item content, score statistics, and per-item responses jointly enable rigorous construct validation, reliability analysis through item consistency and discrimination patterns, and efficiency improvements through identification of redundant or uninformative items. Item-level data is the foundation for AI evaluation science.

To show the norm is feasible and consequential, we construct OPENEVAL¹, an item-level archive of 10.7M+ responses across 160K+ items from widely-used benchmarks, organized under a unified schema the community can contribute to. We use OpenEval to demonstrate how item-level data identifies low-quality items, documents construct misalignment, and recovers validity evidence about benchmarks’ internal structure. At the end, we discuss broader opportunities and address objections around storage, contamination, and author burden.

2 Validity Challenges in AI Benchmarking

Despite their pivotal role, AI benchmarks face validity challenges from two directions, namely internal methodological limitations and external pressures from the rapid evolution of AI systems, both of which are exacerbated by the absence of item-level data.

2.1 Methodological Issues in Benchmark Design

The term *Validity*, specifically *Construct Validity*, was formalized by Cronbach and Meehl [16] in modern psychometrics and is directly applicable to AI evaluation [74]. A *Construct* refers to a theoretical, unobservable attribute or trait (e.g., a perceived AI capability) that a test is intended to measure. Validity thus reflects how well an assessment measures the intended underlying outcome [63] and is crucial for ensuring benchmark quality.

As noted in [70], construct validity is a central concept in psychology, particularly for informing the design of psychological measures. In AI evaluation, however, validity has received limited attention, despite AI capabilities rapidly expanding and benchmark creation involving numerous design decisions, such as target capability identification, construct-to-task operationalization, item adaptation for

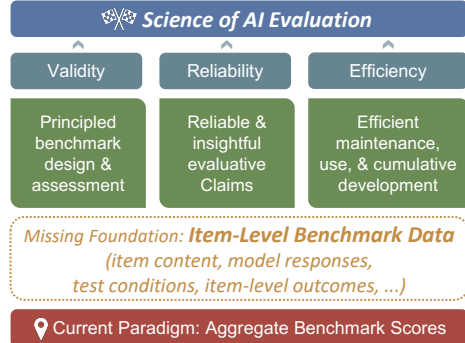


Figure 1: Illustrative overview of our position. Item-level benchmark data serves as the missing foundation between the current aggregate paradigm and the science of AI evaluation.

¹Home page: <https://open-eval.com/>

AI systems, and metric selection. These decisions are often oversimplified due to *a dominant focus on benchmark-level results*, leaving only standard or default settings (e.g., [43, 68]) and insufficient evidence for validity justification [48]. These ambiguities arise when benchmark documentation leaves construct definitions implicit and provides limited item-level evidence supporting the intended score interpretations, making it difficult to articulate and communicate validity concerns consistently, highlighting the need for shared frameworks and guidelines for AI evaluation. Although there are several studies exploring validity-relevant properties (e.g., [79, 73]), research explicitly assessing AI benchmark validity remains limited. Existing analyses suggest that most benchmarks’ validity remains underdeveloped and call for evidence beyond the benchmark level (e.g., [9, 5, 61]).

Benchmarks designed with validity issues act as traps: their claimed goals may sound aligned with users’ intended purposes, yet the underlying constructs being measured could diverge substantially, yielding unreliable conclusions, as demonstrated by [5], that propagate through benchmark selection and all subsequent evaluation practices. Such misalignment can be further exacerbated by confounders unrelated to the intended system ability [78], such as erroneous items, spurious correlations, or unintended shortcuts (e.g., [19, 53]).

Notably, *both construct misalignment and these overlooked confounders are only visible at the item level*, where the content of each item can be scrutinized and its empirical behavior systematically examined. The absence of item-level data hinders reasoning about what actually drives benchmark performance — a question left unresolved from the design stage — which not only wastes the effort invested in benchmark selection, but also leaves the validity of any downstream claim unjustified. Yet existing benchmark data remains largely untapped at this level, leaving benchmark design without the empirical feedback needed to become more principled.

2.2 Pressures from AI Advances on Benchmarking Practices

Beyond the design limitations, the rapid evolution and increasing opacity of AI systems impose external pressures that further erode benchmark validity over time.

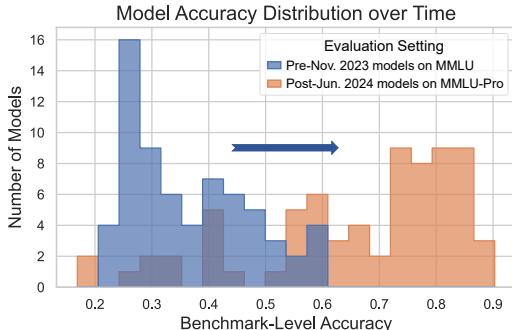


Figure 2: Benchmark-level accuracy distributions for 66 pre–Nov. 2023 models on MMLU and 72 post–Jun. 2024 models on MMLU-PRO.

As AI systems and real-world knowledge co-evolve, benchmarks are subject to various forms of validity degradation [17, 37].

Benchmark saturation. Some benchmarks have gradually become too easy to distinguish between the latest models [55, 18]. For instance, REALTOXICITYPROMPTS [25] is one of the most widely used AI toxicity benchmarks; although by 2023 it could no longer differentiate between multiple versions of GPT [35], it remained widely used beyond 2024.

Outdated knowledge. Benchmarks requiring factual knowledge are particularly time-sensitive, as outdated references render the corresponding test items unreliable [36].

Data contamination. The opacity and information asymmetry in AI system development have made the risk increasingly pervasive. As AI training and evaluation scale, many omnibus benchmarks further aggregate multiple previous benchmarks, making data provenance harder to trace. Both intentional and inadvertent test-train overlap can secretly lead to unfair evaluations unless explicitly reported by developers [80].

All these issues are nearly impossible to detect at benchmark level, and Fig. 2 exemplifies this diagnostic gap: the benchmark-level accuracy distribution has shifted markedly rightward as newer models are evaluated, even though MMLU-PRO was designed to be more challenging. Without item-level inspection, it remains unknown whether the observed improvements reflect genuine capability gains, benchmark saturation, or data contamination, *leaving no valid claims about model capabilities to be drawn*.

This progressive degradation at scale creates a fundamental tension between evaluation efficiency and validity. Manual benchmark updates (e.g., [45, 72]) are believed to generate higher-quality items

but are typically time-consuming and costly. LLM-powered benchmarking (e.g., [38, 22]) improves efficiency but at the expense of validity, as the quality of synthesized test items has been questioned in [10]. More recent benchmark generation techniques, such as adversarial filtering (e.g., [40, 54]), item generation (e.g., [46, 39]), and adaptive difficulty adjustment (e.g., [65]), offer promising directions but require exploring inter-item dynamics and engagement with measurement theories and psychometric methods. Shared item-level data can support further development of dynamic benchmarking and automated item generation by enabling researchers to identify contaminated or outdated items, characterize the constructs they measure, and guide item retention, replacement, and generation as AI systems evolve.

Additionally, a common bottleneck underlying all of these efforts is the lack of shared data infrastructure. Current evaluation results are aggregated locally and not consistently released across leaderboards, making them neither comparable nor easily built upon, and forcing each study to independently assemble and preprocess its data. This duplicates effort across research groups and prevents cumulative progress. Moreover, ensuring the long-term sustainability of a unified item-level data release is itself a real challenge, as the effort required to curate, standardize, and update data across diverse benchmarks is prohibitive for any single research group. That is to say, without *a community-driven, consistent release of item-level benchmark data*, individual efforts to counter these ecosystem pressures remain isolated and unscalable.

3 Item-Level Data: The Missing Foundation for AI Evaluation Science

Given the twofold validity challenges in AI benchmarking and the limits of existing benchmarking practices, **item-level benchmark data is a crucial missing piece in the science of AI evaluation**. Every benchmark run already generates such data: detailed test conditions, the content of each item, the model’s response, and per-response scores and statistics. However, this evidence is routinely discarded once aggregate scores are computed. What remains is a rich yet largely inaccessible foundation for investigating the validity, reliability, and efficiency of AI benchmarking. We argue that releasing item-level data under a unified schema (e.g., Fig. 7) should become a community norm when publishing evaluation results. Reconstituting this evidence into shared infrastructure is a precondition for evaluation to function as a cumulative and open science.

Item-level data enables principled benchmark design and assessment. Item-level data provides an empirical foundation for developing and adapting psychometric methods for AI evaluation. Existing models face challenges in this setting, including limited model samples, violations of statistical assumptions, and distribution shifts as AI systems evolve. Shared item-level evidence enables researchers to examine these limitations and develop measurement models suited to AI evaluation. Through more principled practices such as quantifying item characteristics, identifying decisive benchmark factors, and modeling relationships between items and intended constructs, future benchmarks can better clarify the mapping between AI competencies and evaluation objectives.

Item-level data enables reliable and insightful evaluative claims. Item-level data facilitates a bidirectional attribution that addresses the opacity in AI evaluation caused by the aggregate-focused paradigm. First, observed model performance can be traced to measurable factors, disentangling intended constructs from confounders such as construct-irrelevant shortcuts and data leakage. Second, item properties such as difficulty and discrimination, as well as the latent constructs they collectively reflect, can be investigated for how they manifest across AI systems, revealing where models systematically succeed or fail. Together, these two perspectives ground evaluative claims in empirical item-level evidence, supporting more defensible conclusions about capability and deployment.

Item-level data enables efficient maintenance, use, and cumulative development of AI benchmarks. With increased access to item-level data, score statistics and detailed test cases make it possible to diagnose the validity threats discussed in Sec. 2.2, including benchmark saturation and data contamination, in a timely manner. Item-level analyses can further identify which items have degraded and which remain informative, providing fine-grained guidance for benchmark re-composition, updates, and targeted item generation, saving effort in maintenance rather than multiplying it. Moreover, a consistent, community-wide release of item-level data lowers the costs of data engineering and allows findings to be reproduced, audited, and compared across studies, enabling cumulative progress and prolonging benchmark lifecycles.

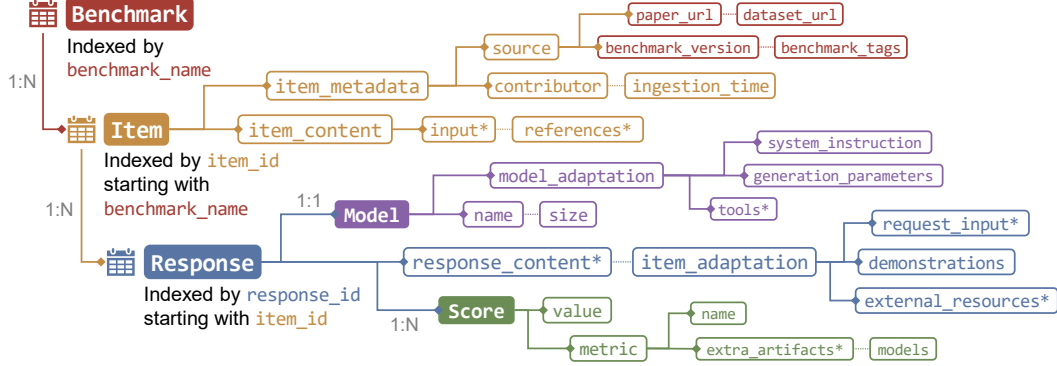


Figure 3: Unified data schema of OPENEVAL. The left column shows the hierarchical storage structure with indexing keys; the right side expands each entity into its constituent fields. Diamond-tipped connectors point from parent to subfield; dashed connectors indicate folded siblings. 1:1 and 1:N denote cardinality, e.g., each response contains exactly one model but may contain multiple scores. Asterisks (*) mark fields accepting values in arbitrary formats.

4 OPENEVAL: An Item-Centered Benchmark Repository

To show what data infrastructure in our position may look like, we propose an item-centered repository, OPENEVAL, designed as a community-driven data infrastructure to standardize benchmark items with model responses, metric scores, and other associated information across different tasks, formats, and evaluation harnesses. OPENEVAL is not the endpoint of our proposal. It is a demonstration that item-level standardization is feasible and a seed for community norms.

There have been many large-scale, high-quality AI benchmark repositories (e.g., HELM [43], Chatbot Arena [12], and Open LLM [23]), some of which release unstructured or semi-structured item-level details alongside the benchmark-level results. While these resources provide a basis for recent AI evaluation research, the shared data infrastructure remains underdeveloped, hindering cumulative progress across studies and groups. OPENEVAL therefore addresses the infrastructure challenge through a **unified, item-centered schema**, as illustrated on the right side of Fig. 3 and detailed in App. C, in which each data entry represents a unique item.

Specifically, the schema is designed around four principles:

- **Unified while remaining flexible.** While effectively standardizing item-level data across different sources, the schema poses relatively loose constraints on fields such as `response_content` and `item_content` to enable incorporation of heterogeneous content. The typed auxiliary arrays (`model_adaptation.tools`, `item_adaptation.external_resources`, and `metric.extra_artifacts`) further accommodate items from various evaluation paradigms, including agent and tool-augmented settings, allowing researchers to focus on research opportunities rather than data engineering overhead.
- **Faithful to evaluation context.** The schema separates the original item content from the adapted input actually given to the model and records the corresponding model configuration, providing a clean snapshot of the test condition for each response rather than collapsing various experimental conditions into a single record. This self-contained, faithful preservation supports reproducibility and enables meaningful comparisons across different experimental setups.
- **Informatively rich and versatile.** The schema archives diverse information at multiple levels, including benchmark-level metadata, test environments, and evaluation metrics, enabling a wide range of practices from provenance tracking and benchmark auditing to coarse- and fine-grained querying by domain, model, or other properties. This richness reflects the paper’s core premise: item-level benchmark data serves the science of AI evaluation not for any single analytical goal, but for a wide range of research opportunities.
- **Scalable for long-term maintenance.** The unified, flexible, and self-contained nature of each entry naturally supports scalability. As shown on the left side of Fig. 3, the repository organizes data into three levels, each indexed by hierarchical identifiers (e.g., `response_id` prefixed by `item_id`). New benchmarks, items, or model responses can be easily incorporated by appending new

entries at the appropriate level without restructuring existing data; this design further facilitates community-driven contributions.

So far, to jump-start the community effort, we have been (1) collecting evaluation results to supplement item-level data coverage across existing benchmark repositories, (2) incorporating interdisciplinary datasets on social aspects of AI, such as CULTURALBENCH [13] and EMOBENCH [60], (3) transforming existing item-level resources from benchmark repositories (e.g., HELM) into OPENEVAL schema, and (4) developing a general AI-assisted converter, complemented by an already published data validator, to accommodate formats beyond those supported by harness-specific converters. OPENEVAL now covers **160K+** items across diverse benchmark datasets, with the number of evaluated models per dataset ranging from 11 to 252, resulting in **10.7M+** item-level responses from 444 distinct models in total, each associated with one or multiple metric scores.

By shifting the norm from reporting aggregate scores to releasing item-level results as a standard part of research publication, we hope that OPENEVAL will grow into a shared data foundation for the science of AI evaluation and responsible AI deployment, leading us toward more rigorous, empirically grounded validation of benchmark-related claims. Beyond the technical infrastructure, we envision community governance mechanisms for communicating item usage history, release status, and contamination reports, alongside shared norms for responsible reuse and auditing (App. B).

5 Empirical Illustrations of Item-Level Analysis

To demonstrate the unique insights enabled by item-level benchmark data, we present illustrative analyses on selected item-level data from OPENEVAL that provide finer-grained understanding of existing AI benchmark datasets, specifically item quality and construct alignment.

5.1 Background and Analytical Approaches

Item-level benchmark analysis has drawn increasing attention in AI evaluation research, with studies examining properties such as item difficulty (e.g., [79, 41]), item discrimination, namely the ability to separate different capability levels (e.g., [31, 42]), item diversity (e.g., [52, 79]), inter-benchmark agreement (e.g., [56, 47]), and downstream performance predictability (e.g., [7, 51, 73]). As discussed in Sec. 2.2, these analyses are typically isolated and difficult to reproduce due to the lack of a large-scale, consistent item-level data release, leaving substantial room for further exploration. Notably, these studies, to varying degrees, draw on psychometric methodology, which offers a well-established toolkit for item-level analysis grounded in decades of test development and validation practice. Two types of analysis from this discipline are particularly relevant to AI evaluation.

The first is item characteristic analysis, which examines whether individual items function as intended with adequate precision for their intended use [2]; *Classical Test Theory* (CTT) and *Item Response Theory* (IRT) are two foundational frameworks for this purpose [14, 28]. CTT measures item difficulty and discrimination directly from observed score matrices with minimal assumptions, making it readily applicable to large-scale, heterogeneous AI benchmark data [27, 50]; IRT provides a more sophisticated, model-based alternative with stronger statistical properties, but has parametric assumptions about the relationship between model ability and item-level performance, with unidimensional formulations requiring careful validation that item responses are driven primarily by a single latent ability [49, 29]. The second is *Item Factor Analysis* (IFA), which provides evidence about a test’s internal structure [4, 58]. The Standards for Educational and Psychological Testing [2] emphasizes that such structural evidence should be provided when a single aggregate score is interpreted as a meaningful summary of examinee ability. In the context of AI evaluation, IFA examines whether a benchmark behaves as a coherent measure of the intended model capability versus reflecting construct-irrelevant variance, such as formatting artifacts, label distribution bias, and data contamination.

5.2 Examining Item Quality via Classical Test Theory

An item’s statistical characteristics such as difficulty and discrimination are routinely examined in psychometric test development for quality assurance. We conduct a CTT analysis of item characteristics from (1) 66 pre-Nov. 2023 models on 567 items in MMLU [32] and (2) 72 post-June 2024 models on 1,000 items in MMLU-PRO [69]. MMLU-PRO [69] is an enhanced variant of MMLU

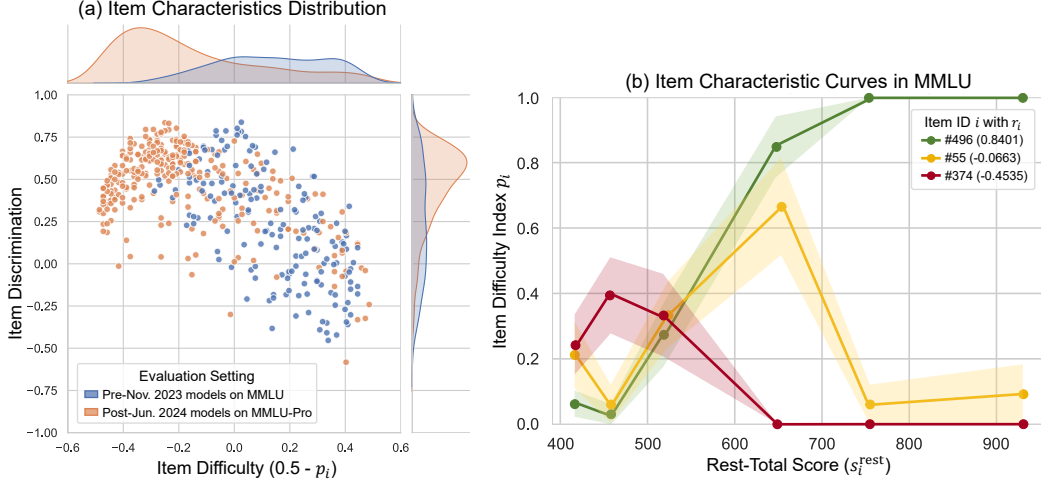


Figure 4: (a) Item characteristics distribution for items from MMLU and MMLU-PRO. Higher item difficulty values correspond to harder items. (b) ICCs for three items in MMLU.

intended to increase difficulty and reduce noise via additional distractors, more careful item curation, and expert item review. CTT item analysis provides an empirical way to evaluate these design claims.

For the i -th item in a benchmark, the *Item Difficulty Index* (p_i) can be estimated as [27] the proportion of the maximum score achieved on the item averaged across models; a larger p_i indicates an easier item. The *Item Discrimination* (r_i) is measured by the Pearson correlation between the item score and the *Rest-Total Score* (s_i^{rest} , sum score on all items except i) across all measured models [33]. Higher r_i indicates that item i can effectively differentiate models with strong vs. weak overall performance on the benchmark, whereas negative or near-zero r_i suggests a potentially problematic item.

Fig. 4(a) shows the distributions of item difficulty and discrimination under the two evaluation settings. (Note that the CTT item difficulties on MMLU and MMLU-PRO are specific to their respective sample of models and are not comparable.) There are two notable observations: (1) The high density of orange observations on the left indicates that a *substantial proportion of MMLU-PRO items have very low difficulty for current models*. In other words, many items are no longer challenging for the 72 post-June 2024 models, suggesting fast benchmark saturation and the need to accelerate benchmark updates. (2) Compared to MMLU, MMLU-PRO *exhibits a smaller proportion of items with low discrimination in the evaluated samples*. This empirical observation is aligned with MMLU-PRO designers’ goal [69] to build a more robust, less noisy benchmark. However, some MMLU-PRO items exhibit poor empirical discrimination despite expert review, warranting further scrutiny for ambiguity, miskeying, construct-irrelevant cues, or measurement of sub-constructs underrepresented in the total score.

To assess the robustness of these comparisons, we conducted supplementary sensitivity analyses in App. D. First, bootstrap resampling and leave-one-developer-out analysis using an expanded pool of 122 models evaluated on the same 1,000 MMLU-PRO items indicate that the item difficulty and discrimination estimates remain relatively stable across the examined changes in model population; second, a difficulty-stratified comparison shows that the observed discrimination advantage of MMLU-PRO persists in most difficulty bins, although differences between the evaluated model samples remain a potential confound.

To further understand how item discrimination manifests, we plot the *Item Characteristic Curves* (ICCs) of three items in MMLU. For each item i , all respondent models are sorted into six equally-sized bins based on their rest-total scores (s_i^{rest}), representing six levels of overall performance. The item difficulty index (p_i) within each bin is then plotted in Fig. 4(b). Intuitively, a high-discrimination item’s p_i should increase with s_i^{rest} , meaning that the item appears easier for higher-performing models, resulting in an overall increasing ICC. This is the case for item #496 which had a high item discrimination ($r_{496}=0.84$), whereas for items #55 and #374 with near-zero or negative r_i , some models that did well on the rest of the benchmark did worse on these items.

These findings can help identify low-quality items in the benchmark, thereby improving the alignment between benchmark outcomes and design goals at a finer granularity. In this sense, detailed examination of item characteristics supports more targeted benchmark maintenance and item generation without requiring wholesale replacement.

5.3 Revealing Construct Alignment via Item Factor Analyses

Table 1: Example BABI item and item counts by reference answers for each cluster in Fig. 5.

Example: Item #1295				
Sheep are afraid of mice. Cats are afraid of mice. Jessica is a sheep. Wolves are afraid of mice. Mice are afraid of wolves. Emily is a wolf. Gertrude is a wolf. Winona is a mouse. Question: What is Emily afraid of?				
Answer: mouse				
Item Answer	Number of Items			
	Cluster #0	Cluster #1	Cluster #2	
Sheep	0	0	240	
Mouse	225	0	0	
Cat	221	3	0	
Wolf	6	326	0	

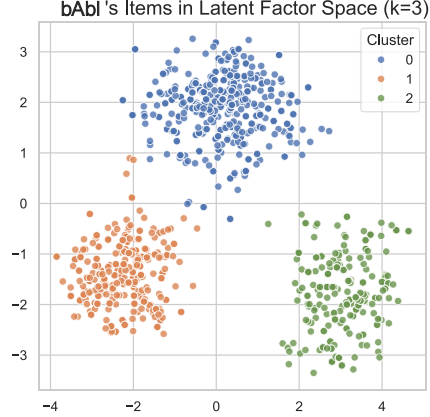


Figure 5: Item clusters on BABI Task 15 based on factor loadings.

As discussed in Sec. 2.1, examination of a benchmark’s internal structure is essential for understanding whether it measures the intended capabilities or irrelevant confounders. Here, we perform variants of conventional IFA for high-dimensional data based on *Singular Value Decomposition* [SVD; 81] and *Generalized Low Rank Models* [GLRM; 66]. We report findings from BABI Task 15 (basic deduction) and MMLU-PRO. Additional analysis results on MMLU are presented in App. E.

BABI Task 15 [71] aims to assess basic deductive reasoning via inheritance of properties. An example item is shown in Table 1. To aid interpretation, we perform K-means clustering on items’ factor loadings on the top 3 factors obtained from SVD-based IFA, with items within each cluster potentially measuring similar sub-constructs. As shown in Fig. 5, 1,000 items in BABI form three distinct clusters. A closer examination raises a construct validity flag: Table 1 shows that item clusters are explained by the reference answer to the item. This finding demonstrates that different models’ performance on BABI is partially explained by models’ propensity to *select specific animals that one is afraid of* (e.g., potentially based on common sense if a model tends to select “wolf”), rather than the intended *basic deduction capability*, suggesting *potential construct misalignment in BABI Task 15*. GLRM-based IFA yields consistent findings.

For MMLU-PRO, we interpret the top four factors retained from GLRM-based IFA after varimax rotation. The top 100 items with the largest absolute loadings on each factor are sent to GPT-5 for interpretation. Table 2 presents a representative item and a sub-construct description for each factor. The four primary dimensions that best explain differences in model performance appear to reflect different higher-level reasoning capabilities, rather than subject domain proficiency. This empirical finding *supports MMLU-PRO’s stated motivation to increase reasoning demands* [69] relative to MMLU. Indeed, items within the same subject domain (e.g., Psychology and Physics in Fig. 13) could differ substantially in loadings on the four factors.

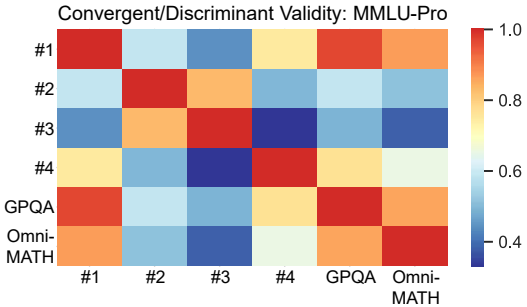


Figure 6: Convergent/discriminant evidence of the four sub-constructs (#1 - #4) on MMLU-PRO.

Table 2: Representative items with large absolute item factor loadings and possible constructs for each GLRM factor in MMLU-PRO. The top 100 representative items are interpreted and summarized into a candidate label by GPT-5, then manually revised.

Factor	Representative Item	Potential Sub-Construct
MMLU-Pro #1	A 10kVA, 2400/240V, single-phase transformer has the following resistances and leakage reactance. Find the primary voltage required to produce 240V at the secondary terminals at full load, when the load power factor is 0.8 power factor lagging/leading.	Formal, quantitative, multi-step modeling
MMLU-Pro #2	What are the principal and selective instruments of control of which the Federal Reserve System makes use?	Domain-specific recall and simple reasoning
MMLU-Pro #3	What is meant by the term “hypothesis testing”?	Conceptual understanding and explanation
MMLU-Pro #4	Which of the following might explain how a price decrease might cause a decrease in quantity demanded and an upward-sloping demand curve?	Applied synthesis and case-based judgment

As an external plausibility check using convergent and discriminant validity evidence [11], we correlate factor subscores (mean scores on the 100 items with the largest absolute loadings) with scores on two external benchmarks: GPQA [59] (graduate-level biology, physics, and chemistry) and OMNI-MATH [24] (Olympiad-level mathematics). Both target high-level formal reasoning, with the former grounded more in applied scientific contexts. We hypothesize that *Factor #1 (formal, quantitative, multi-step modeling) aligns with both formal reasoning benchmarks*, whereas *Factor #4 (applied synthesis and case-based judgment) aligns more with GPQA*. Results in Fig. 6 are broadly consistent with these hypotheses. Further, Factors #2 (domain-specific recall and simple reasoning) and #3 (conceptual understanding and explanation) show weaker correlations with both GPQA and OMNI-MATH, providing discriminant validity evidence. We treat these findings as descriptive rather than definitive evidence supporting these sub-construct interpretations.

6 Broader Implications beyond Benchmarking

The benefits of releasing item-level data extend beyond AI evaluation. Shared infrastructure of this kind could benefit AI research, ground policy in evidence, and democratize AI through participatory evaluation.

For AI Researchers. Releasing such data at scale could contribute to ML theory, where central questions about generalization and learning dynamics are inherently item-level. Methods built on item-level data have advanced principled training: identifying which examples drive learning [64], informing data mixing [15, 75], and grounding curriculum learning [6] through measurable item difficulty. The same per-instance signal supports estimating generalization gaps and attributing predictions to training examples [77].

For Policy Makers. Policy and governance decisions increasingly rely on AI benchmark results to justify claims about model capability, risk, and deployment readiness (e.g., [3, 30, 21]). Item-level data lets policy makers inspect what a benchmark actually measures, enabling contextualized judgments about how much weight a score should carry for a given policy question. The evaluation methods built on such data further equip regulators to audit benchmark design, verify developer claims, and ground policy decisions in independently checkable evidence.

For End Users. The communities affected by AI deployment are largely absent from the evaluations that justify it. Item-level responses, not just items, make user engagement substantive: users can see whether a model actually handles cases like theirs, detect performance gaps that aggregate scores hide, and compare options before deployment [57]. Interfaces built on this infrastructure can lower the technical barrier to participation, making participatory evaluation practical: evaluation as a channel through which users help shape priorities, and through them, model development [44].

7 Alternative Views

Item-level release worsens contamination. A natural objection is that releasing item-level data increases data contamination risk [62, 34]. We argue that openness and contamination prevention need not be contradictory. Information asymmetry about what benchmark items exist and how they have been used complicates both prevention and detection. Standardized releases can help developers identify and filter known benchmark content from training data, while item-level responses and test conditions provide evidence for independent scrutiny of decontamination claims and investigation of suspected contamination. Lowering the cost of external auditing can also help hold developers accountable for undisclosed benchmark use. In addition, our position is compatible with hold-out designs, verified access, and staged release after a fixed embargo period, allowing disclosure practices to account for contamination risks and licensing restrictions. We believe the benefits of independently verifiable evaluation outweigh the potential contamination risks, provided that releases incorporate appropriate safeguards.

Mandating item-level release creates additional burden. Standardized item-level release does require additional effort from benchmark authors and model evaluators. However, because model responses are already generated during evaluation, releasing retained results does not require rerunning inference. The additional work primarily involves conversion, validation, quality control, and uploading. As illustrated by OPENEVAL, shared infrastructure can reduce duplicated effort: contributors submit to a common archive, a unified schema specifies what to release and in what format, and converters for widely used evaluation harnesses reduce the reformatting burden. The cost of individual releases must nevertheless be distinguished from the collective cost of pooling data for analyses across models and benchmarks. Soliciting contributions, coordinating participants, expanding response coverage, and maintaining storage and access require sustained community effort. Standardization makes individual releases easier to combine, reducing the need for each research group to independently assemble and preprocess the same data. Contributors can, in turn, benefit from the accumulated contributions of others, creating incentives for further participation.

Acknowledgments and Disclosure of Funding

This work was supported by Schmidt Sciences.

References

- [1] A. F. Akyürek, M. Y. Kocyigit, S. Paik, and D. T. Wijaya. Challenges in measuring bias via open-ended language generation. In C. Hardmeier, C. Basta, M. R. Costa-jussà, G. Stanovsky, and H. Gonen, editors, *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 76–76, Seattle, Washington, July 2022. Association for Computational Linguistics.
- [2] American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. *Standards for Educational and Psychological Testing*. American Educational Research Association, 2014.
- [3] M. Anderljung, J. Barnhart, A. Korinek, J. Leung, C. O’Keefe, J. Whittlestone, S. Avin, M. Brundage, J. Bullock, D. Cass-Beggs, B. Chang, T. Collins, T. Fist, G. Hadfield, A. Hayes, L. Ho, S. Hooker, E. Horvitz, N. Kolt, J. Schuett, Y. Shavit, D. Siddarth, R. Trager, and K. Wolf. Frontier ai regulation: Managing emerging risks to public safety, 2023.
- [4] T. W. Anderson. *An introduction to multivariate statistical analysis*, volume 2. Wiley, 1958.
- [5] A. M. Bean, R. O. Kearns, A. Romanou, F. S. Hafner, H. Mayne, J. Batzner, N. Foroutan, C. Schmitz, K. Korgul, H. Batra, O. Deb, E. Beharry, C. Emde, T. Foster, A. Gausen, M. Grandury, S. Han, V. Hofmann, L. Ibrahim, H. Kim, H. R. Kirk, F. Lin, G. K.-M. Liu, L. Luettgau, J. Magomere, J. Rystrom, A. Sotnikova, Y. Yang, Y. Zhao, A. Bibi, A. Bosselut, R. Clark, A. Cohan, J. N. Foerster, Y. Gal, S. A. Hale, I. D. Raji, C. Summerfield, P. Torr, C. Ududec, L. Rocher, and A. Mahdi. Measuring what matters: Construct validity in large language model benchmarks. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025.

- [6] Y. Bengio, J. Louradour, R. Collobert, and J. Weston. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning, ICML '09*, page 41–48, New York, NY, USA, 2009. Association for Computing Machinery.
- [7] S. Bhojanam and S. Mehta. Prompt genotyping: Quantifying the evaluation gap between synthetic benchmarks and real LLM performance. In *NeurIPS 2025 Workshop on Evaluating the Evolving LLM Lifecycle: Benchmarks, Emergent Abilities, and Scaling*, 2025.
- [8] S. L. Blodgett, S. Barocas, H. Daumé III, and H. Wallach. Language (technology) is power: A critical survey of “bias” in NLP. In D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online, July 2020. Association for Computational Linguistics.
- [9] S. L. Blodgett, G. Lopez, A. Olteanu, R. Sim, and H. Wallach. Stereotyping Norwegian salmon: An inventory of pitfalls in fairness benchmark datasets. In C. Zong, F. Xia, W. Li, and R. Navigli, editors, *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 1004–1015, Online, Aug. 2021. Association for Computational Linguistics.
- [10] S. R. Bowman and G. Dahl. What will it take to fix benchmarking in natural language understanding? In K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, and Y. Zhou, editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4843–4855, Online, June 2021. Association for Computational Linguistics.
- [11] D. T. Campbell and D. W. Fiske. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56(2):81–105, 1959.
- [12] W.-L. Chiang, L. Zheng, Y. Sheng, A. N. Angelopoulos, T. Li, D. Li, B. Zhu, H. Zhang, M. I. Jordan, J. E. Gonzalez, and I. Stoica. Chatbot arena: an open platform for evaluating llms by human preference. In *Proceedings of the 41st International Conference on Machine Learning, ICML'24*. JMLR.org, 2024.
- [13] Y. Y. Chiu, L. Jiang, B. Y. Lin, C. Y. Park, S. S. Li, S. Ravi, M. Bhatia, M. Antoniak, Y. Tsvetkov, V. Shwartz, and Y. Choi. CulturalBench: A robust, diverse and challenging benchmark for measuring LMs’ cultural knowledge through human-AI red-teaming. In W. Che, J. Nabende, E. Shutova, and M. T. Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 25663–25701, Vienna, Austria, July 2025. Association for Computational Linguistics.
- [14] L. L. Cook and M. J. Pitoniak, editors. *Educational Measurement*. Oxford University Press, 5 edition, 2025.
- [15] I. Covert, W. Ji, T. Hashimoto, and J. Zou. Scaling laws for the value of individual data points in machine learning. In *Proceedings of the 41st International Conference on Machine Learning, ICML'24*. JMLR.org, 2024.
- [16] L. J. Cronbach and P. E. Meehl. Construct validity in psychological tests. *Psychological Bulletin*, 52(4):281–302, 1955.
- [17] M. Dehghani, Y. Tay, A. A. Gritsenko, Z. Zhao, N. Houlsby, F. Diaz, D. Metzler, and O. Vinyals. The benchmark lottery, 2021.
- [18] İ. E. Deveci and D. Ataman. The ouroboros of benchmarking: Reasoning evaluation in an era of saturation. In *NeurIPS 2025 Workshop on Evaluating the Evolving LLM Lifecycle: Benchmarks, Emergent Abilities, and Scaling*, 2025.
- [19] M. Du, V. Manjunatha, R. Jain, R. Deshpande, F. Dernoncourt, J. Gu, T. Sun, and X. Hu. Towards interpreting and mitigating shortcut learning behavior of NLU models. In K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, and Y. Zhou, editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 915–929, Online, June 2021. Association for Computational Linguistics.

- [20] M. Eriksson, E. Purificato, A. Noroozian, J. Vinagre, G. Chaslot, E. Gomez, and D. Fernandez-Llorca. Can we trust ai benchmarks? an interdisciplinary review of current issues in ai evaluation. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 8(1):850–864, Oct. 2025.
- [21] European Parliament and Council of the European Union. Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union, L series, 2024.
- [22] FadillAmir. Benchmarking and standardization of evaluation protocols: A feedback-driven framework using LLM judges to gatekeep and iteratively improve synthetic benchmarks. In *NeurIPS 2025 Workshop on Evaluating the Evolving LLM Lifecycle: Benchmarks, Emergent Abilities, and Scaling*, 2025.
- [23] C. Fourrier, N. Habib, A. Lozovskaya, K. Szafer, and T. Wolf. Open llm leaderboard v2. https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard, 2024.
- [24] B. Gao, F. Song, Z. Yang, Z. Cai, Y. Miao, Q. Dong, L. Li, C. Ma, L. Chen, R. Xu, Z. Tang, B. Wang, D. Zan, S. Quan, G. Zhang, L. Sha, Y. Zhang, X. Ren, T. Liu, and B. Chang. Omni-MATH: A universal olympiad level mathematic benchmark for large language models. In *The Thirteenth International Conference on Learning Representations*, 2025.
- [25] S. Gehman, S. Gururangan, M. Sap, Y. Choi, and N. A. Smith. RealToxicityPrompts: Evaluating neural toxic degeneration in language models. In T. Cohn, Y. He, and Y. Liu, editors, *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3356–3369, Online, Nov. 2020. Association for Computational Linguistics.
- [26] S. Golchin and M. Surdeanu. Time travel in LLMs: Tracing data contamination in large language models. In *The Twelfth International Conference on Learning Representations*, 2024.
- [27] H. Gulliksen. *Theory of Mental Tests*. Wiley Publications in Psychology. John Wiley & Sons, Hoboken, NJ, 1950.
- [28] R. K. Hambleton and R. W. Jones. Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice*, 12(3):38–47, 1993.
- [29] R. K. Hambleton, H. Swaminathan, and H. J. Rogers. *Fundamentals of Item Response Theory*, volume 2 of *Measurement Methods for the Social Sciences*. Sage Publications, Thousand Oaks, CA, 1991.
- [30] A. Hardy, A. Reuel, K. Jafari Meimandi, L. Soder, A. Griffith, D. M. Asmar, S. Koyejo, M. S. Bernstein, and M. J. Kochenderfer. More than marketing? on the information value of ai benchmarks for practitioners. In *Proceedings of the 30th International Conference on Intelligent User Interfaces, IUI ’25*, page 1032–1047, New York, NY, USA, 2025. Association for Computing Machinery.
- [31] D. Heineman, V. Hofmann, I. Magnusson, Y. Gu, N. A. Smith, H. Hajishirzi, K. Lo, and J. Dodge. Signal and noise: A framework for reducing uncertainty in language model evaluation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.
- [32] D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, and J. Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021.
- [33] S. Henrysson. Correction of item-total correlations in item analysis. *Psychometrika*, 28(2):211–218, 1963.
- [34] A. Jacovi, A. Caciularu, O. Goldman, and Y. Goldberg. Stop uploading test data in plain text: Practical strategies for mitigating data contamination by evaluation benchmarks. In H. Bouamor, J. Pino, and K. Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5075–5084, Singapore, Dec. 2023. Association for Computational Linguistics.

- [35] H. Jiang, X. Yi, Z. Wei, Z. Xiao, S. Wang, and X. Xie. Raising the bar: Investigating the values of large language models via generative evolving testing. In *Forty-second International Conference on Machine Learning*, 2025.
- [36] X. Jiang, D. Chang, and X. Xu. Time waits for no benchmark: Exploring the temporal misalignment between static benchmarks, modern LLMs, and the real world. In *NeurIPS 2025 Workshop on Evaluating the Evolving LLM Lifecycle: Benchmarks, Emergent Abilities, and Scaling*, 2025.
- [37] G. Kamradt. There are 4 stages in a benchmark lifecycle. X (formerly Twitter) post, Nov 2025. Accessed: 2026-01-16.
- [38] D. Kiela, M. Bartolo, Y. Nie, D. Kaushik, A. Geiger, Z. Wu, B. Vidgen, G. Prasad, A. Singh, P. Ringshia, Z. Ma, T. Thrush, S. Riedel, Z. Waseem, P. Stenetorp, R. Jia, M. Bansal, C. Potts, and A. Williams. Dynabench: Rethinking benchmarking in NLP. In K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, and Y. Zhou, editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4110–4124, Online, June 2021. Association for Computational Linguistics.
- [39] E. Kim, S. Li, S. Khalil, and H. J. Shin. STAIR-AIG: Optimizing the automated item generation process through human-AI collaboration for critical thinking assessment. In E. Kochmar, B. Alhafni, M. Bexte, J. Burstein, A. Horbach, R. Laarmann-Quante, A. Tack, V. Yaneva, and Z. Yuan, editors, *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 920–930, Vienna, Austria, July 2025. Association for Computational Linguistics.
- [40] R. Le Bras, S. Swayamdipta, C. Bhagavatula, R. Zellers, M. E. Peters, A. Sabharwal, and Y. Choi. Adversarial filters of dataset biases. In *Proceedings of the 37th International Conference on Machine Learning*, ICML’20. JMLR.org, 2020.
- [41] M. Li, H. Jiao, T. Zhou, N. Zhang, S. Peters, and R. W. Lissitz. Item difficulty modeling using fine-tuned small and large language models. In J. Wilson, C. Ormerod, and M. Beiting Parrish, editors, *Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con): Coordinated Session Papers*, pages 48–55, Wyndham Grand Pittsburgh, Downtown, Pittsburgh, Pennsylvania, United States, Oct. 2025. National Council on Measurement in Education (NCME).
- [42] T. Li, W.-L. Chiang, E. Frick, L. Dunlap, T. Wu, B. Zhu, J. E. Gonzalez, and I. Stoica. From crowdsourced data to high-quality benchmarks: Arena-hard and benchbuilder pipeline. In *Forty-second International Conference on Machine Learning*, 2025.
- [43] P. Liang, R. Bommasani, T. Lee, D. Tsipras, D. Soylu, M. Yasunaga, Y. Zhang, D. Narayanan, Y. Wu, A. Kumar, B. Newman, B. Yuan, B. Yan, C. Zhang, C. Cosgrove, C. D. Manning, C. Re, D. Acosta-Navas, D. A. Hudson, E. Zelikman, E. Durmus, F. Ladhak, F. Rong, H. Ren, H. Yao, J. WANG, K. Santhanam, L. Orr, L. Zheng, M. Yuksekgonul, M. Suzgun, N. Kim, N. Guha, N. S. Chatterji, O. Khattab, P. Henderson, Q. Huang, R. A. Chi, S. M. Xie, S. Santurkar, S. Ganguli, T. Hashimoto, T. Icard, T. Zhang, V. Chaudhary, W. Wang, X. Li, Y. Mai, Y. Zhang, and Y. Koreeda. Holistic evaluation of language models. *Transactions on Machine Learning Research*, 2023. Featured Certification, Expert Certification, Outstanding Certification.
- [44] Q. V. Liao and Z. Xiao. Rethinking model evaluation as narrowing the socio-technical gap, 2025.
- [45] B. Y. Lin, Y. Deng, K. Chandu, F. Brahman, A. Ravichander, V. Pyatkin, N. Dziri, R. L. Bras, and Y. Choi. Wildbench: Benchmarking llms with challenging tasks from real users in the wild, 2024.
- [46] F. Lin, S. Xie, Y. Dai, W. Yao, T. Lang, and Y. Zhang. IDGen: Item discrimination induced prompt generation for LLM evaluation. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

- [47] J. Liu, Y. Nam, X. Cui, and S. Swayamdipta. Evaluation under imperfect benchmarks and ratings: A case study in text simplification. In *NeurIPS 2025 Workshop on Evaluating the Evolving LLM Lifecycle: Benchmarks, Emergent Abilities, and Scaling*, 2025.
- [48] Y. L. Liu, S. L. Blodgett, J. Cheung, Q. V. Liao, A. Olteanu, and Z. Xiao. ECBD: Evidence-centered benchmark design for NLP. In L.-W. Ku, A. Martins, and V. Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16349–16365, Bangkok, Thailand, Aug. 2024. Association for Computational Linguistics.
- [49] F. M. Lord. *Applications of Item Response Theory to Practical Testing Problems*. Lawrence Erlbaum Associates, Hillsdale, NJ, 1980.
- [50] F. M. Lord and M. R. Novick. *Statistical Theories of Mental Test Scores*. Addison-Wesley, Reading, MA, 1968.
- [51] I. Magnusson, N. Tai, B. Bogin, D. Heineman, J. D. Hwang, L. Soldaini, A. Bhagia, J. Liu, D. Groeneveld, O. Tafjord, N. A. Smith, P. W. Koh, and J. Dodge. Datadecide: How to predict best pretraining data with small experiments. In *Forty-second International Conference on Machine Learning*, 2025.
- [52] N. Muennighoff, N. Tazi, L. Magne, and N. Reimers. MTEB: Massive text embedding benchmark. In A. Vlachos and I. Augenstein, editors, *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2014–2037, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics.
- [53] O. Nahum, N. Calderon, O. Keller, I. Szpektor, and R. Reichart. Are LLMs better than reported? detecting label errors and mitigating their effect on model performance. In C. Christodoulopoulos, T. Chakraborty, C. Rose, and V. Peng, editors, *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 26782–26809, Suzhou, China, Nov. 2025. Association for Computational Linguistics.
- [54] Y. Nie, A. Williams, E. Dinan, M. Bansal, J. Weston, and D. Kiela. Adversarial NLI: A new benchmark for natural language understanding. In D. Jurafsky, J. Chai, N. Schlueter, and J. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901, Online, July 2020. Association for Computational Linguistics.
- [55] S. Ott, A. Barbosa-Silva, K. Blagec, J. Brauner, and M. Samwald. Mapping global dynamics of benchmark creation and saturation in artificial intelligence. *Nature Communications*, 13(1):6793, 2022.
- [56] Y. Perlitz, A. Gera, O. Arviv, A. Yehudai, E. Bandel, E. Shnarch, M. Shmueli-Scheuer, and L. Choshen. Benchmark agreement testing done right: A guide for LLM benchmark evaluation. In *NeurIPS 2025 Workshop on Evaluating the Evolving LLM Lifecycle: Benchmarks, Emergent Abilities, and Scaling*, 2025.
- [57] I. D. Raji, A. Smart, R. N. White, M. Mitchell, T. Gebru, B. Hutchinson, J. Smith-Loud, D. Theron, and P. Barnes. Closing the ai accountability gap: defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, FAT* ’20*, page 33–44, New York, NY, USA, 2020. Association for Computing Machinery.
- [58] M. D. Reckase. 18 multidimensional item response theory. *Handbook of statistics*, 26:607–642, 2006.
- [59] D. Rein, B. L. Hou, A. C. Stickland, J. Petty, R. Y. Pang, J. Dirani, J. Michael, and S. R. Bowman. GPQA: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*, 2024.
- [60] S. Sabour, S. Liu, Z. Zhang, J. Liu, J. Zhou, A. Sunaryo, T. Lee, R. Mihalcea, and M. Huang. EmoBench: Evaluating the emotional intelligence of large language models. In L.-W. Ku, A. Martins, and V. Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association*

- for *Computational Linguistics (Volume 1: Long Papers)*, pages 5986–6004, Bangkok, Thailand, Aug. 2024. Association for Computational Linguistics.
- [61] O. E. Salaudeen, A. Reuel, A. M. Ahmed, S. Bedi, Z. Robertson, S. Sundar, B. W. Domingue, A. Wang, and S. Koyejo. Measurement to meaning: A validity-centered framework for AI evaluation. In *NeurIPS 2025 Workshop on Evaluating the Evolving LLM Lifecycle: Benchmarks, Emergent Abilities, and Scaling*, 2025.
 - [62] D. Sculley, W. Cukierski, P. Culliton, S. Dane, M. M. Demkin, R. Holbrook, A. Howard, P. T. Mooney, W. Reade, M. Risdal, and N. Keating. Position: AI competitions provide the gold standard for empirical rigor in genAI evaluation. In *Forty-second International Conference on Machine Learning Position Paper Track*, 2025.
 - [63] H. S. Son. Validity evaluation for the data used for artificial intelligence system. In Y. Bi, R. Bhatia, and S. Kapoor, editors, *Intelligent Systems and Applications*, pages 362–369, Cham, 2020. Springer International Publishing.
 - [64] S. Swayamdipta, R. Schwartz, N. Lourie, Y. Wang, H. Hajishirzi, N. A. Smith, and Y. Choi. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In B. Webber, T. Cohn, Y. He, and Y. Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9275–9293, Online, Nov. 2020. Association for Computational Linguistics.
 - [65] S. T. Truong, Y. Tu, P. Liang, B. Li, and S. Koyejo. Reliable and efficient amortized model-based evaluation. In *Forty-second International Conference on Machine Learning*, 2025.
 - [66] M. Udell, C. Horn, R. Zadeh, S. Boyd, et al. Generalized low rank models. *Foundations and Trends® in Machine Learning*, 9(1):1–118, 2016.
 - [67] K. L. Wagstaff. Machine learning that matters. In *Proceedings of the 29th International Conference on International Conference on Machine Learning*, ICML’12, page 1851–1856, Madison, WI, USA, 2012. Omnipress.
 - [68] A. Wang, Y. Pruksachatkun, N. Nangia, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman. *SuperGLUE: a stickier benchmark for general-purpose language understanding systems*. Curran Associates Inc., Red Hook, NY, USA, 2019.
 - [69] Y. Wang, X. Ma, G. Zhang, Y. Ni, A. Chandra, S. Guo, W. Ren, A. Arulraj, X. He, Z. Jiang, T. Li, M. Ku, K. Wang, A. Zhuang, R. Fan, X. Yue, and W. Chen. MMLU-pro: A more robust and challenging multi-task language understanding benchmark. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024.
 - [70] D. Westen and R. Rosenthal. Quantifying construct validity: Two simple measures. *Journal of Personality and Social Psychology*, 84(3):608–618, 2003.
 - [71] J. Weston, A. Bordes, S. Chopra, A. M. Rush, B. Van Merriënboer, A. Joulin, and T. Mikolov. Towards ai-complete question answering: A set of prerequisite toy tasks. *arXiv preprint arXiv:1502.05698*, 2015.
 - [72] C. White, S. Dooley, M. Roberts, A. Pal, B. Feuer, S. Jain, R. Shwartz-Ziv, N. Jain, K. Sai-fullah, S. Dey, Shubh-Agrawal, S. S. Sandha, S. V. Naidu, C. Hegde, Y. LeCun, T. Goldstein, W. Neiswanger, and M. Goldblum. Livebench: A challenging, contamination-limited LLM benchmark. In *The Thirteenth International Conference on Learning Representations*, 2025.
 - [73] S. Wu, H. Bao, S. Li, A. Holtzman, and J. A. Evans. Mapping overlaps in benchmarks through perplexity in the wild, 2025.
 - [74] Z. Xiao, S. Zhang, V. Lai, and Q. Liao. Evaluating evaluation metrics: A framework for analyzing NLG evaluation metrics using measurement theory. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.
 - [75] S. M. Xie, H. Pham, X. Dong, N. Du, H. Liu, Y. Lu, P. Liang, Q. V. Le, T. Ma, and A. W. Yu. Doremi: Optimizing data mixtures speeds up language model pretraining. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

- [76] C. Xu, N. Yan, S. Guan, C. Jin, Y. Mei, Y. Guo, and T. Kechadi. DCR: Quantifying data contamination in LLMs evaluation. In C. Christodoulopoulos, T. Chakraborty, C. Rose, and V. Peng, editors, *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 23002–23020, Suzhou, China, Nov. 2025. Association for Computational Linguistics.
- [77] X. Xu, Z. Wu, R. Qiao, A. Verma, Y. Shu, J. Wang, X. Niu, Z. He, J. Chen, Z. Zhou, G. K. R. Lau, H. Dao, L. Agussurja, R. H. L. Sim, X. Lin, W. Hu, Z. Dai, P. W. Koh, and B. K. H. Low. Position paper: Data-centric AI in the age of large language models. In Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11895–11913, Miami, Florida, USA, Nov. 2024. Association for Computational Linguistics.
- [78] Z. Xu, S. Xie, Q. Lv, S. Xiao, L. Song, S. Wenjuan, and F. Lin. Diagnosing failures in large language models’ answers: Integrating error attribution into evaluation framework. In W. Che, J. Nabende, E. Shutova, and M. T. Pilehvar, editors, *Findings of the Association for Computational Linguistics: ACL 2025*, pages 21148–21165, Vienna, Austria, July 2025. Association for Computational Linguistics.
- [79] J. Yao, P. Jin, K. Bao, Q. Yu, K. Bhardwaj, C. Su, J. Wang, Y. ZHU, S. Devare, D. Mosk-Aoyama, Z. Dong, V. K. Srinivasan, Y. Zhang, O. Kuchaiev, J. Jiao, and B. Zhu. The measure of all measures: Quantifying LLM benchmark quality. In *NeurIPS 2025 Workshop on Evaluating the Evolving LLM Lifecycle: Benchmarks, Emergent Abilities, and Scaling*, 2025.
- [80] A. K. Zhang, K. Klyman, Y. Mai, Y. Levine, Y. Zhang, R. Bommasani, and P. Liang. Position: Language model developers should report train-test overlap. In *Forty-second International Conference on Machine Learning Position Paper Track*, 2025.
- [81] H. Zhang, Y. Chen, and X. Li. A note on exploratory item factor analysis by singular value decomposition. *Psychometrika*, 85(2):358–372, 2020.

A Limitations

The limitations discussed here pertain to how we present and substantiate our position in this paper, rather than to the value or potential of item-level benchmark data itself.

First, the empirical illustrations in Sec. 5 draw primarily on classical test theory and item factor analysis, representing only a subset of the analytical possibilities enabled by item-level data. Applying psychometric methods to AI benchmarking data also requires caution, as their underlying assumptions may not hold: shared training corpora and model lineages can introduce dependencies among models, while benchmark multidimensionality can complicate the interpretation of rest-total discrimination. Although our sensitivity analyses (App. D) support the stability of CTT estimates, they do not establish model independence. The GPT-assisted factor interpretations in Sec. 5.3 also remain exploratory. Broader methodological coverage, including item response theory, differential item functioning analysis, and diagnostic classification models, would further demonstrate the range of research practices that item-level data can support.

Second, although Secs. 2, 4, & 7 discuss the tension between open data access and data contamination risk, and App. B elaborates potential safeguards and governance mechanisms, their effectiveness and long-term sustainability remain to be established in practice. Implementing and evaluating access controls, staged releases, and community reporting processes will be necessary to assess how well these mechanisms balance transparency, reproducibility, privacy, and benchmark integrity.

B Data Accessibility and Governance

Standardized item-level releases must balance transparency and reproducibility with privacy, licensing restrictions, and benchmark integrity. Our position advocates maximizing access whenever legally and ethically feasible, while allowing release mechanisms to accommodate benchmark-specific constraints. Here, we elaborate how standardized releases can support contamination prevention and auditing, and outline complementary safeguards and community governance mechanisms.

Supporting Contamination Prevention and Auditing. Information asymmetry about which benchmark items exist, how they have been used, and what uses are permitted complicates contamination prevention and detection. Standardized item content and provenance records can help developers identify known benchmark material and filter it from training data. Item-level responses and test conditions further support independent scrutiny of decontamination claims by enabling researchers to investigate item-level contamination signals. Such signals warrant further investigation rather than conclusive evidence of contamination or its absence. Making such evidence accessible also lowers the cost of developing external auditing methods and investigating undisclosed benchmark use.

Controlled Release Mechanisms.

- *Hold-out designs:* A portion of benchmark items can be withheld from public release and reserved for independent, periodic re-evaluation. Comparisons between performance on public and held-out items can help investigate potential overfitting or contamination, provided that differences in item difficulty and content are taken into account. This design preserves a protected evaluation resource while allowing the released portion to support item-level analysis and methodological development.
- *Verified access:* For content that cannot be released publicly but can be shared under appropriate conditions, access can be restricted to researchers who register and agree to specified usage terms, with additional identity or affiliation verification where warranted. Such arrangements can specify permitted uses, redistribution restrictions, and reporting obligations. Access remains subject to the underlying privacy and licensing constraints; registration alone does not remove these restrictions.
- *Staged release:* Items can be released after a predefined embargo period, such as one evaluation cycle, to balance timely community access with the value of temporarily protected test content. The release schedule and access conditions should be documented in advance so that contributors and users can track when and how the data will become available. Content that remains subject to privacy or licensing restrictions would continue to require appropriate access controls after the embargo ends.

Community Governance and Long-Term Sustainability. We envision a coordinated community process for maintaining item-level data standards and promoting transparency. A shared communication channel would allow contributors and users to report known benchmark usage, staged-release status, and suspected contamination, supported by community norms for disclosure and auditing. Reports should distinguish suspected from substantiated contamination and allow supporting evidence and corrections to be recorded, reducing information asymmetry without claiming complete visibility into undisclosed data use. Regarding the schema governance, a lead organizing team would coordinate schema updates based on community feedback, with version control and backward-compatibility requirements to limit fragmentation. Harness-specific converters, general conversion tools, and schema validation can reduce submission errors, while additional quality control is needed to ensure content accuracy, provenance, and metadata consistency. Long-term sustainability also requires resources for storage, access, version management, corrections, and contributor support. Drawing on practices from established repositories, we are exploring federated governance under common standards and central coordination. Shared infrastructure can distribute responsibilities and reduce duplicated effort, but sustained funding and maintenance remain necessary.

C OpenEval Schema

Fig. 7 displays the data schema of OpenEval, including required and optional fields, data type constraints, and field descriptions. The `response`, `model`, and `score` objects folded in Fig. 7 are detailed in Figs. 8, 9, and 10, respectively.

```
{
  "item_id": "[str | auto] UNIQUE IDENTIFIER FOR THE ITEM",
  "item_metadata": {
    "ingestion_time": "[str | auto] TIMESTAMP OF WHEN THE ITEM WAS INGESTED (ISO 8601 FORMAT)",
    "contributor": {
      "name": "[str | optional] NAME OF THE CONTRIBUTOR",
      "email": "[str | optional] EMAIL ADDRESS OF THE CONTRIBUTOR",
      "affiliation": "[str | optional] AFFILIATION OF THE CONTRIBUTOR"
    },
    "source": {
      "benchmark_name": "[str | non-empty] NAME OF THE BENCHMARK OR DATASET",
      "benchmark_version": "[str | required] VERSION OF THE BENCHMARK OR DATASET",
      "paper_url": "[str | optional] URL TO THE BENCHMARK PAPER",
      "dataset_url": "[str | optional] URL TO THE BENCHMARK DATASET",
      "benchmark_tags": "[list[str] | optional] TAGS OR KEYWORDS ASSOCIATED WITH THE BENCHMARK OR DATASET"
    }
  },
  "item_content": {
    "input": "[list[str,dict] | non-empty] CONTENT OF THE ITEM, SUCH AS A QUESTION, A DIALOGUE, OR MULTIPLE CHOICE OPTIONS",
    "references": "[list[str,dict] | required] REFERENCES FOR THE EVALUATION METRIC"
  },
  "responses": [...],
  "schema_version": "[str | auto] OPENEVAL SCHEMA VERSION"
}
```

Figure 7: OPENEVAL data schema (item). The value of each field specifies the data type, presence requirement, and a short explanation of the field. Specifically, `auto` means the field is auto-generated, `optional` means the field can be absent from the data entry, `required` means the data contributors are required to specify this field, and `non-empty` means the field cannot be left blank.

```

{
  "response_id": "[str | auto] UNIQUE IDENTIFIER FOR THE RESPONSE",
  "model": {...
},
  "item_adaptation": {
    "request_input": "[list[str,dict] | non-empty] ACTUAL INPUT PROVIDED TO THE MODEL
    FOR THIS RESPONSE, SUCH AS THE ADAPTED QUESTION OR DIALOGUE",
    "demonstrations": "[list[str,dict] | optional] DEMONSTRATIONS OR EXAMPLES
    PROVIDED TO THE MODEL FOR THIS RESPONSE",
    "external_resources": [
      {
        "type": "[str | required] TYPE OF THE EXTERNAL RESOURCE",
        "content": "[any | required] AVAILABLE EXTERNAL RESOURCES FROM THE TEST
        ENVIRONMENT"
      }
    ]
  },
  "response_content": "[list[str,dict] | non-empty] CONTENT OF THE RESPONSE, SUCH AS
  THE MODEL'S ANSWER OR OUTPUT",
  "scores": [...
]
}

```

Figure 8: OPENEVAL data schema (response). Each response object is an element of the responses field in Fig. 7.

```

{
  "name": "[str | non-empty] NAME OF THE RESPONDENT MODEL",
  "size": "[str | optional] SIZE OF THE RESPONDENT MODEL",
  "model_adaptation": {
    "system_instruction": "[str | required] SYSTEM INSTRUCTION OR PROMPT USED TO
    ADAPT THE ITEM FOR THIS RESPONSE",
    "generation_parameters": {
      "temperature": "[float | required] TEMPERATURE USED FOR GENERATION",
      "do_sample": "[bool | required] WHETHER OR NOT TO USE SAMPLING, FALSE MEANS
      GREEDY DECODING",
      "top_k": "[int | required] TOP-K VALUE USED FOR GENERATION",
      "top_p": "[float | required] TOP-P VALUE USED FOR GENERATION",
      "max_tokens": "[int | required] MAXIMUM NUMBER OF TOKENS GENERATED"
    },
    "tools": [
      {
        "type": "[str | required] TYPE OF THE TOOL, SUCH AS A SEARCH ENGINE",
        "content": "[any | required] TOOL AVAILABLE TO THE MODEL"
      }
    ]
  }
}

```

Figure 9: OPENEVAL data schema (model). The model object here corresponds to the model field in Fig. 8.

```

{
  "metric": {
    "name": "[str | non-empty] NAME OF THE EVALUATION METRIC",
    "models": "[list[str] | required] MODELS OR ALGORITHMS USED TO COMPUTE THE EVALUATION METRIC",
    "extra_artifacts": [
      {
        "type": "[str | required] TYPE OF THE ARTIFACT",
        "content": "[any | required] ADDITIONAL ARTIFACT FOR SCORING THE RESPONSE"
      }
    ]
  },
  "value": "[int,float,bool | non-empty] NUMERIC VALUE OF THE EVALUATION METRIC"
}

```

Figure 10: OPENEVAL data schema (score). Each score object is an element of the `scores` field in Fig. 8.

D Sensitivity Analyses of CTT Estimates

To assess the sensitivity of the item difficulty and discrimination estimates to the evaluated model population, we expanded the original sample of 72 models to 122 models evaluated on the same 1,000 MMLU-PRO items. The original sample spans 14 developers and 20 model families, while the expanded sample covers 17 developers. We conducted two complementary analyses: bootstrap resampling and leave-one-developer-out analysis.

Bootstrap Resampling. We generated 1,000 bootstrap samples, each by drawing 72 models with replacement from the expanded pool of 122 models, and re-estimated item difficulty and discrimination in each sample. Table 3 reports the mean item-level standard error and 90% confidence interval width, together with the mean correlation across items between the resampled estimates and those obtained from the original 72-model sample. The average correlations were 0.9916 for difficulty and 0.9732 for discrimination, indicating that the relative ordering of item estimates remained largely stable under the examined changes in model composition.

Leave-one-developer-out Analysis. To further examine whether the estimates were disproportionately influenced by any single developer, we excluded all models from each of the 17 developers in turn and re-estimated item difficulty and discrimination using the remaining models. For each exclusion, we computed the Pearson correlation across items between the resulting estimates and the full 122-model estimates, using items with defined estimates in both samples. As shown in Table 3, these correlations averaged 0.9988 for difficulty and 0.9973 for discrimination. Table 4 reports the results for the five largest developers; all correlations exceeded 0.988.

Table 3: Sensitivity of CTT estimates on 1,000 MMLU-PRO items. The correlation reference is the original 72-model sample for bootstrap resampling and the full 122-model sample for leave-one-developer-out analysis. For leave-one-developer-out analysis, SEs and interval widths are jackknife-style descriptive summaries.

Analysis	Estimate	Correlation	SE	90% CI Width
Bootstrap	Difficulty	0.9916	0.0523	0.1714
	Discrimination	0.9732	0.0691	0.2295
Leave-one-developer-out	Difficulty	0.9988	0.0718	0.2361
	Discrimination	0.9973	0.0646	0.2127

Regarding the jackknife-style approximation in the leave-one-developer-out analysis, we calculated the standard error for each item j as follows:

$$\widehat{\text{SE}}_j = \frac{D-1}{\sqrt{D}} \text{SD}_{d=1,\dots,D} \left(\hat{\theta}_{j,-d} \right), \quad (1)$$

Table 4: Stability of CTT estimates after excluding each of the five largest model developers. Correlations are computed across items against the full 122-model estimates; model counts indicate the number of models excluded.

Developer	Model Count	Full-Sample Correlation	
		Difficulty	Discrimination
Qwen	27	0.9951	0.9899
OpenAI	21	0.9936	0.9882
Google	19	0.9976	0.9924
Anthropic	10	0.9980	0.9952
Meta	8	0.9993	0.9976

where $\hat{\theta}_{j,-d}$ denotes the item’s difficulty or discrimination estimate after excluding developer d , $D = 17$, and SD is the sample standard deviation. The approximate 90% CI width was then computed as $2 \times 1.645 \times \widehat{SE}_j$, and both quantities were averaged across items. These quantities should be interpreted as descriptive sensitivity summaries rather than calibrated confidence intervals.

Difficulty-Stratified Discrimination Comparison. To examine whether the observed discrimination differences between MMLU and MMLU-PRO persist within similar ranges of observed item difficulty, we grouped items by their difficulty index p_i into five bins of width 0.2 and compared item discrimination within each bin using Mann–Whitney U tests. We excluded 22 zero-variance MMLU-PRO items for which discrimination was undefined.

As shown in Table 5, MMLU-PRO exhibits higher mean discrimination in all five bins. Among the first four bins, three show statistically significant differences ($p < 0.01$), while the difference in $[0.4, 0.6)$ is not significant ($p = 0.333$). The final bin contains only two MMLU items, limiting the interpretability of that comparison. These results suggest that the observed discrimination advantage is not solely attributable to differences in the overall distribution of observed item difficulty.

Table 5: Item discrimination within difficulty-index bins for MMLU and MMLU-PRO. A larger difficulty-index p_i indicates an easier item. The p -values are from within-bin Mann–Whitney U tests; 22 zero-variance MMLU-PRO items were excluded.

p_i Bin	Dataset	Bin Size	Mean Disc.	p -value
(0.0, 0.2)	MMLU	144	-0.0174	2.48×10^{-3}
	MMLU-PRO	86	0.0615	
[0.2, 0.4)	MMLU	172	0.1444	3.43×10^{-6}
	MMLU-PRO	86	0.2933	
[0.4, 0.6)	MMLU	175	0.4161	3.33×10^{-1}
	MMLU-PRO	152	0.4708	
[0.6, 0.8)	MMLU	74	0.4605	7.16×10^{-13}
	MMLU-PRO	281	0.5876	
[0.8, 1.0)	MMLU	2	0.3803	1.18×10^{-1}
	MMLU-PRO	373	0.4967	

Together, these analyses indicate that the CTT estimates, particularly their relative ordering across items, remain relatively stable and are not disproportionately driven by any single developer. These findings do not establish independence among models or isolate benchmark design effects from differences between the model samples evaluated on MMLU and MMLU-PRO and thus remain descriptive.

E Additional IFA Results

Table 6 and Fig. 11 report the results of the GLRM analysis conducted in Sec. 5.3 on MMLU. Fig. 12 shows item clusters from four benchmark datasets in HELM, revealed by K-means clustering over item factor loadings derived from GLRM. Fig. 13 illustrates that items within the same subject or dataset, despite sharing the same label, can emphasize different aspects when their maximum item factor loadings differ.

Table 6: Representative items with large absolute item factor loadings and possible constructs for each GLRM factor in MMLU. The top 100 representative items are interpreted and summarized into a candidate label by GPT-5, then manually revised.

Factor	Representative Item	Potential Sub-Construct
MMLU #1	Which one of the following statements best describes the algebraic representation of the fitted regression line?	Domain-specific canonical framework knowledge
MMLU #2	Based on the paper “SoK: SSL and HTTPS: Revisiting past challenges and evaluating certificates trust model enhancements”, which of the following statements are false?	Applied synthesis and case-based judgment
MMLU #3	Find the product of the given polynomials in the given polynomial ring. $f(x) = 4x - 5, g(x) = 2x^2 - 4x + 2$ in $\mathbb{Z}_8[x]$.	Formal, quantitative, multi-step modeling
MMLU #4	Which of the following statements is true concerning the population regression function (PRF) and sample regression function (SRF)?	Domain-specific recall and simple reasoning
MMLU #5	The () is categorized as an unknown segment of the Deep Web which has been purposely kept hidden and is inaccessible using standard web browsers.	Conceptual understanding and explanation

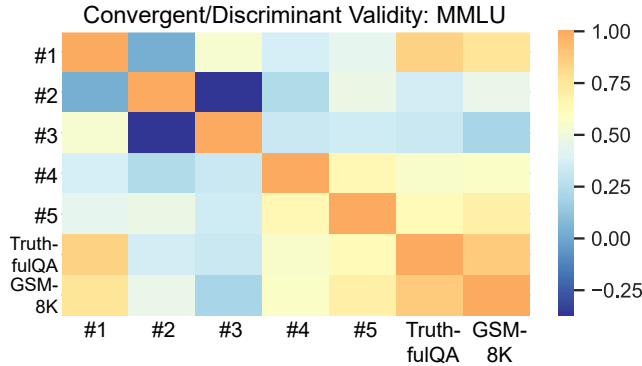


Figure 11: Convergent/discriminant evidence of the five sub-constructs (#1 - # 5) on MMLU.

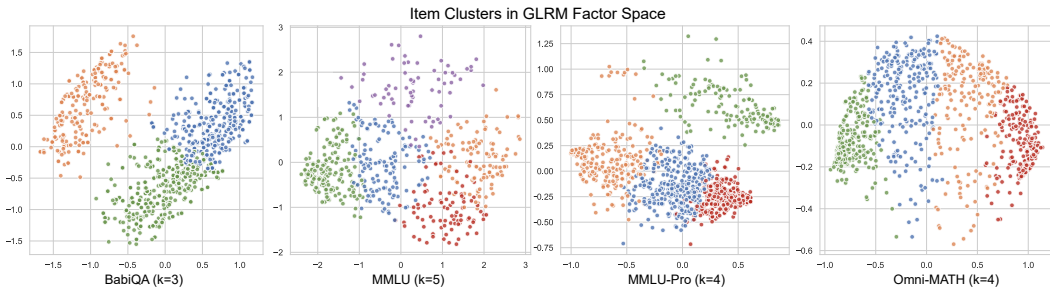


Figure 12: Clusters from four benchmark datasets in HELM revealed by K-means clustering over item factor loadings from GLRM.

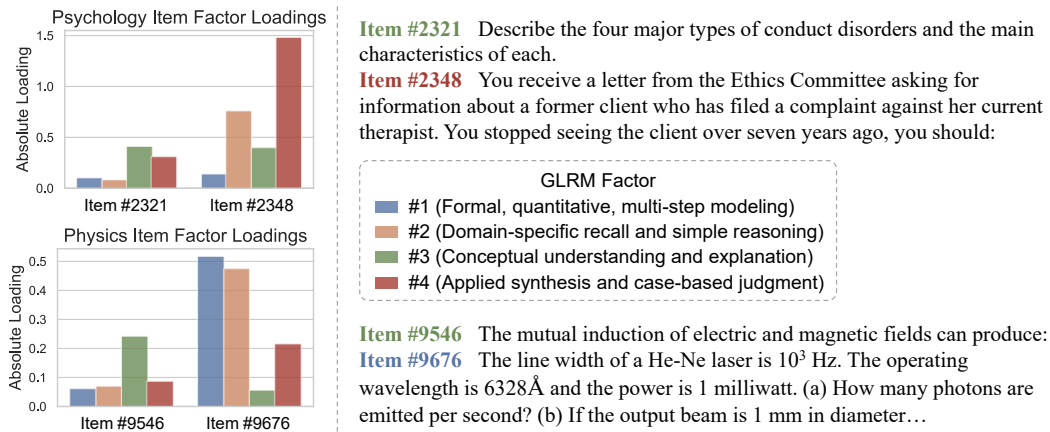


Figure 13: Example items with different maximum factor loadings within the same subject (psychology and physics) in MMLU-PRO.