From Leaderboards to Decisions: A Psychometric Perspective on Criterion-Referenced Evaluation of Large Language Models

Chengyuan Yao[1], Zhen Xu[1], Jiayu Zheng[2] and Renzhe Yu[1]

[1]Teachers College, Columbia University
[2]Johns Hopkins University

Abstract

Large language models (LLMs) are increasingly deployed in high-stakes domains, where stakeholders need to decide whether and under what conditions a model can be trusted and deployed. Such judgments are analogous to criterion-referenced assessment in human measurement, where decisions are made relative to predefined standards rather than relative rankings. However, prevailing LLM evaluation practices, which are mostly centered on leaderboards and isolated benchmarks, are ill-suited for criterion-referenced decision making under heterogeneous conditions, and benchmark scores are often treated as decisive evidence despite concerns about their validity and context sensitivity. In this paper, we propose a psychometric perspective on criterion-referenced LLM evaluation, in which target LLM properties (e.g. fairness) are conceptualized as latent constructs and evaluation is treated as a measurement problem. From this perspective, benchmark scores are viewed as imperfect indicators of an underlying constructs conditioned on specific evaluation settings. We discuss the potential of this conceptualization for informing the interpretation of benchmark results and for addressing persistent validity and sensitivity challenges in current evaluation practice. We further discuss how existing psychometric frameworks can support LLM evaluation and identify open challenges for future work.

Keywords

LLM Evaluation, Psychometrics, Measurement, Criterion-referenced Assessments, Responsible AI

1. Introduction

As large language models (LLMs) are increasingly deployed in real-world settings, evaluation results have become a primary source of evidence for decisions about model readiness and deployment. In many of these settings, stakeholders are not merely comparing models, but determining whether a model meets a required standard for use. For example, the Preparedness Framework released by OpenAI explicitly states that model safety is evaluated prior to deployment and that a model will not be released if it exceeds a predefined risk threshold [1]. Similarly, in the context of transport category aircraft, NASA has indicated that existing evaluation evidence is insufficient to support the use of generative AI for engineering design and certification practices [2]. In the social sciences and other high-stakes domains, the use of AI systems is likewise subject to requirements related to safety, fairness, and other normative considerations [3, 4].

Such judgments are inherently thresholded and resemble criterion-referenced assessments in psychological measurement, where certification depends on whether performance evidence meets predefined standards [5]. Decades of research have shown that criterion-referenced decisions place strong demands on the validity, consistency, and interpretability of the underlying evidence [6, 7, 8]. However, in LLM evaluation, comparable requirements and validation practices remain underdeveloped. Prevailing approaches rely heavily on leaderboards and isolated benchmarks, which are often treated as if performance on a single benchmark directly reflects whether a model possesses a target property. Yet benchmark scores alone frequently fail to support generalizable claims without explicit validation [9].

LLM Psychometrics: Applying Assessment Theories to Evaluate LLMs and LLM-Supported Learners, LAK 2026 Workshop
cy2706@tc.columbia.edu (C. Yao); zx2393@tc.columbia.edu (Z. Xu); jzheng59@jh.edu (J. Zheng); renzheyu@tc.columbia.edu (R. Yu)
0009-0003-4498-2675 (C. Yao); 0009-0004-3131-910X (Z. Xu); 0000-0001-7719-5270 (J. Zheng); 0000-0002-2375-3537 (R. Yu)
© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073
published 2026-11-15

CEUR-WS.org/Vol-4330/paper7.pdf

Beyond the particular dataset, evaluation conditions, and scoring procedures on which they are built, it is often unclear what a given benchmark score actually represents. This ambiguity is particularly consequential for properties such as fairness and safety, where acceptable performance is inherently context-dependent [10] and where unmodeled shifts in usage conditions can alter who is helped or harmed by a system.

These challenges are further compounded by the sensitivity of LLM performance to evaluation conditions. Observed performance can vary substantially across prompting strategies, task formats, contextual framings, and other methodological choices [11, 12, 13]. When such sources of variation are not explicitly accounted for, benchmark outcomes may reflect evaluation-induced effects rather than underlying model properties. Under these conditions, benchmark results can yield unstable or even contradictory conclusions about whether a model meets a given standard. Moreover, benchmarks targeting similar properties are frequently developed independently, with heterogeneous task designs and evaluation taxonomy.

In this paper, we provide a psychometric perspective on criterion-referenced LLM evaluation. Under this view, benchmark scores are not endpoints of evaluation, but imperfect indicators of an underlying property (a latent construct) that cannot be directly observed. We illustrate the potential of this conceptualization for addressing validity and sensitivity challenges in current evaluation practice. We also discuss the promise of existing psychometric frameworks for evaluation of LLMs, as well as open challenges for future work.

2. Criterion-Referenced Assessment: From Psychological Measurement to LLM Evaluation

In psychological measurement literature, a fundamental distinction is commonly drawn between norm-referenced and criterion-referenced assessment. Norm-referenced assessment is designed to compare individuals or systems relative to one another, with scores interpreted in terms of rank or percentile standing [14]. In contrast, criterion-referenced assessment is concerned with determining whether performance meets predefined standards or criteria, independent of how others perform [14]. The primary interpretive goal for criterion-referenced assessment is therefore not relative ranking, but classification: whether an individual is deemed to have achieved an acceptable level of competence, mastery, or readiness.

2.1. Criterion-Referenced Assessment in Psychological Measurement

Criterion-referenced assessment is a well-established paradigm in psychological measurement for determining whether individuals meet predefined standards of performance. It is widely used in educational assessment to evaluate mastery of specified learning objectives and in professional certification and licensure to determine whether candidates satisfy minimum competency requirements for practice. For example, examinations such as the United States Medical Licensing Examination (USMLE) are used to assess whether examinees meet a common professional standard for medical practice [15], while standardized language proficiency tests such as TOEFL and IELTS are used to determine whether individuals meet minimum language requirements for academic study [16].

Criterion-referenced assessment is typically conceptualized as a multi-stage measurement process rather than a single act of scoring. Instrument construction constitutes the first stage and involves both the conceptualization of the target construct and its operationalization through assessment design. Conceptualization specifies the latent construct to be measured and the relevant performance domain, while operationalization translates this specification into concrete measurement instruments. In large-scale assessments such as USMLE or TOEFL, these design choices are guided by domain expertise and define the scope of claims that assessment outcomes are intended to support.

Following construction, instrument evaluation focuses on validity and reliability. Validity concerns the extent to which observed scores support the intended interpretations and uses of an assessment,

whereas reliability concerns the consistency and stability of measurement [17].

The final stage involves score interpretation and inference, in which evaluated performance evidence is used to support criterion-referenced decisions. Interpretation requires the specification of decision rules, such as cut scores, that translate scores into categorical judgments. In psychological measurement, these demands are addressed through established processes of standard setting, validation, and evaluation of decision reliability [18]. In large-scale testing programs, such as those administered by Educational Testing Service (ETS), cut scores are determined through structured procedures involving qualified domain experts and measurement professionals. These procedures are typically accompanied by empirical evaluations of classification error and systematic documentation of the rationale and consequences associated with the chosen standards.

In addition, criterion-referenced decisions are not purely technical but inherently and context-dependent. The same proficiency score may be sufficient for undergraduate admission but insufficient for teaching assistant certification. For example, in U.S. universities, international students may be required to achieve a higher speaking subscore (e.g., a minimum of 24 on the TOEFL speaking section) to meet teaching assistant eligibility requirements, even if their overall score satisfies general admission criteria [19].

2.2. Criterion-Referenced Evaluation in Large Language Models

In practice, evaluation of large language models is also frequently used in a criterion-referenced manner. Many properties of interest, particularly responsible AI properties such as safety and fairness, are evaluated not to rank models relative to one another, but to determine whether predefined standards are met, thereby informing deployment decisions and access control [20]. For example, safety evaluations are routinely used to decide whether a model can be released to the public and whether additional mitigations are required prior to deployment [1]. Similarly, fairness evaluations are invoked to assess whether a model's behavior meets ethical expectations. In educational contexts, LLM evaluations are likewise used to judge whether model outputs meet minimum standards for classroom use or instructional support.

Despite this criterion-referenced use, current benchmark-based evaluation practices exhibit limitations. One limitation concerns construct validity. Research has suggested that contemporary machine learning evaluation often moves directly from high-level normative concepts to measurement instruments with limited conceptual systematization in between [21]. In addition, recent work has argued that limitations in benchmark validity constrain the extent to which benchmark outcomes can support generalizable claims about model properties [9]. Consequently, the validity arguments required to interpret benchmark scores as evidence for an underlying normative property are frequently underdeveloped. As a result, without validation, benchmark scores can provide a weak basis for criterion-referenced judgments about whether a model satisfies a predefined normative standard.

A second limitation concerns the sensitivity of evaluation outcomes to evaluation conditions. Prior research has shown that observed LLM performance can vary across prompting strategies, task formats, contextual framings, and decoding settings [11, 12, 13]. From a measurement perspective, such heterogeneity indicates that multiple sources of systematic variance influence observed outcomes, rather than reflecting a single stable property of the model. However, current evaluation practice rarely models these sources of variation explicitly. Consequently, criterion-referenced decisions based on benchmark scores may be unstable, as pass–fail judgments can change across reasonable evaluation conditions.

3. A Psychometric Perspective on Criterion-Referenced LLM Evaluation

We conceptualize criterion-referenced LLM evaluation as a measurement problem and illustrate how a psychometric perspective can help address the limitations identified in the previous section. Under criterion-referenced use, properties of interest, such as safety, fairness, or pedagogical appropriateness,

are treated as latent constructs that cannot be directly observed, but must be inferred from performance on evaluation instruments. Accordingly, benchmarks function as measurement instruments and provide imperfect indicators of the target property that are subject to measurement error and method effects.

In response to the first limitation concerning construct validity, which arises at the stages of instrument construction and evaluation, psychometric practice emphasizes explicit conceptualization of the target construct to be measured and careful alignment between construct definitions and their operationalization. Recent work, building on foundational work in measurement theory [22], has brought measurement concepts into the evaluation of generative AI systems by framing evaluation as a form of social measurement and adapting psychometric frameworks for construct conceptualization and measurement processes. These frameworks offer a set of lenses for interrogating whether evaluation instruments validly represent the intended target capability and whether the resulting measurements support their intended interpretations and uses [21].

The second limitation concerns the sensitivity of evaluation outcomes to evaluation conditions. Generalizability Theory (G-Theory) provides a natural framework for modeling such sources of variation by treating evaluation conditions as facets of measurement [23]. In its simplest form, observed benchmark outcomes can be expressed as

$$X_{\text{obs}} = \mu_{\text{true}} + \sum_{f \in \mathcal{F}} \eta_f + \sum_{f \in \mathcal{F}} (\mu \times \eta)_f + \varepsilon, \tag{1}$$

where X_{obs} denotes the observed benchmark outcome, μ_{true} represents the model's underlying latent ability on the target property, η_f captures systematic variation introduced by evaluation facet f, $(\mu \times \eta)_f$ represents interaction effects reflecting differential sensitivity of models to specific facets, and ε denotes residual or unsystematic error.

In the context of LLM evaluation, relevant facets may include, but are not limited to:

- **Prompting strategies**: zero-shot, few-shot, chain-of-thought, red-teaming, etc.;
- **Decoding parameters**: temperature, top-k / top-p, etc.;
- **Evaluation format**: multiple-choice questions, generative tasks etc.;
- **Contextual or domain variance**: topic domain (e.g., medical, political, educational), as well as cultural or linguistic context.

These facets can influence observed performance even when the underlying ability on the target property remains unchanged. Therefore, disentangling these effects supports more reasonable interpretation of benchmark outcomes and provides a basis for assessing the robustness of criterion-referenced judgments under plausible variations in evaluation conditions.

4. Promise of Existing Psychometric Frameworks

When properties are treated as latent constructs, a wide range of psychometric frameworks becomes applicable. These frameworks offer principled tools for modeling the relationship between observed task performance and underlying abilities. We introduce several commonly used frameworks and discuss their potential for LLM evaluation. While our focus is on criterion-referenced settings, these frameworks are not limited to such uses and can also support the evaluation of general model capabilities.

Exploratory Factor Analysis (EFA). Exploratory Factor Analysis is commonly used in psychological measurement to examine whether a set of observed indicators reflects a shared underlying construct or multiple distinct dimensions [24]. In the context of LLM evaluation, EFA can be applied to assess whether benchmarks intended to measure the same property exhibit a coherent latent structure. For instance, EFA can reveal whether performance across a collection of safety- or fairness-related benchmarks is dominated by a single general factor or instead separates into multiple sub-dimensions. Furthermore, benchmarks that exhibit weak or unstable factor loadings may capture construct-irrelevant variance or operationalize qualitatively distinct aspects of model behavior that diverge from the target construct.

Multi-Trait Multi-Method (MTMM) Analysis. Multi-Trait Multi-Method analysis provides a classical framework for distinguishing variance attributable to the target construct from variance introduced by the measurement method. In psychological measurement, MTMM designs are routinely used to evaluate convergent and divergent validity across multiple instruments and formats [25]. Analogously, in LLM evaluation, MTMM can be used to examine whether benchmarks targeting the same property provide consistent evidence across different task formats, such as multiple-choice versus open-ended generation. At the same time, MTMM can help identify method effects by clarifying when discrepancies in observed performance are driven by evaluation format rather than differences in the underlying ability.

Item Response Theory (IRT). Item Response Theory (IRT) jointly characterizes respondents and items within a common measurement framework and models observed performance as a function of item characteristics, such as difficulty and discrimination, and a respondent's latent construct level. In educational and psychological measurement, once item characteristics have been established, this formulation enables principled prediction of how individuals at different construct levels are expected to perform across items of varying difficulty. In LLM evaluation, we treat a benchmark suite as a measurement instrument, with individual benchmark tasks serving as measurement items. Under this convention, once the target latent construct level of a given model has been estimated and item characteristics have been established, IRT enables inference about how the model is expected to perform on previously unobserved items within the same construct domain. In addition, IRT provides tools for evaluating item functioning, helping to identify poorly functioning or redundant items based on their informativeness. This perspective supports more efficient measurement designs by leveraging item informativeness with respect to the latent property, including adaptive item selection strategies such as computerized adaptive testing. Under such designs, items are selected conditional on a model's current estimated property level to maximize expected information while reducing evaluation burden [26, 27].

Many-Facet Rasch Models (MFR). Many-Facet Rasch Models extend standard item-based measurement models by explicitly incorporating multiple sources of systematic variation, such as raters, tasks, or contextual conditions [28]. When adapted to LLM evaluation, a similar approach can be used to account for systematic variances. By estimating facet effects directly, MFR-style models yield adjusted estimates that better reflect construct-level behavior rather than artifacts of specific evaluation conditions.

5. Challenges

High-dimensional regime. In the recent Microsoft Research Fellowship call, one of the research challenges is the high-dimensional regime of applying psychometric methods to LLM evaluation [29]. Classical psychometric methods are typically developed for contexts involving large numbers of examinees and a comparatively limited set of items. LLM evaluation, however, exhibits the reverse configuration: a small number of models assessed across a broad and continuously growing set of tasks and benchmarks. This 'few models × many tests' structure raises identifiability, stability, and overfitting concerns for latent variable modeling. Therefore, adapting psychometric estimation to this high-dimensional setting can be an important direction for future work. Recent research [30] notes that additional models can be obtained by evaluating ablated variants of the same model or by testing multiple versions of a model family on shared tasks. Beyond these approaches, additional strategies and methodological extensions remain to be explored.

Data contamination. Psychometric models typically assume that measurement items have not been previously encountered by the learner and that item difficulty reflects genuine novelty. In the context of LLM evaluation, however, many existing benchmarks are likely to be partially included in pretraining or alignment data. Such contamination complicates psychometric interpretation: a benchmark may appear easy not because the model has internalized the underlying construct, but because it has effectively

memorized the content. As a result, standard assumptions about item difficulty, discrimination, or diagnostic value may no longer hold. These concerns point to the need for contamination-aware measurement in LLM evaluation. Future work may explore strategies such as constructing held-out or regenerable benchmark items, estimating separate item parameters for memorized versus new items, and using synthetic or procedurally generated evaluation tasks that can be refreshed over time.

Declaration on Generative AI

During the preparation of this work, the author(s) used Chat-GPT-5.1 in order to: Grammar and spelling check. The author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] OpenAI, Preparedness Framework (Version 2), Technical Report, OpenAI, 2025. URL: https://cdn.openai.com/pdf/18a02b5d-6b67-4cec-ab64-68cdfbddebcd/preparedness-framework-v2.pdf, last updated: April 15, 2025.

[2] M. S. Graydon, S. M. Lehman, Examining proposed uses of llms to produce or assess assurance arguments (2025).

[3] N. A. Smuha, Regulation 2024/1689 of the eur. parl. & council of june 13, 2024 (eu artificial intelligence act), International Legal Materials (2025) 1–148.

[4] U. E. Sector, Ai and education: guidance for policy-makers, United Nations Educational, Scientific and Cultural Organization: Paris, France (2021).

[5] C. W. Smith, Criterion-referenced assessment, 1973.

[6] A. E. R. Association, A. P. Association, N. C. on Measurement in Education, et al., Standards for educational and psychological testing, (No Title) (1985).

[7] S. Messick, Validity. em r. linn (org.), educational measurement.(13-103), New York, NY: American Council on Education and Macmillan Publishing Company (1989).

[8] H. Huynh, On the reliability of decisions in domain-referenced testing, Journal of educational measurement (1976) 253–264.

[9] O. Salaudeen, A. Reuel, A. Ahmed, S. Bedi, Z. Robertson, S. Sundar, B. Domingue, A. Wang, S. Koyejo, Measurement to meaning: A validity-centered framework for ai evaluation, arXiv preprint arXiv:2505.10573 (2025).

[10] A. Z. Jacobs, H. Wallach, Measurement and fairness, in: Proceedings of the 2021 ACM conference on fairness, accountability, and transparency, 2021, pp. 375–385.

[11] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, et al., Chain-of-thought prompting elicits reasoning in large language models, Advances in neural information processing systems 35 (2022) 24824–24837.

[12] S. Singh, A. Alyakin, D. A. Alber, J. Stryker, A. P. S. Tong, K. Sangwon, N. Goff, M. De La Paz, M. Hernandez-Rovira, K. Y. Park, et al., The pitfalls of multiple-choice questions in generative ai and medical education, Scientific reports 15 (2025) 42096.

[13] W. Zhou, S. Zhang, H. Poon, M. Chen, Context-faithful prompting for large language models, in: H. Bouamor, J. Pino, K. Bali (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2023, Association for Computational Linguistics, Singapore, 2023, pp. 14544–14556. URL: https://aclanthology.org/2023.findings-emnlp.968/. doi:10.18653/v1/2023.findings-emnlp.968.

[14] D. J. Weiss, M. L. Davison, Review of test theory and methods, Technical Report, Computerized Adaptive Testing Laboratory, Psychometric Methods Program ..., 1981.

[15] United States Medical Licensing Examination (USMLE), Usmle service transition, https://www.usmle.org, 2025. Accessed January 2026.

[16] Columbia University, English proficiency requirements, https://undergrad.admissions.columbia.edu/apply/international/english-proficiency, n.d. Columbia Undergraduate Admissions.
[17] L. Cohen, L. Manion, K. Morrison, Validity and reliability, in: Research methods in education, Routledge, 2017, pp. 245–284.
[18] M. Ziek y, M. Perie, A Primer on Setting Cut Scores on Tests of Educational Achievement, Technical Report, Educational Testing Service, Princeton, NJ, 2008. Including excerpts from *Passing Scores: A Manual for Setting Standards of Performance on Educational and Occupational Tests* by Samuel Livingston and Michael Zieky.
[19] Graduate College, University of Illinois Urbana-Champaign, Additional instructions for international applicants, https://grad.illinois.edu/admissions/application-instructions/additional-instructions-international-applicants, 2025. Accessed January 2026.
[20] T. Shevlane, S. Farquhar, B. Garfinkel, M. Phuong, J. Whittlestone, J. Leung, D. Kokotajlo, N. Marchal, M. Anderljung, N. Kolt, et al., Model evaluation for extreme risks, arXiv preprint arXiv:2305.15324 (2023).
[21] H. Wallach, M. Desai, N. Pangakis, A. F. Cooper, A. Wang, S. Barocas, A. Chouldechova, C. Atalla, S. L. Blodgett, E. Corvi, et al., Evaluating generative ai systems is a social science measurement challenge, arXiv preprint arXiv:2411.10939 (2024).
[22] R. Adcock, D. Collier, Measurement validity: A shared standard for qualitative and quantitative research, American political science review 95 (2001) 529–546.
[23] R. J. Shavelson, N. M. Webb, Generalizability theory, in: Handbook of complementary methods in education research, Routledge, 2012, pp. 309–322.
[24] M. Norris, L. Lecavalier, Evaluating the use of exploratory factor analysis in developmental disability psychological research, Journal of autism and developmental disorders 40 (2010) 8–20.
[25] D. T. Campbell, D. W. Fiske, Convergent and discriminant validation by the multitrait-multimethod matrix., Psychological bulletin 56 (1959) 81.
[26] H. Jiang, X. Yi, Z. Wei, Z. Xiao, S. Wang, X. Xie, Raising the bar: Investigating the values of large language models via generative evolving testing, arXiv preprint arXiv:2406.14230 (2024).
[27] R. R. Meijer, M. L. Nering, Computerized adaptive testing: Overview and introduction, Applied psychological measurement 23 (1999) 187–194.
[28] T. Eckes, Introduction to many-facet Rasch measurement, Peter Lang, 2023.
[29] Microsoft Research, Microsoft research fellowship, https://www.microsoft.com/en-us/research/academic-program/microsoft-research-fellowship/, 2025. Accessed: 2025-12-04.
[30] R. Burnell, H. Hao, A. R. Conway, J. H. Orallo, Revealing the structure of language model capabilities, arXiv preprint arXiv:2306.10062 (2023).