

Review of 445 LLM benchmarks finds widespread construct-validity weaknesses

Source copy kept by wheresthe.ai for the story "Review of 445 LLM benchmarks finds widespread construct-validity weaknesses" (<https://wheresthe.ai/d/oxford-benchmark-construct-validity-2025-11/>).

Source: <https://arxiv.org/abs/2511.04703>

Text saved 18 Oct 2026 (the page as published).

SHA-256 of the saved text: 18ad79a459b0b327718b4abae356dfb513b4be22c69de7388ea3439afd9e951

-->

Computer Science > Computation and Language

arXiv:2511.04703 (cs)

[Submitted on 3 Nov 2025]

Title: Measuring what Matters: Construct Validity in Large Language Model Benchmarks

Authors: Andrew M. Bean , Ryan Othniel Kearns , Angelika Romanou , Franziska Sofia Hafner , Harry Mayne , Jan Batzner , Negar Foroutan , Chris Schmitz , Karolina Korgul , Hunar Batra , Oishi Deb , Emma Beharry , Cornelius Emde , Thomas Foster , Anna Gausen , María Grandury , Simeng Han , Valentin Hofmann , Lujain Ibrahim , Hazel Kim , Hannah Rose Kirk , Fangru Lin , Gabrielle Kaili-May Liu , Lennart Luetzgau , Jabez Magomere , Jonathan Rystrom , Anna Sotnikova , Yushi Yang , Yilun Zhao , Adel Bibi , Antoine Bosselut , Ronald Clark , Arman Cohan , Jakob Foerster , Yarin Gal , Scott A. Hale , Inioluwa Deborah Raji , Christopher Summerfield , Philip H.S. Torr , Cozmin Ududec , Luc Rocher , Adam Mahdi View a PDF of the paper titled Measuring what Matters: Construct Validity in Large Language Model Benchmarks, by Andrew M. Bean and 41 other authors

View PDF HTML (experimental)

Abstract: Evaluating large language models (LLMs) is crucial for both assessing their capabilities and identifying safety or robustness issues prior to deployment. Reliably measuring abstract and complex phenomena such as 'safety' and 'robustness' requires strong construct validity, that is, having measures that represent what matters to the phenomenon. With a team of 29 expert reviewers, we conduct a systematic review of 445 LLM benchmarks from leading conferences in natural language processing and machine learning. Across the reviewed articles, we find patterns related to the measured phenomena, tasks, and scoring metrics which undermine the validity of the resulting claims. To address these shortcomings, we provide eight key recommendations and detailed actionable guidance to researchers and practitioners in developing LLM benchmarks.

Comments: 39th Conference on Neural Information Processing Systems (NeurIPS 2025) Track on Datasets and Benchmarks

Subjects:

Computation and Language (cs.CL) ; Artificial Intelligence (cs.AI)

Cite as: arXiv:2511.04703 [cs.CL]

(or arXiv:2511.04703v1 [cs.CL] for this version)

<https://doi.org/10.48550/arXiv.2511.04703>

Focus to learn more

arXiv-issued DOI via DataCite

Submission history From: Ryan Kearns [view email] [v1] Mon, 3 Nov 2025 17:39:40 UTC (722 KB)

Full-text links: Access Paper:

View a PDF of the paper titled Measuring what Matters: Construct Validity in Large Language Model Benchmarks,

by Andrew M. Bean and 41 other authors View PDF HTML (experimental) TeX Source

view license

Current browse context:

cs.CL

< prev

| next >

new | recent | 2025-11

Change to browse by:

cs cs.AI

References & Citations

NASA ADS Google Scholar

Semantic Scholar

export BibTeX citation Loading...

BibTeX formatted citation

×

loading...

Data provided by:

Bookmark

Bibliographic Tools

Bibliographic and Citation Tools

Bibliographic Explorer Toggle

Bibliographic Explorer (What is the Explorer?)

Connected Papers Toggle

Connected Papers (What is Connected Papers?)

Litmaps Toggle

Litmaps (What is Litmaps?)

scite.ai Toggle

scite Smart Citations (What are Smart Citations?)

Code, Data, Media

Code, Data and Media Associated with this Article

alphaXiv Toggle

alphaXiv (What is alphaXiv?)

Links to Code Toggle

CatalyzeX Code Finder for Papers (What is CatalyzeX?)

DagsHub Toggle

DagsHub (What is DagsHub?)

GotitPub Toggle

[Gotit.pub \(What is GotitPub? \)](#)

[Huggingface Toggle](#)

[Hugging Face \(What is Huggingface? \)](#)

[ScienceCast Toggle](#)

[ScienceCast \(What is ScienceCast? \)](#)

[Demos](#)

[Demos](#)

[Replicate Toggle](#)

[Replicate \(What is Replicate? \)](#)

[Spaces Toggle](#)

[Hugging Face Spaces \(What is Spaces? \)](#)

[Spaces Toggle](#)

[TXYZ.AI \(What is TXYZ.AI? \)](#)

[Related Papers](#)

[Recommenders and Search Tools](#)

[Link to Influence Flower](#)

[Influence Flower \(What are Influence Flowers? \)](#)

[Core recommender toggle](#)

[CORE Recommender \(What is CORE? \)](#)

[Author](#)

[Venue](#)

[Institution](#)

[Topic](#)

[About arXivLabs](#)

[arXivLabs: experimental projects with community collaborators](#)

arXivLabs is a framework that allows collaborators to develop and share new arXiv features directly on our website.

Both individuals and organizations that work with arXivLabs have embraced and accepted our values of openness, community, excellence, and user data privacy. arXiv is committed to these values and only works with partners that adhere to them.

Have an idea for a project that will add value for arXiv's community? [Learn more about arXivLabs](#) .

[Which authors of this paper are endorsers? | Disable MathJax \(What is MathJax? \)](#)