Position: Science of AI Evaluation Requires Item-level Benchmark Data

Han Jiang [1] Susu Zhang [2] Xiaoyuan Yi [3] Xing Xie [3] Ziang Xiao* [1]

Abstract

AI evaluations have become the primary evidence for deploying generative AI systems across high-stakes domains. However, current evaluation paradigms often exhibit systemic validity failures. These issues, ranging from unjustified design choices to misaligned metrics, remain intractable without a principled framework for gathering validity evidence and conducting granular diagnostic analysis. In this position paper, we argue that **item-level AI benchmark data** is essential for establishing a rigorous science of AI evaluation. Item-level analysis enables fine-grained diagnostics and principled validation of benchmarks. We substantiate this position by dissecting current validity failures and revisiting evaluation paradigms across computer science and psychometrics. Through illustrative analyses of item properties and latent constructs, we demonstrate the unique insights afforded by item-level data. To catalyze community-wide adoption, we introduce `OpenEval`, a growing repository of item-level benchmark data designed supporting evidence-centered AI evaluation.

1. Introduction

Generative AI has drastically expanded automation, reshaping the social and economic fabric (Zhao et al., 2025). As these models move into high-stakes deployments, the risk of unpredictable behavior necessitates rigorous oversight and regulation. Consequently, AI evaluation, which is currently dominated by benchmarking practices (Eriksson et al., 2025), is essential for understanding model capabilities, informing AI policy, and guiding responsible deployment across domains.

Benchmarks have a long history as standardized tests for comparing systems, such as computing power (e.g., Dongarra et al., 1979) and component-level efficiency (e.g., UL Enterprise, 2002). Contemporary AI benchmarks primarily rely on curated datasets designed to operationalize specific capabilities, including reasoning (e.g., Valmeekam et al., 2022), tool use (e.g., Huang et al., 2024), social skills (e.g., Choi et al., 2023), and domain knowledge (e.g., Petroni et al., 2021). Characterized by capability-to-task breakdowns, task-specific metrics, and score aggregation (Liu et al., 2024), this paradigm provides a standardized, empirical basis for comparing AI system performance.

However, providing a defensible picture of model capabilities is increasingly difficult as the validity of AI benchmarks called into question (Raji et al., 2021). A central limitation is that critical design choices — including capability definitions, content curation, and metric selection — often lack transparency or formal justification (Liu et al., 2024). This opacity undermines the validity evidence (Blodgett et al., 2021; Liu et al., 2024) needed to support the interpretations of results, making it unclear whether benchmarks genuinely measure their intended constructs, despite explicit metadata or task descriptions (Akyürek et al., 2022). Furthermore, when the motivations for specific design paths are not clearly articulated, benchmarks become prone to redundancy (Polo et al., 2024; Ott et al., 2022).

Compounding these internal design limitations is the rapid evolution and opacity of modern AI systems, which continue to impose new pressures on benchmarking practices. The speed of AI development manifests as benchmark saturation (Ott et al., 2022), rapidly outdated content (Jiang et al., 2025b), and widespread data contamination (Xu et al., 2025a), rendering the resulting scores uninformative or misleading for deployment decisions (Golchin & Surdeanu, 2024). Further, the proliferation of model capabilities and task settings has created a widening socio-technical gap between technical solutions and real-world requirements (Wagstaff, 2012; Liao & Xiao, 2025). This lack of clear rationale also leads to redundancy, as new benchmarks often duplicate existing efforts without adding unique evaluative insights (Blodgett et al., 2020).

Crucially, many validity issues are not diagnosable from benchmark analyses with the model scores at the benchmark-level. Benchmark-level scores do not provide a sufficient

[1]Department of Computer Science, Johns Hopkins University, Baltimore, MD, USA [2]Department of Psychology, University of Illinois Urbana-Champaign, Urbana, IL, USA [3]Social Computing Group, Microsoft Research Asia, Beijing, China. Correspondence to: Ziang Xiao <ziang.xiao@jhu.edu>.

Preprint. February 24, 2026.

evidentiary basis for evaluating item quality, construct coverage, or potential confounders. Foundational questions—including whether items effectively differentiate model capabilities, how construct-irrelevant nuisance factors drive performance, or whether gains reflect genuine reasoning rather than artifacts—are inherently item-level inquiries. Without item-level response data, our field lacks the empirical evidence required to evaluate and curate effective benchmarks. We therefore argue that **broader access to and analysis of item-level AI benchmark data are essential for establishing a more scientifically grounded, evidence-centered approach to AI evaluation.**

Item-level assessment data have long been significant to test development and validation across education, psychology, and other social sciences disciplines concerned with measuring human abilities. Adapting these established practices to AI evaluation, through a community-wide push for broader data access, could transform fragmented results into cumulative empirical evidence. Leveraging the rich information from item content, score statistics, and per-item model responses, item-level analyses complement existing paradigms to enable more rigorous benchmark validation, granular diagnostics of benchmark characteristics, and more informed benchmark design.

In this position paper, we analyze validity failures in AI benchmarking and link them directly to the current neglect of item-level data in Sec. 2. By contrasting paradigms in computer science and psychometrics, we motivate an item-level shift (Sec. 3) and formalize our position (Sec. 4). In Sec. 5, we present illustrative item-level analyses on selected AI benchmark datasets to highlight the unique insights afforded by item-level data. Finally, we introduce a growing, large-scale item-level benchmark data repository, `OpenEval`, in Sec. 6 to encourage further community effort and action, and discuss broader opportunities beyond benchmarking enabled by item-level benchmark data in Sec. 7.

2. Validity Issues in AI Benchmarking

Despite their pivotal role in AI evaluation, current AI benchmarks face methodological and practical validity challenges that stem from a neglect of item-level data.

2.1. Validity Problems in Benchmark Design

The term *validity*, specifically *construct validity*, was formalized by Cronbach & Meehl (1955) in modern psychometrics and is directly applicable to AI evaluation (Xiao et al., 2023). A *construct* refers to a theoretical, unobservable attribute or trait (e.g., a perceived AI capability) that a test is intended to measure. Validity reflects *how well an assessment measures the intended underlying outcome* (Son, 2020) and is critical for ensuring benchmark quality.

As noted by Westen & Rosenthal (2003), construct validity is a central concept in psychology, particularly for informing the design of psychological measures. In AI evaluation, however, it has received limited attention, despite benchmark creation involving numerous design decisions, including target user selection, construct-to-task operationalization, and scoring metrics. These decisions are often oversimplified due to *a dominant focus on benchmark-level results*, leaving only standard or default settings (e.g., Liang et al., 2023; Wang et al., 2019) and insufficient evidence for validity justification (Liu et al., 2024).

These hidden ambiguities have resulted in a lack of a common language for communicating construct validity within the AI community. Although there are several studies exploring validity-relevant properties (e.g., Yao et al., 2025a; Wu et al., 2025), research explicitly assessing the construct validity of AI benchmarks remains limited. Existing analyses suggest that many benchmarks' construct validity has yet to improve and that detailed validity evidence should be extracted not only from benchmark-level information (e.g., Blodgett et al., 2021; Bean et al., 2025; Salaudeen et al., 2025), highlighting the need to examine item-level data.

2.2. Validity Degradation over the Benchmark Lifecycle

Once released, a benchmark inevitably experiences validity degradation over its lifecycle. Many popular static benchmarks are approaching their *sunset* stage, where results and conclusion derived from them are no longer informative or reliable (Dehghani et al., 2021; Kamradt, 2025).

First, as AI systems and real-world knowledge evolve, some benchmarks have gradually saturated, with most items becoming too easy to distinguish between current models (Ott et al., 2022; Deveci & Ataman, 2025). For instance, `RealToxicityPrompts` (Gehman et al., 2020) is one of the most widely used AI toxicity benchmarks. Although by 2023 it could no longer differentiate between multiple versions of GPT (Jiang et al., 2025a), it remained widely used in AI safety research beyond 2024. Other benchmarks requiring factual knowledge are more time-sensitive: once references become outdated, the corresponding test items risk yielding unreliable assessments (Jiang et al., 2025b).

Another problem is data contamination, which is subtler due to the lack of transparency of AI system development. As AI training and evaluation scale, many omnibus benchmarks now aggregate multiple previous benchmarks, particularly general-purpose suites such as `BigBench` (Ghazal et al., 2013), `SuperGLUE` (Wang et al., 2019), and `MMLU` (Hendrycks et al., 2021; Wang et al., 2024). This growing aggregation makes data provenance harder to trace, and even inadvertent test–train overlap

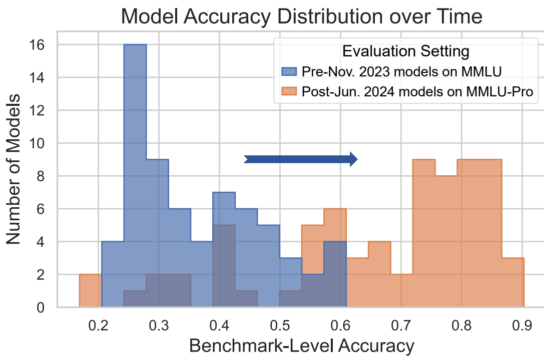


Figure 1. Benchmark-level accuracy distributions for 66 pre–Nov. 2023 models on `MMLU` and 72 post–Jun. 2024 models on `MMLU-Pro`. Results are from HELM-Classic and HELM-Capabilities.

can lead to unfair evaluations, which are nearly impossible to detect at benchmark level without explicit reporting by developers (Zhang et al., 2025).

These issues are difficult to diagnose and address without item-level details. As shown in Fig. 1, while both AI systems and benchmarks are advancing, the average accuracy distribution on `MMLU` continues to shift upward. Without further inspection of individual test items, it remains unknown whether the observed improvements reflect genuine capability gains, benchmark saturation, or data contamination, not to mention which items should be removed or retained. To mitigate these issues, some benchmarks are periodically updated manually (e.g., Lin et al., 2024a; White et al., 2025), but such approaches are time-consuming and costly. LLM-powered benchmarking (e.g., Kiela et al., 2021; FadillAmir, 2025) improves efficiency, yet the validity of the synthesized test items has been questioned by Bowman & Dahl (2021). More targeted, item-level methods are still lacking to compensate for the benchmark validity degradation.

2.3. Validity Challenges for Future Benchmarks

Moreover, emerging AI capabilities and evolving user needs continuously reshape the requirements for benchmark selection, analysis, and creation.

As AI systems exhibit increasingly diverse capabilities and users demand a widening range of use cases, benchmark selection becomes more complex, which requires item-level analyses that reveal underlying constructs and provide stronger evidentiary insight. However, even well-chosen benchmarks can yield invalid conclusions if the analysis is flawed. As shown by Bean et al. (2025), simply relying on aggregate benchmark scores, i.e., the final results may produce misleading conclusions about model capabilities. Beyond the validity degradation discussed in Sec. 2.2, the final results can be influenced by many confounders unralted to system ability (Xu et al., 2025b), such as erroneous items, spurious correlations, or unintended shortcuts exploited by models (e.g., Du et al., 2021; Nahum et al., 2025). The absence of item-level information hinders reasoning about what actually drives benchmark performance, leaving the validity of benchmark analyses unjustified.

Moreover, since current evaluation results are aggregated and not consistently released across leaderboards, they are neither comparable nor easily built upon, which ultimately limits the scalability of benchmark analysis. This further underscores the necessity of a consistent, item-level release of benchmark data.

A common scenario, however, is: what if no existing benchmarks adequately meet user needs? Traditional manual curation is too inefficient to be viable. Automatic benchmark generation techniques, such as adversarial filtering (e.g., Le Bras et al., 2020; Nie et al., 2020), item generation (e.g., Lin et al., 2024b; Kim et al., 2025), and difficulty adjustment (e.g., Truong et al., 2025), requires exploring inter-item dynamics. Meanwhile, theoretically grounded validation for these dynamic, adaptive benchmarks appears largely absent from the science of AI evaluation, as do principles for guiding the construction of such benchmarks. Lessons from existing benchmark data are largely ignored at the item level, which needs systematic analysis to inspire the creation of more future-proof evaluations.

3. Current Approaches to Benchmark Analysis

Benchmark analysis provides guidance on interpreting and extrapolating the outcomes of various benchmarking practices, which is indispensable for evaluation.

3.1. Benchmark Analysis in AI

In AI research, benchmark analysis has traditionally focused on aggregate comparisons that provide a high-level overview of model performance, such as leaderboards ranking many models on maintained benchmarks (e.g., Ni et al., 2024; Chiang et al., 2024); task-specific assessments analyzing system performance on selected tasks (e.g., Li et al., 2024; Yao et al., 2025b); or technical reports showcasing the capability sets of one or a series of models (e.g., OpenAI et al., 2024; Grattafiori et al., 2024). In these analyses, benchmarks serve as tools rather than research objects, and the validity of measurements depends on the employed benchmarks' scientific adequacy, which is rarely examined.

To fill this gap, some studies have begun to examine benchmark quality through qualitative meta-analyses by expert reviewers (e.g., Blodgett et al., 2021; Reuel et al., 2024; Bean et al., 2025) and to propose benchmark design goals. Several conceptual frameworks have been introduced to guide such qualitative analyses of benchmark validity (e.g., Liu et al., 2024; Salaudeen et al., 2025). Item-level inspection is implicitly involved in these analysis, and most findings suggest that the overall construct validity of many AI benchmarks remains concerning.

Further quantification of benchmark validity has prompted a growing number of studies to analyze item-level benchmark data, often focusing on test data quality or enhancing evaluation efficiency through item selection. For example, item difficulty estimation is common in such item-level analyses (e.g., Yao et al., 2025a; Li et al., 2025a), as more difficult items are intuitively better at distinguishing stronger models. The ability to separate different capability levels has also been treated as a benchmark characteristic (e.g., Heineman et al., 2025; Li et al., 2025b). Other properties includes item diversity (e.g., Muennighoff et al., 2023; Yao et al., 2025a), inter-benchmark agreement (e.g., Perlitz et al., 2025; Liu et al., 2025), and downstream performance predictability (e.g., Bhojanam & Mehta, 2025; Magnusson et al., 2025; Wu et al., 2025).

Despite significant progress, the lack of a large-scale, consistent item-level data release means these analyses are typically isolated, irreproducible and largely constrained to relatively monotonic scenarios, leaving substantial room for further exploration.

3.2. Item Analysis Practice in Psychometrics

Psychometrics treats item-level analysis as foundational to test construction, revision, and validation. According to the Standards for Educational and Psychological Testing (American Educational Research Association et al., 2014) — a consensus-based guideline articulating best practices for developing tests, assembling validity evidence, and documenting technical quality for responsible score interpretation — test validity is established through an iterative cycle that combines (i) qualitative item review and (ii) empirical item pretesting before operational use. During pretesting, items are field-tested on a representative sample, and the collected item-level response data are analyzed to verify that items function as intended and that the resulting score has adequate precision/reliability for its intended decisions.

The Standards further emphasize that when a single aggregate score is interpreted as a meaningful summary, developers should provide evidence about a test's internal structure, i.e., whether items behave coherently as indicators of the intended construct rather than reflecting multiple unrelated dimensions. Such evidence is commonly statistically examined with item factor analysis (Anderson et al., 1958; Reckase, 2006). For AI benchmarks, the direct analogue is to examine whether a benchmark behaves as a coherent measure of the intended capability versus reflecting construct-irrelevant variance (construct contamination). In benchmarks, potential sources of construct-irrelevant variance include formatting/annotation artifacts, answer-key idiosyncrasies, and leakage. A practical empirical flag is an item whose score is weakly, non-monotonically, or negatively related to performance on the rest of the benchmark.

Item-level evidence is also central to item screening and maintenance. Using field-test data and item statistics from classical test theory or item response theory (Cook & Pitoniak, 2025), developers screen for item difficulty coverage, diagnose misfit, and decide on item revision/replacement and form assembly to achieve adequate precision for the intended use. When items are reused over time, monitoring shifts in item statistics is also part of routine maintenance and can flag potential leakage. These analyses guide the detection of item misfit (i.e., empirical deviation from test/model assumptions) and inform item revision, replacement, and test assembly decisions, so that the resulting aggregate score attains adequate precision for the intended application.

From this perspective, reporting only aggregate LLM benchmark scores is analogous to reporting a single exam average without access to the item-level evidence needed to evaluate item functioning, dimensionality, and construct-irrelevant variance. Consequently, claims about benchmark's measurement quality are empirically underdetermined. More importantly, this perspective points toward a concrete opportunity for LLM research: psychometric test construction practices and statistical tools provide a well-understood and empirically-driven toolkit for analyzing benchmark items, model abilities, and their interaction. Adopting these practices to LLM benchmarks would enable measurement theory-grounded, evidence-based analyses of benchmark validity, item quality, and the structure of model competence. These analyses can directly inform benchmark revision and validation in ways that aggregate scores alone do not allow.

4. A Missing Foundation for the Science of AI Evaluation

Given the persistent validity issues in AI benchmarking and the limits in existing benchmark analysis approaches, **item-level benchmark data is a crucial missing piece in the science of AI evaluation**. It lies in every AI evaluation practice: detailed test conditions, the content of each item and model response, and per-response scores and statistics, forming a rich yet underexplored foundation for investigating the reliability and validity of AI benchmarking.

Item-level data enables fine-grained benchmark analysis and valid, insightful AI evaluation. By providing insights into AI systems beyond what is available from aggregate final results, item-level data facilitates the attribution of model behavior to measurable factors. This allows rigorous examination of how specific item properties are perceived and manifested in AI systems, thereby informing interpretability

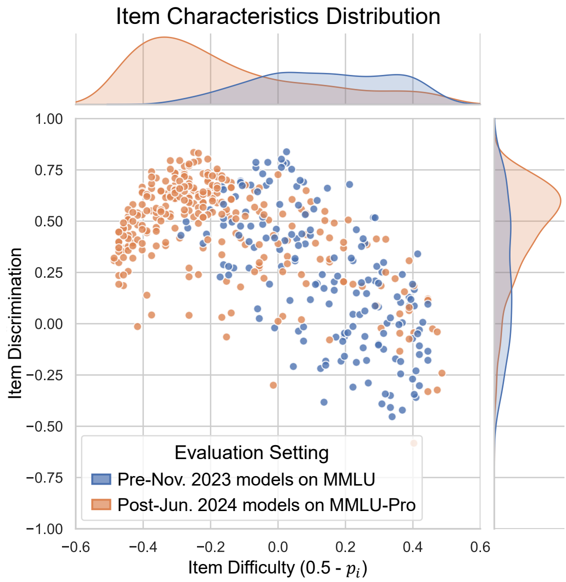


Figure 2. Item characteristic distributions for MMLU and MMLU-Pro. Item difficulty is transformed to $\mathrm{Diff}_i = 0.5 - p_i$, with higher values correspond to harder items.

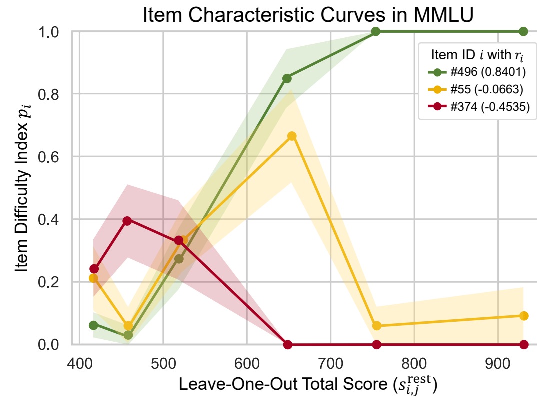


Figure 3. ICCs for three items in `MMLU`.

and alignment research. Furthermore, it is key to uncovering systematic strengths and weaknesses of AI systems across latent factors with improved construct validity, such as cognitive skills and reasoning types, which in turn supports reasoning about capability application and generalization.

Item-level data enables the principled design of AI benchmark. It is a prerequisite for developing measurement theories for AI systems, including principled notions of ability, reliability, and validity, analogous to those in psychometrics. These theories, combined with the rich evidence from item-level data, make it possible to scrutinize and enhance benchmark validity. With principled practices, such as quantifying item characteristics, identifying decisive benchmark factors, and modeling relationships between items and intended constructs, future benchmarks can clarify the dynamic mapping between AI system mechanisms and user interests.

Item-level data enables efficient maintenance and fair administration of AI benchmarks. With increased access to item-level data, the potential validity risks discussed in Sec. 2.2, including benchmark saturation and data contamination, can be timely diagnosed via item analyses using score statistics and detailed test cases. Such analyses can also effectively identifies and groups items of interest, providing valuable guidance for benchmark re-composition and updates. In addition, releasing item-level data enhances the transparency and scalability of AI benchmark analysis, allowing findings to be replicated, extended, and compared across studies, thereby prolonging benchmark lifecycles.

5. Empirical Illustrations

To illustrate the unique insights enabled by item-level benchmark data, we leverage item-level resources from HELM-Classic (v0.3.0) and HELM-Capabilities (Liang et al., 2023) to examine item characteristics and benchmark sub-constructs decomposition.

5.1. Item Characteristics from CTT

An item's statistical characteristics such as difficulty and discrimination are routinely examined in psychometric test development for quality assurance. We conduct a Classical Test Theory (CTT) analysis of item characteristics from (1) 66 pre-Nov. 2023 models on 567 items in `MMLU` (Hendrycks et al., 2021) and (2) 72 post–June 2024 models on 1,000 items in `MMLU-Pro` (Wang et al., 2024). `MMLU-Pro` (Wang et al., 2024) was introduced as an enhanced variant of `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 (p_i) can be estimated as (Gulliksen, 1950) 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 (i.e., sum score on all items except i) across all measured models (Henrysson, 1963). Higher r_i indicates that item i can well-differentiate models with strong vs. weak overall performance on the benchmark, whereas negative or near-zero r_i suggests a potentially problematic item.

Fig. 2 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

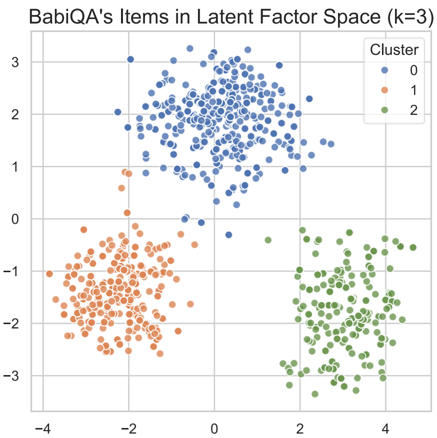


Figure 4. Item clusters on `BabiQA` based on factor loadings.

Table 1. Example `BabiQA` item and the numbers of items with different answer keys in each cluster in Fig. 4.

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

	Number of Items		
Item Answer	Cluster #0	Cluster #1	Cluster #2
Sheep	0	0	240
Mouse	225	0	0
Cat	221	3	0
Wolf	6	326	0

orange observations on the left indicates that a substantial proportion of `MMLU-Pro` items have very low difficulty. In other words, many items are no longer challenging for the 72 post-June 2024 models, suggesting fast benchmark saturation. (2) Compared to `MMLU`, item quality substantially improved on the `MMLU-Pro` with much fewer items with low or negative discrimination. This empirical observation is aligned with `MMLU-Pro` designers' goal (Wang et al., 2024) to build a more robust, less noisy benchmark. However, some `MMLU-Pro` items still showed poor discrimination. Although these items remained in `MMLU-Pro` after expert item review, their poor empirical discrimination merits additional scrutiny (e.g., for ambiguity, miskeying, or construct-irrelevant cues).

To further understand how item discrimination manifests, we plot the item characteristic curves (ICCs) of three items in `MMLU`. For each item i, all models are sorted into six equally-sized bins based on rest-total scores excluding item i. The expected item i score ($\%$ of max score) within each bin is then plotted in Fig. 3. Intuitively, a high discrimination item's expected score should increase with the total score on the remainder of the test, resulting in a monotonically increasing ICC. This was 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, 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 the design goals at a finer granularity, and supporting the maintenance and generation of benchmark data.

5.2. Examining Benchmark Internal Structure via IFA

As discussed in Sec. 2, examination of benchmark internal structure is essential for understanding what capabilities (or sub-dimensions) it actually measures. Here, we performed item factor analysis (IFA), using variants of conventional IFA for high-dimensional data based on singular value decomposition (SVD; Zhang et al., 2020) and Generalized Low Rank Models (GLRM; Udell et al., 2016). We report findings from `BabiQA` task 15 (basic deduction) and `MMLU-Pro`. Additional analysis results on `MMLU` are presented in App. A.

`BabiQA` task 15 (Westen & Rosenthal, 2003) aims to assess basic deductive reasoning via inheritance of properties. An example item is shown in Table 1. To aid interpretations, we performed K-means clustering to items' factor loadings on the top 3 factors obtained from SVD-based IFA, with items within each cluster measuring similar sub-dimensions. As shown in Fig. 4, 1,000 items in `BabiQA` formed three distinct clusters. A closer examination raised a construct validity flag: Table 1 shows that item clusters were explained by the answer key to the item, suggesting that different models' performance on `BabiQA` Benchmark were 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*. GLRM-based IFA yielded consistent findings.

For `MMLU-Pro`, we interpret the top 4 factors retained from GLRM-based IFA after varimax rotation. The top 100 items with the largest absolute loadings on each factor were sent to GPT-5 for interpretation. Table 2 presents a representative item for each factor and the sub-dimension interpretation. The four primary dimensions that best explained 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 (Wang et al., 2024) relative to `MMLU`. Indeed, items within the same subject domain (e.g., Psychology and Physics in Fig. 9) could differ substantially in loadings on the four factors.

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 (Campbell & Fiske, 1959), we correlate factor subscores (mean scores on the top-100-loading items) with scores on two external benchmarks: `GPQA` (Rein et al., 2024) (graduate-level biology, physics, and chemistry) and `Omni-MATH` (Gao et al., 2025) (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 benchmarks, whereas Factor #4 (applied synthesis and case-based judgment) aligns more with `GPQA`. Results in Fig. 5 are broadly consistent with these hypotheses. Further, factors #2 (domain-specific recall and simple reasoning) and #3 (conceptual understanding and explanation) showed weak 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.

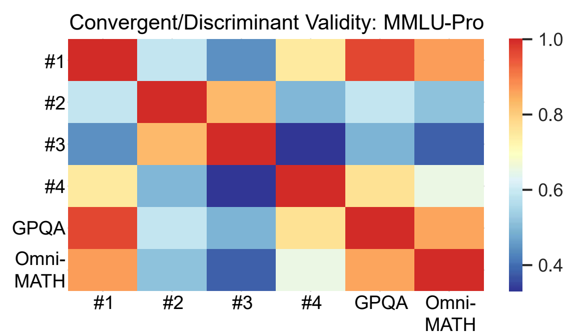


Figure 5. Convergent/discriminant evidence of the four sub-constructs (#1 - # 4) on `MMLU-Pro`.

6. OpenEval: Item-Level Benchmark Data Repository

To improve access to item-level data, we build a growing repository, `OpenEval`, designed to neatly organize existing benchmark items together with model responses, scores, and other associated information across different benchmark releases.

There have been many large-scale, high-quality AI benchmark repositories (e.g., Liang et al., 2023; Chiang et al., 2024), some of which release unstructured or semi-structured item-level details alongside the benchmark-level results. To facilitate item-level benchmark analysis and other forms of exploration, `OpenEval` features a ***scalable, item-centric schema*** in which each data entry represents a unique item occurrence. The schema is illustrated in Fig. 6.

Currently, we curated item-level resources from HELM and the OpenLLM Leaderboard v2 (Fourrier et al., 2024). We have been (1) collecting evaluation results to reduce the sparsity of the dataset-model matrix, and (2) incorporating external and interdisciplinary datasets. `OpenEval` now covers over **225k** items from **64** benchmark datasets, with the number of evaluated models per dataset ranging from dozens to thousands, resulting in more than **8M** item-level responses and scores in total.

We welcome additional data contributions from the community, envisioning a standardized, consistent, and traceable provenance for as many AI benchmark items as possible. We hope that `OpenEval` will grow into a foundation for the science of AI evaluation and responsible AI deployment.

7. Unlocking Possibilities beyond AI Evaluation

Beyond benchmark analysis and AI evaluation, item-level data has the potential to deliver both technical and practical benefits across disciplines and society.

Data-Centric Methodology for AI Development Access to item-level system performance enables investigation of

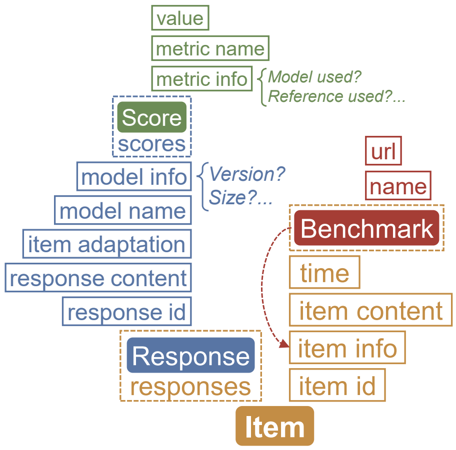


Figure 6. Schema for data entries in `OpenEval`.

AI learning trajectories across samples with varying properties, informing decisions about training data composition, training paradigms, and the choice of proxy tasks and evaluation metrics. Moreover, item-level data supports a shift toward data-driven research paradigms in many machine learning subfields (Xu et al., 2024), including statistical learning, generalization theory, and counterfactual ML, where aggregate-level analyses are inherently insufficient.

Domain-Grounded and Interdisciplinary Research Item-level benchmark data opens opportunities for domain-grounded research across multiple disciplines. First, it enables domain practitioners, such as linguists, clinicians, educators, and legal scholars, to interpret how AI systems perceive their domain, thereby informing better operationalization and deployment. Second, it supports the generation and validation of high-quality, domain-specific data for research in which AI is not the primary object, facilitating cross-disciplinary studies and methodological rigor.

Evidence-Based AI Governance and Auditing Policy and governance decisions increasingly rely on benchmark results to justify claims about model capability, risk, and readiness for deployment. Item-level data provides the necessary evidentiary grounding, allowing regulators and stakeholders to trace aggregate claims back to concrete data examples, error patterns, and coverage gaps. By making the referenced benchmark analyses transparent and reproducible, item-level data supports more accountable and informed decision-making in AI deployment and oversight.

Communicative and Public-Facing Uses Individual items make it possible to exemplify what AI systems can and cannot achieve in specific, relatable terms. This facilitates clearer communication with non-expert audiences, allowing public audit, and promotes more responsible narratives about AI capabilities, progress and societal risks, thereby supporting broader AI democratization.

8. Alternative Views

Item-level AI benchmark data has existed at a large scale for long, but for various reasons, attention to such data remains limited, as does its utilization. Here, we address some recent positions that may be partially opposed to ours.

We should stop uploading benchmark data in plain text to reduce the risk of data contamination. In addition to this position, Jacovi et al. (2023) propose three strategies to mitigate data contamination, including the use of encryption and closed APIs to limit access to benchmark data. In our view, such reduced accessibility would undermine the transparency of AI evaluation, widen the information gap between owners of closed-source resources and others, and ultimately exacerbate the unfairness caused by data contamination. In contrast, if we further increase this transparency by releasing comprehensive, item-level benchmark data for each study, the information gap could be reduced, and the disclosed test items and results could help more researchers detect data contamination.

Data contamination makes AI benchmarks unreliable; AI competitions should be the gold standard. Sculley et al. (2025) argue that the pitfall in benchmarking can be avoided by replacing static benchmarks with competitions that prioritize novelty and robustness over reproducibility. However, reproducibility is essential for the reusability, scalability, and comparability of studies, and ultimately for the efficiency of evaluation practices. It is more efficient to leverage the existing richness of item-level information for directly alleviation of data contamination, thereby improving the validity and reliability of AI benchmarking without requiring excessive resources. Moreover, only through item-level analyses can we obtain the insightful conclusions that support the novelty required for generating the robust test data for AI competitions.

Do not interpret AI benchmark performance as human-like intelligence; develop AI-specific evaluations instead. This position, proposed by Sühr et al. (2025), argues that AI benchmarking could be designed entirely from scratch to suit the context. Although it aligns with our position regarding the need for principled AI benchmark design, it underestimates the value of existing item-level benchmark data due to the underlying validity issues. Being human-centric is not a flaw, as the science of AI evaluation essentially serves human society; rather, the limitation lies in current benchmarking paradigms, which fail to make the intended constructs explicit. Therefore, item-level data provides a significant empirical foundation for justifying the interpretation of benchmark results and for understanding both what is being measured and what we aim to predict.

References

Akyürek, A. F., Kocyigit, M. Y., Paik, S., and Wijaya, D. T. Challenges in measuring bias via open-ended language generation. In Hardmeier, C., Basta, C., Costa-jussà, M. R., Stanovsky, G., and Gonen, H. (eds.), *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pp. 76–76, Seattle, Washington, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.gebnlp-1.9. URL https://aclanthology.org/2022.gebnlp-1.9/.

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.

Anderson, T. W., Anderson, T. W., Anderson, T. W., Anderson, T. W., and Mathématicien, E.-U. *An introduction to multivariate statistical analysis*, volume 2. Wiley New York, 1958.

Bean, A. M., Kearns, R. O., Romanou, A., Hafner, F. S., Mayne, H., Batzner, J., Foroutan, N., Schmitz, C., Korgul, K., Batra, H., Deb, O., Beharry, E., Emde, C., Foster, T., Gausen, A., Grandury, M., Han, S., Hofmann, V., Ibrahim, L., Kim, H., Kirk, H. R., Lin, F., Liu, G. K.-M., Luettgau, L., Magomere, J., Rystrøm, J., Sotnikova, A., Yang, Y., Zhao, Y., Bibi, A., Bosselut, A., Clark, R., Cohan, A., Foerster, J. N., Gal, Y., Hale, S. A., Raji, I. D., Summerfield, C., Torr, P., Ududec, C., Rocher, L., and Mahdi, A. 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. URL https://openreview.net/forum?id=mdA5lVvNcU.

Bhojanam, S. and Mehta, S. 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. URL https://openreview.net/forum?id=WCNAcIKREG.

Blodgett, S. L., Barocas, S., Daumé III, H., and Wallach, H. Language (technology) is power: A critical survey of "bias" in NLP. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J. (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5454–5476, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.485. URL https://aclanthology.org/2020.acl-main.485/.

Blodgett, S. L., Lopez, G., Olteanu, A., Sim, R., and Wallach, H. Stereotyping Norwegian salmon: An inventory of pitfalls in fairness benchmark datasets. In Zong, C., Xia, F., Li, W., and Navigli, R. (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. 1004–1015, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.81. URL https://aclanthology.org/2021.acl-long.81/.

Bowman, S. R. and Dahl, G. What will it take to fix benchmarking in natural language understanding? In Toutanova, K., Rumshisky, A., Zettlemoyer, L., Hakkani-Tur, D., Beltagy, I., Bethard, S., Cotterell, R., Chakraborty, T., and Zhou, Y. (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4843–4855, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.385. URL https://aclanthology.org/2021.naacl-main.385/.

Campbell, D. T. and Fiske, D. W. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56(2):81–105, 1959. ISSN 0033-2909 (Print), 1939-1455 (Electronic). doi: 10.1037/h0046016. URL https://doi.org/10.1037/h0046016.

Chiang, W.-L., Zheng, L., Sheng, Y., Angelopoulos, A. N., Li, T., Li, D., Zhu, B., Zhang, H., Jordan, M. I., Gonzalez, J. E., and Stoica, I. 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.

Choi, M., Pei, J., Kumar, S., Shu, C., and Jurgens, D. Do LLMs understand social knowledge? evaluating the sociability of large language models with SocKET benchmark. In Bouamor, H., Pino, J., and Bali, K. (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 11370–11403, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.699. URL https://aclanthology.org/2023.emnlp-main.699/.

Cook, L. L. and Pitoniak, M. J. (eds.). *Educational Measurement*. Oxford University Press, 5 edition, 2025. ISBN 978-0-19-765496-5. URL https://global.oup.com/academic/product/educational-measurement-9780197654965.

Cronbach, L. J. and Meehl, P. E. Construct validity in psychological tests. *Psychological Bulletin*, 52(4):281–302,

1955. ISSN 1939-1455 (Electronic); 0033-2909 (Print). doi: https://doi.org/10.1037/h0040957. 60 references. (PsycInfo Database Record (c) 2025 APA, all rights reserved).

Dehghani, M., Tay, Y., Gritsenko, A. A., Zhao, Z., Houlsby, N., Diaz, F., Metzler, D., and Vinyals, O. The benchmark lottery, 2021. URL https://arxiv.org/abs/2107.07002.

Deveci, İ. E. and Ataman, D. 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. URL https://openreview.net/forum?id=0zDiyIGCFT.

Dongarra, J. J., Moler, C. B., Bunch, J. R., and Stewart, G. W. *LINPACK Users' Guide*. Society for Industrial and Applied Mathematics, 1979. doi: 10.1137/1.9781611971811. URL https://epubs.siam.org/doi/abs/10.1137/1.9781611971811.

Du, M., Manjunatha, V., Jain, R., Deshpande, R., Dernoncourt, F., Gu, J., Sun, T., and Hu, X. Towards interpreting and mitigating shortcut learning behavior of NLU models. In Toutanova, K., Rumshisky, A., Zettlemoyer, L., Hakkani-Tur, D., Beltagy, I., Bethard, S., Cotterell, R., Chakraborty, T., and Zhou, Y. (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 915–929, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.71. URL https://aclanthology.org/2021.naacl-main.71/.

Eriksson, M., Purificato, E., Noroozian, A., Vinagre, J., Chaslot, G., Gomez, E., and Fernandez-Llorca, D. 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. doi: 10.1609/aies.v8i1.36595. URL https://ojs.aaai.org/index.php/AIES/article/view/36595.

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. URL https://openreview.net/forum?id=dJ01PpOozV.

Fourrier, C., Habib, N., Lozovskaya, A., Szafer, K., and Wolf, T. Open llm leaderboard v2. https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard, 2024.

Gao, B., Song, F., Yang, Z., Cai, Z., Miao, Y., Dong, Q., Li, L., Ma, C., Chen, L., Xu, R., Tang, Z., Wang, B., Zan, D., Quan, S., Zhang, G., Sha, L., Zhang, Y., Ren, X., Liu, T., and Chang, B. Omni-MATH: A universal olympiad level mathematic benchmark for large language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=yaqPf0KAlN.

Gehman, S., Gururangan, S., Sap, M., Choi, Y., and Smith, N. A. RealToxicityPrompts: Evaluating neural toxic degeneration in language models. In Cohn, T., He, Y., and Liu, Y. (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 3356–3369, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.301. URL https://aclanthology.org/2020.findings-emnlp.301/.

Ghazal, A., Rabl, T., Hu, M., Raab, F., Poess, M., Crolotte, A., and Jacobsen, H.-A. Bigbench: towards an industry standard benchmark for big data analytics. In *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data*, SIGMOD '13, pp. 1197–1208, New York, NY, USA, 2013. Association for Computing Machinery. ISBN 9781450320375. doi: 10.1145/2463676.2463712. URL https://doi.org/10.1145/2463676.2463712.

Golchin, S. and Surdeanu, M. Time travel in LLMs: Tracing data contamination in large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=2Rwq6c3tvr.

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., Yang, A., Fan, A., Goyal, A., Hartshorn, A., Yang, A., Mitra, A., Sravankumar, A., Korenev, A., Hinsvark, A., Rao, A., Zhang, A., Rodriguez, A., Gregerson, A., Spataru, A., Roziere, B., Biron, B., Tang, B., Chern, B., Caucheteux, C., Nayak, C., Bi, C., Marra, C., McConnell, C., Keller, C., Touret, C., Wu, C., Wong, C., Ferrer, C. C., Nikolaidis, C., Allonsius, D., Song, D., Pintz, D., Livshits, D., Wyatt, D., Esiobu, D., Choudhary, D., Mahajan, D., Garcia-Olano, D., Perino, D., Hupkes, D., Lakomkin, E., AlBadawy, E., Lobanova, E., Dinan, E., Smith, E. M., Radenovic, F., Guzmán, F., Zhang, F., Synnaeve, G., Lee, G., Anderson, G. L., Thattai, G., Nail, G., Mialon, G., Pang, G., Cucurell, G., Nguyen, H., Korevaar, H., Xu, H., Touvron, H., Zarov, I., Ibarra, I. A., Kloumann, I., Misra, I., Evtimov, I., Zhang, J., Copet, J., Lee, J., Geffert, J., Vranes, J., Park, J., Mahadeokar, J., Shah, J., van der Linde, J., Billock, J., Hong, J., Lee, J., Fu, J., Chi, J., Huang, J., Liu, J., Wang, J., Yu, J., Bitton, J., Spisak, J., Park, J., Rocca, J., Johnstun, J., Saxe, J., Jia, J., Alwala, K. V., Prasad, K., Upasani, K., Plawiak, K., Li, K., Heafield, K., Stone, K., El-Arini, K., Iyer, K., Malik, K., Chiu, K., Bhalla, K., Lakhotia, K., Rantala-Yeary, L., van der Maaten, L., Chen, L., Tan, L., Jenkins, L., Martin, L., Madaan, L., Malo, L., Blecher, L., Landzaat, L., de Oliveira, L., Muzzi, M., Pasupuleti, M., Singh, M., Paluri, M., Kardas, M., Tsimpoukelli, M., Oldham, M., Rita, M., Pavlova, M., Kambadur, M., Lewis, M., Si, M., Singh, M. K., Hassan, M., Goyal, N., Torabi, N., Bashlykov, N., Bogoychev, N., Chatterji, N., Zhang, N., Duchenne, O., Çelebi, O., Alrassy, P., Zhang, P., Li, P., Vasic, P., Weng, P., Bhargava, P., Dubal, P., Krishnan, P., Koura, P. S., Xu, P., He, Q., Dong, Q., Srinivasan, R., Ganapathy, R., Calderer, R., Cabral, R. S., Stojnic, R., Raileanu, R., Maheswari, R., Girdhar, R., Patel, R., Sauvestre, R., Polidoro, R., Sumbaly, R., Taylor, R., Silva, R., Hou, R., Wang, R., Hosseini, S., Chennabasappa, S., Singh, S., Bell, S., Kim, S. S., Edunov, S., Nie, S., Narang, S., Raparthy, S., Shen, S., Wan, S., Bhosale, S., Zhang, S., Vandenhende, S., Batra, S., Whitman, S., Sootla, S., Collot, S., Gururangan, S., Borodinsky, S., Herman, T., Fowler, T., Sheasha, T., Georgiou, T., Scialom, T., Speckbacher, T., Mihaylov, T., Xiao, T., Karn, U., Goswami, V., Gupta, V., Ramanathan, V., Kerkez, V., Gonguet, V., Do, V., Vogeti, V., Albiero, V., Petrovic, V., Chu, W., Xiong, W., Fu, W., Meers, W., Martinet, X., Wang, X., Wang, X., Tan, X. E., Xia, X., Xie, X., Jia, X., Wang, X., Goldschlag, Y., Gaur, Y., Babaei, Y., Wen, Y., Song, Y., Zhang, Y., Li, Y., Mao, Y., Coudert, Z. D., Yan, Z., Chen, Z., Papakipos, Z., Singh, A., Srivastava, A., Jain, A., Kelsey, A., Shajnfeld, A., Gangidi, A., Victoria, A., Goldstand, A., Menon, A., Sharma, A., Boesenberg, A., Baevski, A., Feinstein, A., Kallet, A., Sangani, A., Teo, A., Yunus, A., Lupu, A., Alvarado, A., Caples, A., Gu, A., Ho, A., Poulton, A., Ryan, A., Ramchandani, A., Dong, A., Franco, A., Goyal, A., Saraf, A., Chowdhury, A., Gabriel, A., Bharambe, A., Eisenman, A., Yazdan, A., James, B., Maurer, B., Leonhardi, B., Huang, B., Loyd, B., Paola, B. D., Paranjape, B., Liu, B., Wu, B., Ni, B., Hancock, B., Wasti, B., Spence, B., Stojkovic, B., Gamido, B., Montalvo, B., Parker, C., Burton, C., Mejia, C., Liu, C., Wang, C., Kim, C., Zhou, C., Hu, C., Chu, C.-H., Cai, C., Tindal, C., Feichtenhofer, C., Gao, C., Civin, D., Beaty, D., Kreymer, D., Li, D., Adkins, D., Xu, D., Testuggine, D., David, D., Parikh, D., Liskovich, D., Foss, D., Wang, D., Le, D., Holland, D., Dowling, E., Jamil, E., Montgomery, E., Presani, E., Hahn, E., Wood, E., Le, E.-T., Brinkman, E., Arcaute, E., Dunbar, E., Smothers, E., Sun, F., Kreuk, F., Tian, F., Kokkinos, F., Ozgenel, F., Caggioni, F., Kanayet, F., Seide, F., Florez, G. M., Schwarz, G., Badeer, G., Swee, G., Halpern, G., Herman, G., Sizov, G., Guangyi, Zhang, Lakshminarayanan, G., Inan, H., Shojanazeri, H., Zou, H., Wang, H., Zha, H., Habeeb, H., Rudolph, H., Suk, H., Aspegren, H., Goldman, H., Zhan, H., Damlaj, I., Molybog, I., Tufanov, I., Leontiadis, I., Veliche, I.-E., Gat, I., Weissman, J., Geboski, J., Kohli, J., Lam, J., Asher, J., Gaya, J.-B., Marcus, J., Tang, J., Chan, J., Zhen, J., Reizenstein, J., Teboul, J., Zhong, J., Jin, J., Yang, J., Cummings, J., Carvill, J., Shepard, J., McPhie, J., Torres, J., Ginsburg, J., Wang, J., Wu, K., U, K. H., Saxena, K., Khandelwal, K., Zand, K., Matosich, K., Veeraraghavan, K., Michelena, K., Li, K., Jagadeesh, K., Huang, K., Chawla, K., Huang, K., Chen, L., Garg, L., A, L., Silva, L., Bell, L., Zhang, L., Guo, L., Yu, L., Moshkovich, L., Wehrstedt, L., Khabsa, M., Avalani, M., Bhatt, M., Mankus, M., Hasson, M., Lennie, M., Reso, M., Groshev, M., Naumov, M., Lathi, M., Keneally, M., Liu, M., Seltzer, M. L., Valko, M., Restrepo, M., Patel, M., Vyatskov, M., Samvelyan, M., Clark, M., Macey, M., Wang, M., Hermoso, M. J., Metanat, M., Rastegari, M., Bansal, M., Santhanam, N., Parks, N., White, N., Bawa, N., Singhal, N., Egebo, N., Usunier, N., Mehta, N., Laptev, N. P., Dong, N., Cheng, N., Chernoguz, O., Hart, O., Salpekar, O., Kalinli, O., Kent, P., Parekh, P., Saab, P., Balaji, P., Rittner, P., Bontrager, P., Roux, P., Dollar, P., Zvyagina, P., Ratanchandani, P., Yuvraj, P., Liang, Q., Alao, R., Rodriguez, R., Ayub, R., Murthy, R., Nayani, R., Mitra, R., Parthasarathy, R., Li, R., Hogan, R., Battey, R., Wang, R., Howes, R., Rinott, R., Mehta, S., Siby, S., Bondu, S. J., Datta, S., Chugh, S., Hunt, S., Dhillon, S., Sidorov, S., Pan, S., Mahajan, S., Verma, S., Yamamoto, S., Ramaswamy, S., Lindsay, S., Lindsay, S., Feng, S., Lin, S., Zha, S. C., Patil, S., Shankar, S., Zhang, S., Zhang, S., Wang, S., Agarwal, S., Sajuyigbe, S., Chintala, S., Max, S., Chen, S., Kehoe, S., Satterfield, S., Govindaprasad, S., Gupta, S., Deng, S., Cho, S., Virk, S., Subramanian, S., Choudhury, S., Goldman, S., Remez, T., Glaser, T., Best, T., Koehler, T., Robinson, T., Li, T., Zhang, T., Matthews, T., Chou, T., Shaked, T., Vontimitta, V., Ajayi, V., Montanez, V., Mohan, V., Kumar, V. S., Mangla, V., Ionescu, V., Poenaru, V., Mihailescu, V. T., Ivanov, V., Li, W., Wang, W., Jiang, W., Bouaziz, W., Constable, W., Tang, X., Wu, X., Wang, X., Wu, X., Gao, X., Kleinman, Y., Chen, Y., Hu, Y., Jia, Y., Qi, Y., Li, Y., Zhang, Y., Zhang, Y., Adi, Y., Nam, Y., Yu, Wang, Zhao, Y., Hao, Y., Qian, Y., Li, Y., He, Y., Rait, Z., DeVito, Z., Rosnbrick, Z., Wen, Z., Yang, Z., Zhao, Z., and Ma, Z. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Gulliksen, H. *Theory of Mental Tests*. Wiley Publications in Psychology. John Wiley & Sons, Hoboken, NJ, 1950. doi: 10.1037/13240-000.

Heineman, D., Hofmann, V., Magnusson, I., Gu, Y., Smith, N. A., Hajishirzi, H., Lo, K., and Dodge, J. Signal and noise: A framework for reducing uncertainty in language model evaluation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=sAFottNlra.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

Henrysson, S. Correction of item-total correlations in item analysis. *Psychometrika*, 28(2):211–218, 1963. doi: 10.1007/BF02289618.

Huang, Y., Shi, J., Li, Y., Fan, C., Wu, S., Zhang, Q., Liu, Y., Zhou, P., Wan, Y., Gong, N. Z., and Sun, L. Metatool benchmark for large language models: Deciding whether to use tools and which to use. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=R0c2qtalgG.

Jacovi, A., Caciularu, A., Goldman, O., and Goldberg, Y. Stop uploading test data in plain text: Practical strategies for mitigating data contamination by evaluation benchmarks. In Bouamor, H., Pino, J., and Bali, K. (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 5075–5084, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.308. URL https://aclanthology.org/2023.emnlp-main.308/.

Jiang, H., Yi, X., Wei, Z., Xiao, Z., Wang, S., and Xie, X. Raising the bar: Investigating the values of large language models via generative evolving testing. In *Forty-second International Conference on Machine Learning*, 2025a. URL https://openreview.net/forum?id=0REM9ydeLZ.

Jiang, X., Chang, D., and Xu, X. 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*, 2025b. URL https://openreview.net/forum?id=WRMjVoJfCY.

Kamradt, G. There are 4 stages in a benchmark lifecycle. X (formerly Twitter) post, Nov 2025. URL https://x.com/GregKamradt/status/1991920852443836749. Accessed: 2026-01-16.

Kiela, D., Bartolo, M., Nie, Y., Kaushik, D., Geiger, A., Wu, Z., Vidgen, B., Prasad, G., Singh, A., Ringshia, P., Ma, Z., Thrush, T., Riedel, S., Waseem, Z., Stenetorp, P., Jia, R., Bansal, M., Potts, C., and Williams, A. Dynabench: Rethinking benchmarking in NLP. In Toutanova, K., Rumshisky, A., Zettlemoyer, L., Hakkani-Tur, D., Beltagy, I., Bethard, S., Cotterell, R., Chakraborty, T., and Zhou, Y. (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4110–4124, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.324. URL https://aclanthology.org/2021.naacl-main.324/.

Kim, E., Li, S., Khalil, S., and Shin, H. J. STAIR-AIG: Optimizing the automated item generation process through human-AI collaboration for critical thinking assessment. In Kochmar, E., Alhafni, B., Bexte, M., Burstein, J., Horbach, A., Laarmann-Quante, R., Tack, A., Yaneva, V., and Yuan, Z. (eds.), *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pp. 920–930, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-270-1. doi: 10.18653/v1/2025.bea-1.69. URL https://aclanthology.org/2025.bea-1.69/.

Le Bras, R., Swayamdipta, S., Bhagavatula, C., Zellers, R., Peters, M. E., Sabharwal, A., and Choi, Y. Adversarial filters of dataset biases. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org, 2020.

Li, F., Hogg, D. C., and Cohn, A. G. Advancing spatial reasoning in large language models: An in-depth evaluation and enhancement using the stepgame benchmark. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(17):18500–18507, Mar. 2024. doi: 10.1609/aaai.v38i17.29811. URL https://ojs.aaai.org/index.php/AAAI/article/view/29811.

Li, M., Jiao, H., Zhou, T., Zhang, N., Peters, S., and Lissitz, R. W. Item difficulty modeling using fine-tuned small and large language models. In Wilson, J., Ormerod, C., and Beiting Parrish, M. (eds.), *Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con): Coordinated Session Papers*, pp. 48–55, Wyndham Grand Pittsburgh, Downtown, Pittsburgh, Pennsylvania, United States, October 2025a. National Council on Measurement in Education (NCME). ISBN 979-8-218-84230-7. URL https://aclanthology.org/2025.aimecon-sessions.5/.

Li, T., Chiang, W.-L., Frick, E., Dunlap, L., Wu, T., Zhu, B., Gonzalez, J. E., and Stoica, I. From crowdsourced data to high-quality benchmarks: Arena-hard and benchbuilder pipeline. In *Forty-second International Con-*

ference on Machine Learning, 2025b. URL https://openreview.net/forum?id=KfTf9vFvSn.

Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., Zhang, Y., Narayanan, D., Wu, Y., Kumar, A., Newman, B., Yuan, B., Yan, B., Zhang, C., Cosgrove, C., Manning, C. D., Re, C., Acosta-Navas, D., Hudson, D. A., Zelikman, E., Durmus, E., Ladhak, F., Rong, F., Ren, H., Yao, H., WANG, J., Santhanam, K., Orr, L., Zheng, L., Yuksekgonul, M., Suzgun, M., Kim, N., Guha, N., Chatterji, N. S., Khattab, O., Henderson, P., Huang, Q., Chi, R. A., Xie, S. M., Santurkar, S., Ganguli, S., Hashimoto, T., Icard, T., Zhang, T., Chaudhary, V., Wang, W., Li, X., Mai, Y., Zhang, Y., and Koreeda, Y. Holistic evaluation of language models. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=iO4LZibEqW. Featured Certification, Expert Certification, Outstanding Certification.

Liao, Q. V. and Xiao, Z. Rethinking model evaluation as narrowing the socio-technical gap, 2025. URL https://arxiv.org/abs/2306.03100.

Lin, B. Y., Deng, Y., Chandu, K., Brahman, F., Ravichander, A., Pyatkin, V., Dziri, N., Bras, R. L., and Choi, Y. Wildbench: Benchmarking llms with challenging tasks from real users in the wild, 2024a. URL https://arxiv.org/abs/2406.04770.

Lin, F., Xie, S., Dai, Y., Yao, W., Lang, T., and Zhang, Y. IDGen: Item discrimination induced prompt generation for LLM evaluation. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024b. URL https://openreview.net/forum?id=zv4UISZzp5.

Liu, J., Nam, Y., Cui, X., and Swayamdipta, S. 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. URL https://openreview.net/forum?id=Ok4RwKoM39.

Liu, Y. L., Blodgett, S. L., Cheung, J., Liao, Q. V., Olteanu, A., and Xiao, Z. ECBD: Evidence-centered benchmark design for NLP. In Ku, L.-W., Martins, A., and Srikumar, V. (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16349–16365, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.861. URL https://aclanthology.org/2024.acl-long.861/.

Magnusson, I., Tai, N., Bogin, B., Heineman, D., Hwang, J. D., Soldaini, L., Bhagia, A., Liu, J., Groeneveld, D., Tafjord, O., Smith, N. A., Koh, P. W., and Dodge, J. Datadecide: How to predict best pretraining data with small experiments. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=p9YlQPF8fE.

Muennighoff, N., Tazi, N., Magne, L., and Reimers, N. MTEB: Massive text embedding benchmark. In Vlachos, A. and Augenstein, I. (eds.), *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 2014–2037, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.eacl-main.148. URL https://aclanthology.org/2023.eacl-main.148/.

Nahum, O., Calderon, N., Keller, O., Szpektor, I., and Reichart, R. Are LLMs better than reported? detecting label errors and mitigating their effect on model performance. In Christodoulopoulos, C., Chakraborty, T., Rose, C., and Peng, V. (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 26782–26809, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1360. URL https://aclanthology.org/2025.emnlp-main.1360/.

Ni, J., Xue, F., Yue, X., Deng, Y., Shah, M., Jain, K., Neubig, G., and You, Y. Mixeval: Deriving wisdom of the crowd from LLM benchmark mixtures. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=6A29LUZhfv.

Nie, Y., Williams, A., Dinan, E., Bansal, M., Weston, J., and Kiela, D. Adversarial NLI: A new benchmark for natural language understanding. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J. (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4885–4901, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.441. URL https://aclanthology.org/2020.acl-main.441/.

OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., Bello, I., Berdine, J., Bernadett-Shapiro, G., Berner, C., Bogdonoff, L., Boiko, O., Boyd, M., Brakman, A.-L., Brockman, G., Brooks, T., Brundage, M., Button, K., Cai, T., Campbell, R., Cann, A., Carey, B., Carlson, C., Carmichael, R., Chan, B., Chang, C., Chantzis, F., Chen, D., Chen, S., Chen, R., Chen, J., Chen, M., Chess, B., Cho, C., Chu, C., Chung, H. W., Cummings, D., Currier, J., Dai, Y., Decareaux, C., Degry, T., Deutsch, N., Deville, D., Dhar, A., Dohan, D., Dowling, S., Dunning, S., Ecoffet, A., Eleti, A., Eloundou, T., Farhi, D., Fedus, L., Felix, N., Fishman, S. P., Forte, J., Fulford, I., Gao, L., Georges, E., Gibson, C., Goel, V., Gogineni, T., Goh, G., Gontijo-Lopes, R., Gordon, J., Grafstein, M., Gray, S., Greene, R., Gross, J., Gu, S. S., Guo, Y., Hallacy, C., Han, J., Harris, J., He, Y., Heaton, M., Heidecke, J., Hesse, C., Hickey, A., Hickey, W., Hoeschele, P., Houghton, B., Hsu, K., Hu, S., Hu, X., Huizinga, J., Jain, S., Jain, S., Jang, J., Jiang, A., Jiang, R., Jin, H., Jin, D., Jomoto, S., Jonn, B., Jun, H., Kaftan, T., Łukasz Kaiser, Kamali, A., Kanitscheider, I., Keskar, N. S., Khan, T., Kilpatrick, L., Kim, J. W., Kim, C., Kim, Y., Kirchner, J. H., Kiros, J., Knight, M., Kokotajlo, D., Łukasz Kondraciuk, Kondrich, A., Konstantinidis, A., Kosic, K., Krueger, G., Kuo, V., Lampe, M., Lan, I., Lee, T., Leike, J., Leung, J., Levy, D., Li, C. M., Lim, R., Lin, M., Lin, S., Litwin, M., Lopez, T., Lowe, R., Lue, P., Makanju, A., Malfacini, K., Manning, S., Markov, T., Markovski, Y., Martin, B., Mayer, K., Mayne, A., McGrew, B., McKinney, S. M., McLeavey, C., McMillan, P., McNeil, J., Medina, D., Mehta, A., Menick, J., Metz, L., Mishchenko, A., Mishkin, P., Monaco, V., Morikawa, E., Mossing, D., Mu, T., Murati, M., Murk, O., Mély, D., Nair, A., Nakano, R., Nayak, R., Neelakantan, A., Ngo, R., Noh, H., Ouyang, L., O'Keefe, C., Pachocki, J., Paino, A., Palermo, J., Pantuliano, A., Parascandolo, G., Parish, J., Parparita, E., Passos, A., Pavlov, M., Peng, A., Perelman, A., de Avila Belbute Peres, F., Petrov, M., de Oliveira Pinto, H. P., Michael, Pokorny, Pokrass, M., Pong, V. H., Powell, T., Power, A., Power, B., Proehl, E., Puri, R., Radford, A., Rae, J., Ramesh, A., Raymond, C., Real, F., Rimbach, K., Ross, C., Rotsted, B., Roussez, H., Ryder, N., Saltarelli, M., Sanders, T., Santurkar, S., Sastry, G., Schmidt, H., Schnurr, D., Schulman, J., Selsam, D., Sheppard, K., Sherbakov, T., Shieh, J., Shoker, S., Shyam, P., Sidor, S., Sigler, E., Simens, M., Sitkin, J., Slama, K., Sohl, I., Sokolowsky, B., Song, Y., Staudacher, N., Such, F. P., Summers, N., Sutskever, I., Tang, J., Tezak, N., Thompson, M. B., Tillet, P., Tootoonchian, A., Tseng, E., Tuggle, P., Turley, N., Tworek, J., Uribe, J. F. C., Vallone, A., Vijayvergiya, A., Voss, C., Wainwright, C., Wang, J. J., Wang, A., Wang, B., Ward, J., Wei, J., Weinmann, C., Welihinda, A., Welinder, P., Weng, J., Weng, L., Wiethoff, M., Willner, D., Winter, C., Wolrich, S., Wong, H., Workman, L., Wu, S., Wu, J., Wu, M., Xiao, K., Xu, T., Yoo, S., Yu, K., Yuan, Q., Zaremba, W., Zellers, R., Zhang, C., Zhang, M., Zhao, S., Zheng, T., Zhuang, J., Zhuk, W., and Zoph, B. Gpt-4 technical report, 2024. URL https://arxiv.org/abs/2303.08774.

Ott, S., Barbosa-Silva, A., Blagec, K., Brauner, J., and Samwald, M. Mapping global dynamics of benchmark creation and saturation in artificial intelligence. *Nature Communications*, 13(1):6793, 2022. ISSN 2041-1723. doi: 10.1038/s41467-022-34591-0.

Perlitz, Y., Gera, A., Arviv, O., Yehudai, A., Bandel, E., Shnarch, E., Shmueli-Scheuer, M., and Choshen, L. 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. URL https://openreview.net/forum?id=JezpMjurwV.

Petroni, F., Piktus, A., Fan, A., Lewis, P., Yazdani, M., De Cao, N., Thorne, J., Jernite, Y., Karpukhin, V., Maillard, J., Plachouras, V., Rocktäschel, T., and Riedel, S. KILT: a benchmark for knowledge intensive language tasks. In Toutanova, K., Rumshisky, A., Zettlemoyer, L., Hakkani-Tur, D., Beltagy, I., Bethard, S., Cotterell, R., Chakraborty, T., and Zhou, Y. (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2523–2544, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.200. URL https://aclanthology.org/2021.naacl-main.200/.

Polo, F. M., Weber, L., Choshen, L., Sun, Y., Xu, G., and Yurochkin, M. tinybenchmarks: evaluating llms with fewer examples. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Raji, D., Denton, E., Bender, E. M., Hanna, A., and Paullada, A. Ai and the everything in the whole wide world benchmark. In Vanschoren, J. and Yeung, S. (eds.), *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/file/084b6fbb10729ed4da8c3d3f5a3ae7c9-Paper-round2.pdf.

Reckase, M. D. 18 multidimensional item response theory. *Handbook of statistics*, 26:607–642, 2006.

Rein, D., Hou, B. L., Stickland, A. C., Petty, J., Pang, R. Y., Dirani, J., Michael, J., and Bowman, S. R. GPQA: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=Ti67584b98.

Reuel, A., Hardy, A., Smith, C., Lamparth, M., Hardy, M., and Kochenderfer, M. Betterbench: Assessing AI benchmarks, uncovering issues, and establishing best practices. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*,

2024. URL https://openreview.net/forum?id=hcOq2buakM.

Salaudeen, O. E., Reuel, A., Ahmed, A. M., Bedi, S., Robertson, Z., Sundar, S., Domingue, B. W., Wang, A., and Koyejo, S. 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. URL https://openreview.net/forum?id=2Bw6uC49QF.

Sculley, D., Cukierski, W., Culliton, P., Dane, S., Demkin, M. M., Holbrook, R., Howard, A., Mooney, P. T., Reade, W., Risdal, M., and Keating, N. 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. URL https://openreview.net/forum?id=Rxd2TpV6Eg.

Son, H. S. Validity evaluation for the data used for artificial intelligence system. In Bi, Y., Bhatia, R., and Kapoor, S. (eds.), *Intelligent Systems and Applications*, pp. 362–369, Cham, 2020. Springer International Publishing. ISBN 978-3-030-29516-5.

Sühr, T., Dorner, F. E., Salaudeen, O., Kelava, A., and Samadi, S. Stop evaluating ai with human tests, develop principled, ai-specific tests instead, 2025. URL https://arxiv.org/abs/2507.23009.

Truong, S. T., Tu, Y., Liang, P., Li, B., and Koyejo, S. Reliable and efficient amortized model-based evaluation. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=HDbWrsgkB9.

Udell, M., Horn, C., Zadeh, R., Boyd, S., et al. Generalized low rank models. *Foundations and Trends® in Machine Learning*, 9(1):1–118, 2016.

UL Enterprise. Pcmark 2002. Benchmark software, 2002. URL https://benchmarks.ul.com/legacy-benchmarks.

Valmeekam, K., Olmo, A., Sreedharan, S., and Kambhampati, S. Large language models still can't plan (a benchmark for LLMs on planning and reasoning about change). In *NeurIPS 2022 Foundation Models for Decision Making Workshop*, 2022. URL https://openreview.net/forum?id=wUU-7XTL5XO.

Wagstaff, K. L. Machine learning that matters. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, ICML'12, pp. 1851–1856, Madison, WI, USA, 2012. Omnipress. ISBN 9781450312851.

Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. R. *SuperGLUE: a stickier benchmark for general-purpose language understanding systems*. Curran Associates Inc., Red Hook, NY, USA, 2019.

Wang, Y., Ma, X., Zhang, G., Ni, Y., Chandra, A., Guo, S., Ren, W., Arulraj, A., He, X., Jiang, Z., Li, T., Ku, M., Wang, K., Zhuang, A., Fan, R., Yue, X., and Chen, W. 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. URL https://openreview.net/forum?id=y10DM6R2r3.

Westen, D. and Rosenthal, R. Quantifying construct validity: Two simple measures. *Journal of Personality and Social Psychology*, 84(3):608–618, 2003. ISSN 0022-3514. doi: 10.1037/0022-3514.84.3.608. URL https://doi.org/10.1037/0022-3514.84.3.608.

White, C., Dooley, S., Roberts, M., Pal, A., Feuer, B., Jain, S., Shwartz-Ziv, R., Jain, N., Saifullah, K., Dey, S., Shubh-Agrawal, Sandha, S. S., Naidu, S. V., Hegde, C., LeCun, Y., Goldstein, T., Neiswanger, W., and Goldblum, M. Livebench: A challenging, contamination-limited LLM benchmark. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=sKYHBTAxVa.

Wu, S., Bao, H., Li, S., Holtzman, A., and Evans, J. A. Mapping overlaps in benchmarks through perplexity in the wild, 2025. URL https://arxiv.org/abs/2509.23488.

Xiao, Z., Zhang, S., Lai, V., and Liao, Q. 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. URL https://openreview.net/forum?id=KfJffhdWO1.

Xu, C., Yan, N., Guan, S., Jin, C., Mei, Y., Guo, Y., and Kechadi, T. DCR: Quantifying data contamination in LLMs evaluation. In Christodoulopoulos, C., Chakraborty, T., Rose, C., and Peng, V. (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 23002–23020, Suzhou, China, November 2025a. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1173. URL https://aclanthology.org/2025.emnlp-main.1173/.

Xu, X., Wu, Z., Qiao, R., Verma, A., Shu, Y., Wang, J., Niu, X., He, Z., Chen, J., Zhou, Z., Lau, G. K. R., Dao, H., Agussurja, L., Sim, R. H. L., Lin, X., Hu, W., Dai, Z., Koh, P. W., and Low, B. K. H. Position paper: Data-centric AI in the age of large language models. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N. (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 11895–11913, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.695. URL https://aclanthology.org/2024.findings-emnlp.695/.

Xu, Z., Xie, S., Lv, Q., Xiao, S., Song, L., Wenjuan, S., and Lin, F. Diagnosing failures in large language models' answers: Integrating error attribution into evaluation framework. In Che, W., Nabende, J., Shutova, E., and Pilehvar, M. T. (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 21148–21165, Vienna, Austria, July 2025b. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.1089. URL https://aclanthology.org/2025.findings-acl.1089/.

Yao, J., Jin, P., Bao, K., Yu, Q., Bhardwaj, K., Su, C., Wang, J., ZHU, Y., Devare, S., Mosk-Aoyama, D., Dong, Z., Srinivasan, V. K., Zhang, Y., Kuchaiev, O., Jiao, J., and Zhu, B. The measure of all measures: Quantifying LLM benchmark quality. In *NeurIPS 2025 Workshop on Evaluating the Evolving LLM Lifecycle: Benchmarks, Emergent Abilities, and Scaling*, 2025a. URL https://openreview.net/forum?id=HpnGmTIs7Z.

Yao, S., Shinn, N., Razavi, P., and Narasimhan, K. R. τ-bench: A benchmark for Tool-Agent-User interaction in real-world domains. In *The Thirteenth International Conference on Learning Representations*, 2025b. URL https://openreview.net/forum?id=roNSXZpUDN.

Zhang, A. K., Klyman, K., Mai, Y., Levine, Y., Zhang, Y., Bommasani, R., and Liang, P. Position: Language model developers should report train-test overlap. In *Forty-second International Conference on Machine Learning Position Paper Track*, 2025. URL https://openreview.net/forum?id=J5MmGPWKfb.

Zhang, H., Chen, Y., and Li, X. A note on exploratory item factor analysis by singular value decomposition. *Psychometrika*, 85(2):358–372, 2020.

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., Liu, P., Nie, J.-Y., and Wen, J.-R. A survey of large language models, 2025. URL https://arxiv.org/abs/2303.18223.

A. Additional Analysis Results

Here, we present additional analysis results complementing Sec. 5. Table 3 and Fig. 7 report the results of the GLRM analysis conducted in Sec. 5.2 on MMLU. Fig. 8 shows item clusters from four benchmark datasets in HELM, revealed by K-means clustering over item factor loadings derived from GLRM. Fig. 9 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 3. 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 ChatGPT, 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

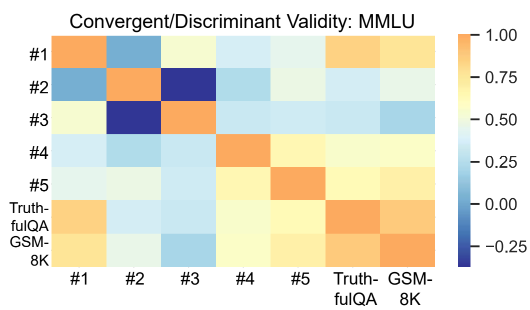


Figure 7. Convergent/discriminant evidence of the four sub-constructs (#1 - # 5) on MMLU.

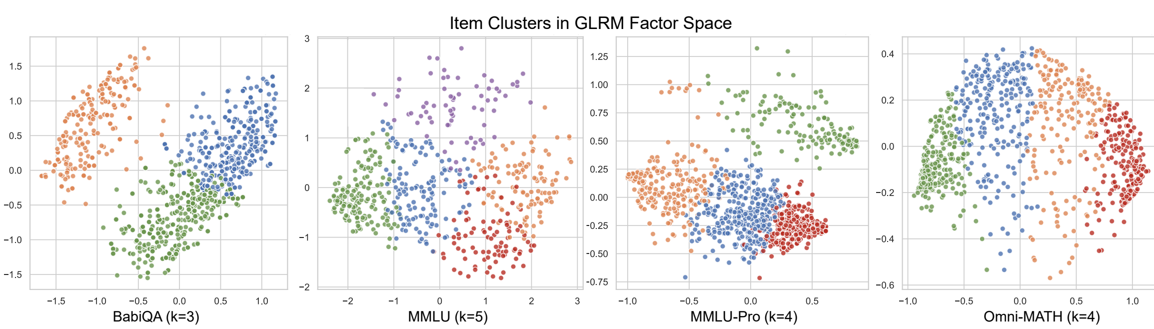
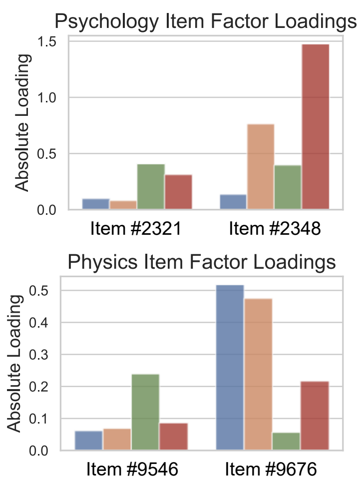


Figure 8. Clusters from four benchmark datasets in HELM revealed by K-means clustering over item factor loadings from GLRM.



Item #2321 Describe the four major types of conduct disorders and the main characteristics of each.
Item #2348 You receive a letter from the Ethics Committee asking for information about a former client who has filed a complaint against her current therapist. You stopped seeing the client over seven years ago, you should:

GLRM Factor
- #1 (Formal, quantitative, multi-step modeling)
- #2 (Domain-specific recall and simple reasoning)
- #3 (Conceptual understanding and explanation)
- #4 (Applied synthesis and case-based judgment)

Item #9546 The mutual induction of electric and magnetic fields can produce:
Item #9676 The line width of a He-Ne laser is 10^3 Hz. The operating wavelength is 6328Å and the power is 1 milliwatt. (a) How many photons are emitted per second? (b) If the output beam is 1 mm in diameter…

Figure 9. Example items with different maximum factor loadings within the same subject (psychology and physics) in `MMLU-Pro`.