A Tutorial on Measurement Science for Language Model Evaluation

Sang Truong[1], **Serena Wang**[2], **Rodolfo Corona**[3], **Sanmi Koyejo**[1]
[1]Stanford University, [2]University of British Columbia, [3]Harvard University
sttruong@cs.stanford.edu

Abstract

Benchmark scores are used to support claims about language model capabilities, comparisons, and suitability for deployment. The evidence needed for these claims depends on what is measured, how observations are collected, and how evaluation shapes model developer behavior. This tutorial brings together measurement theory, statistical modeling, and mechanism design to examine six common claims in language model evaluation. Part I introduces measurement principles for interpreting benchmark results and assessing the scope of the claims they support. Part II examines how evaluations shape model developer incentives and how their design can encourage model improvements that serve diverse user needs. For each claim, we explain when it can hold, where it breaks, and how to formulate a better-supported interpretation. Worked examples and short discussions connect these principles to practical evaluation decisions. Participants leave with practical tools for stating evaluation targets, selecting relevant evidence, and reporting the limits of conclusions drawn from benchmark scores.

1 Description and Learning Outcomes

Benchmarks are a central instrument through which the NLP community compares language models and assesses their suitability for use. Yet an observed score and the inference drawn from it are different objects. "The model gets 87% on this test" describes performance under a particular protocol. Claims that a model reasons well, outperforms another model, or is suitable for a deployment setting require additional evidence. Broad benchmark coverage alone does not establish that the evaluation measures the intended construct or supports the intended use (Messick, 1995; Raji et al., 2021; Liang et al., 2023).

This tutorial draws on measurement theory, statistical inference, and mechanism design to examine what benchmark results justify and what evaluations reward. We ask when evidence from an evaluation supports its intended use, and how evaluation design shapes model development. The emphasis is on drawing defensible conclusions and designing evaluations that encourage improvements serving diverse user needs.

The tutorial is organized around six recurring claims in two parts. **Part I: Measurement Theory** examines construct validity, comparable and efficient measurement, variation across evaluation conditions, and deployment relevance. **Part II: Incentive-aware Robust Evaluation** examines how evaluation design can reward progress that serves users. Across both parts, each claim follows the same progression: *what holds, where it breaks, and a better-supported claim*. We distinguish sampling uncertainty, which can sometimes be reduced by collecting more observations, from specification problems, for which more observations of the wrong target do not supply the missing validity evidence.

Learning outcomes. Participants will learn to:

- State the construct, target population, protocol, and deployment decision behind an evaluation claim.
- Assess whether benchmark aggregation supports the named capability.
- Explain how linked calibration and adaptive item selection support efficient comparisons between models evaluated on different questions.
- Use response patterns to prioritize item review and distinguish shared, model-specific, and item-specific language effects.
- Check whether the evaluated outcome and system match the intended deployment, and identify incentives that can improve scores without improving that outcome.

Proceedings of the 5th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 15th International Joint Conference on Natural Language Processing: Tutorial Abstracts, pages 20–26
November 6-10, 2026 ©2026 Association for Computational Linguistics

These skills are relevant to benchmark authors, researchers comparing models, and practitioners selecting systems. The tutorial emphasizes the conditions under which an evaluation is informative and the evidence needed to extend its interpretation. It connects foundational work in psychometrics with contemporary NLP case studies and economic models of evaluation.

2 Related Tutorials

The tutorial builds on complementary treatments of evaluation metrics (Khapra and Sai, 2021), evaluation practice and reproducibility (Sutawika and Schoelkopf, 2024), item response theory for NLP (Lalor et al., 2024), human-centered evaluation (Blodgett et al., 2024), and the science of benchmarking (Ma et al., 2025). Our organizing focus is the inference a practitioner wants to draw from a score. Six concrete claims connect construct validity, common-scale comparisons, evaluation conditions, deployment decisions, and incentives, with worked examples that make their assumptions and limitations explicit.

3 Target Audience and Prerequisites

The intended audience is CL/NLP researchers, students, and practitioners who build benchmarks, report model comparisons, or use evaluation results to select systems. We assume familiarity with language model evaluation and basic probability and statistics, including averages, variance, and uncertainty. Familiarity with supervised learning and maximum-likelihood estimation is helpful for the statistical examples. No prior background in psychometrics or mechanism design is required; item response theory is introduced through worked examples. Discussion prompts require no programming setup. Python familiarity is useful only for optional notebook material.

4 Tutorial Outline

The tutorial follows the sequence in Table 1: 180 minutes of instruction and discussion, with a 30-minute break between C3 and C4. Corona presents C1 and C4, Truong presents C2 and C3, and Wang presents C5 and C6, with 50 minutes per instructor. Koyejo leads the 15-minute opening and 15-minute closing. Each claim combines a motivating example, the relevant measurement or incentive concepts, and a more defensible interpretation.

Segment	Minutes
Opening: evaluation as measurement	15
C1–C3: constructs, scales, language effects	75
Break	30
C4–C6: deployment and incentives	75
Closing synthesis and discussion	15
Total	210

Table 1: Three-and-a-half-hour tutorial schedule: 180 minutes of instruction and discussion and a 30-minute break.

Opening: Evaluation as Measurement. We motivate the distinction between an observed score and a claim about capability or deployment. Examples of benchmark defects, incomplete reporting, and cultural assumptions introduce the need to state what is being measured (Burnell et al., 2023; Singh et al., 2025). The opening distinguishes sampling error from specification error and introduces the six questions below.

4.1 Part I: Measurement Theory

C1: Does averaging benchmark scores robustly establish a capability? Averaging can reduce benchmark-specific variation while preserving what the benchmarks share. That shared component need not be the capability named by their labels. We introduce convergent and discriminant validity and examine evidence about benchmarks grouped under capability and safety concepts (Messick, 1995; Desai et al., 2026; Ren et al., 2024). An application example considers how coverage and weights determine what an aggregate healthcare score represents (Bedi et al., 2026; Arora et al., 2025). Participants learn to ask both whether benchmarks agree and whether that agreement supports the intended construct, rather than relying on a shared label alone.

C2: Can we compare models reliably with fewer evaluations? Taking a coherent construct as the starting point, we ask how to compare models evaluated on different question subsets. Expensive agent trajectories motivate collecting fewer responses per model. We introduce the Rasch model as a common scale for model ability and item difficulty, emphasizing linked calibration as the basis for comparison. Amortized calibration and adaptive selection using Fisher information illustrate how to obtain informative measurements with fewer evaluations (Lord, 1980; Truong et al., 2025a). We then show how expected response associations can guide item inspection: correlation-

based diagnostics in *Fantastic Bugs* help prioritize review and uncover benchmark or grading problems (Truong et al., 2025b). A statistical flag calls for inspection; it does not establish a flaw on its own. The takeaway is that a response model can support efficient measurement and identify questions worth inspecting, subject to adequate calibration and model fit.

C3: After translation, are scores still comparable? Observed jailbreak rates vary across languages, and English is not always the least vulnerable language (Zhang et al., 2026). This empirical pattern motivates adding language as an evaluation facet. For exposition, we extend C2's Rasch model with language effects to distinguish a shared language shift from model–language and prompt–language effects. The case study shows that model families can shift in opposite directions and individual prompts can depart substantially from the shared shift; large differences can persist even for accurate translations. Participants learn to examine which aspects of the evaluation contribute to a score gap before interpreting it as a change in model safety. Language provides a concrete example of how averages can obscure systematic variation across evaluation conditions.

C4: Does a leaderboard winner fit my product? A stable ranking does not by itself establish that a model meets a deployment requirement, and model's ranking can change depending on the components it's paired with in a system. We examine whether the scored outcome matches the deployment outcome and whether the evaluated system matches the system being deployed. Examples consider differences between benchmark success and task completion (Whitfill et al., 2026; Tonja et al., 2026), as well as changes to the system through prompts, tools, or interaction protocols (Feng et al., 2025; Mizrahi et al., 2024; Liu et al., 2026). Assessment of benchmark ontology, content, and form helps participants articulate which aspects of the deployment are represented and which are missing (Corona et al., 2026). Participants learn to specify a deployment outcome and seek evidence for that outcome rather than assuming a leaderboard position answers the product question.

4.2 Part II: Incentive-aware Robust Evaluation

C5: Is evaluation gaming mostly cheating or data leakage? We broaden the discussion beyond contamination to repeated model releases, optimization for evaluated tasks, tailoring to particular user populations, and selective disclosure of evaluations. These behaviors can arise from following the incentives an evaluation creates, even without breaking its rules. Goodhart's law and mechanism design provide a language for separating developer objectives from the outcomes an evaluator wants to encourage (Goodhart, 1984; Manheim and Garrabrant, 2018; Wang et al., 2025). Participants learn to describe how a score could improve while a relevant capability, population, or deployment objective is neglected.

C6: Do private, uncontaminated test sets solve gaming? Private holdouts address some failure modes, but evaluation design must also account for adaptation, information asymmetry, and heterogeneous preferences. We introduce an evaluator–developer game and explain how randomized evaluation can change incentives for capability coverage under explicit assumptions (Dwork et al., 2015; Truong et al., 2026). Preference leaderboards illustrate a second problem: optimizing average user preference may neglect substantial minority groups. We discuss how pluralistic rankings can represent these preferences, with a brief overview of robust evaluation and counterfactual metrics (Haghtalab et al., 2026; Khalaf et al., 2026; Wang et al., 2024). The emphasis is on whether an evaluation rewards useful model development and supports decisions across user groups, under the assumptions each design requires.

Closing Synthesis and Discussion. We combine the six claims into an evaluation checklist: define the claim and target population; inspect the items and scoring; check calibration and variation across evaluation conditions; seek deployment evidence; examine incentives; and report item-level responses and the full evaluation protocol. We briefly discuss open questions about changing model behavior, validating automated scorers, and testing generalization to future tasks. The closing asks participants what evidence would make a benchmark score useful for their own setting and how they would report the supported interpretation and its limits.

Reading List. Recommended starting points are Messick (1995); Desai et al. (2026) for validity; Lord (1980); Truong et al. (2025a,b) for measurement and item diagnostics; Zhang et al. (2026); Corona et al. (2026) for evaluation conditions and

deployment; and Dwork et al. (2015); Wang et al. (2025); Haghtalab et al. (2026) for incentives and robust evaluation. The reading list combines foundational work, research by the instructors, and work by other authors. An expanded bibliography accompanies the tutorial materials.

5 Diversity and Accessibility

Linguistic, cultural, and population diversity are technical concerns throughout the tutorial. C3 examines how language effects vary across models and prompts in multilingual safety evaluation; C4 asks whose tasks and deployment contexts an evaluation represents; and C6 examines how aggregate preferences can overlook minority user groups. Benchmarks developed in English-speaking, Western contexts provide motivating examples of a broader problem: evidence from one setting need not support the same interpretation in another. The instructors bring expertise in computer science, statistics, policy, economics, and measurement. Discussion prompts require no local software installation, and publicly available slides and readings support participation beyond the conference.

6 Tutorial Materials

The tutorial website, `https://aimslab.stanford.edu/tutorial`, provides the bibliography and a central location for slides and supplementary resources. The instructional materials comprise worked examples, discussion prompts, and the closing evaluation checklist. Any optional notebooks will be linked from the same site; they are not required for participation. We will make the slides and bibliography publicly available.

7 Ethics Statement

Validity judgments depend on the intended use and on the perspectives of affected populations. Statistical diagnostics and automated audits can inform human review but cannot determine cultural appropriateness or justify deployment on their own. We discuss gaming to motivate defenses and make their assumptions explicit; the same insights could also inform strategic behavior. Better measurement can support harmful applications as well as beneficial ones, so measurement quality does not substitute for evaluating the acceptability of an application. Throughout the tutorial, we emphasize uncertainty, representation, and limits on the conclusions an evaluation supports.

Instructor Biographies

Sang Truong (Stanford University, US) is a Ph.D. candidate in Computer Science at the Stanford Artificial Intelligence Laboratory. His research develops foundational and applied methods for AI measurement sciences, drawing on tools from measurement theory, probabilistic machine learning, and psychometrics to support the scientific development of safe AI systems that serve people across diverse backgrounds and needs. He has co-organized the 1st workshop on AI Measurement Science at the Conference on Language Modeling, the 1st and 2nd Workshops on Language Models for Underserved Communities (LM4UC) at NAACL 2025 and AAAI 2026, as well as the Data-centric Machine Learning Research Workshops at ICML 2024 and ICLR 2024.

Serena Wang (University of British Columbia; serenalw@cs.ubc.ca) is an Assistant Professor of Computer Science at the University of British Columbia. She previously held a postdoctoral fellowship at Harvard University. Wang's research focuses on incentives in the AI evaluation ecosystem, including mechanism design and information design approaches to robust evaluation design. Her recently published work includes strategic evaluation to mitigate gaming behavior (Wang et al., 2024), robust leaderboard design under preference heterogeneity (Procaccia et al., 2025; Khalaf et al., 2026; Haghtalab et al., 2026), and incentives for information sharing between model developers and evaluators (Wang et al., 2025). Wang is a Canada CIFAR AI Chair, and a recipient of the Harvard Center for Research on Computation and Society Postdoctoral Fellowship, the Apple Scholars in AI/ML PhD Fellowship and the NSF Graduate Research Fellowship. She has previously organized workshops at AAAI and Harvard, and given tutorials at the Open Data Science Conference and Grace Hopper Conference.

Rodolfo Corona (Harvard University; rcorona@hks.harvard.edu) is a Postdoctoral Fellow in Technology and Geopolitics at the Harvard Kennedy School Belfer Center. He received his Ph.D. in Computer Science from UC Berkeley, where he was advised by Dan Klein and Trevor Darrell. His research spans grounded language learning, evaluations, and AI governance, having presented his work at ACL, NAACL, and NeurIPS. Rodolfo's current research is focused on AI benchmarks, studying how the cultural, regional, and deployment contexts of AI systems impact the validity of their evaluations. Rodolfo has taught international audiences, particularly in Latin American conferences. He has served as Hackathon and Practicals Chair for SALA 2026, as an organizer and instructor for the KHIPU 2025 practicals, and on the advisory board for RIIAA 6.0. He has served as a graduate student instructor for UC Berkeley's CS 288 (graduate Natural Language Processing) and CS 188 (Artificial Intelligence) courses. Rodolfo is a recipient of the NSF Graduate Research Fellowship, the Ford Foundation Fellowship, the UC Berkeley Chancellor's Fellowship, and a Fulbright Student Research Grant.

Sanmi Koyejo (Stanford University; sanmi@cs.stanford.edu) is an associate professor in the Department of Computer Science at Stanford University. He leads the Stanford Trustworthy Artificial Intelligence Research (STAIR) lab, focusing on the principles and practice of trustworthy machine learning. His research on emergent behaviors, scaling laws, and novel evaluation methodologies has been recognized with several awards, including the NeurIPS, ACL, and UAI Outstanding Paper Awards, the Presidential Early Career Award for Scientists and Engineers, a Skip Ellis Early Career Award, a Sloan Fellowship, and an NSF CAREER Award. He served as co-General Chair of NeurIPS 2022, Workshop Chair at NeurIPS 2021 and ICLR 2021, and has co-organized workshops including ICML 2024 Workshop on Theoretical Foundations of Foundation Models (TF2M) and ICLR 2024 Workshop on Reliable and Responsible Foundation Models (RRFM).

References

Rahul K Arora, Jason Wei, Rebecca Soskin Hicks, Preston Bowman, Joaquin Quiñonero-Candela, Foivos Tsimpourlas, Michael Sharman, Meghan Shah, Andrea Vallone, Alex Beutel, and 1 others. 2025. Healthbench: Evaluating large language models towards improved human health. *arXiv preprint arXiv:2505.08775*.

Suhana Bedi, Hejie Cui, Miguel Fuentes, Alyssa Unell, Michael Wornow, Juan M Banda, Nikesh Kotecha, Timothy Keyes, Yifan Mai, Mert Oez, and 1 others. 2026. Holistic evaluation of large language models for medical tasks with medhelm. *Nature medicine*, 32(3):943–951.

Su Lin Blodgett, Jackie Chi Kit Cheung, Vera Liao, and Ziang Xiao. 2024. Human-centered evaluation of language technologies. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts (EMNLP)*, pages 39–43.

Ryan Burnell, Wout Schellaert, John Burden, Tomer D. Ullman, Fernando Martinez-Plumed, Joshua B. Tenenbaum, Danaja Rutar, Lucy G. Cheke, Jascha Sohl-Dickstein, Melanie Mitchell, Douwe Kiela, Murray Shanahan, Ellen M. Voorhees, Anthony G. Cohn, Joel Z. Leibo, and Jose Hernandez-Orallo. 2023. Rethink reporting of evaluation results in AI. *Science*, 380:136–138.

Rodolfo Corona, Sang T. Truong, Nhi N. Truong, Ritwik Gupta, Fred Philippy, Atnafu Lambebo Tonja, Mena Attia, Fahim Faisal, Kaushal Kumar Maurya, Belu Ticona, Sumaya Nur Adan, Fazl Barez, Omar U. Florez, Supheakmungkol Sarin, Aseem Srivastava, Xiaoyuan Yi, Nick Haber, Dan Klein, Thamar Solorio, and 3 others. 2026. Uplifting human decision making in AI evaluation by automating benchmark validity analysis. In *Workshop on Combining Theory and Benchmarks at ICML*.

Meera Desai, Sang T. Truong, Hanna Wallach, Alex Chouldechova, A. Feder Cooper, Jean Garcia-Gathright, Daniel E. Ho, Abigail Z. Jacobs, Sanmi Koyejo, Nicholas Pangakis, and Angelina Wang. 2026. What AI benchmarks actually measure: Adapting convergent and discriminant validity to interrogate fifty-six AI benchmarks. In *Conference on Language Modeling*.

Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Roth. 2015. The reusable holdout: Preserving validity in adaptive data analysis. *Science*.

Haiwen Feng, Long Lian, Lisa Dunlap, Jiahao Shu, XuDong Wang, Renhao Wang, Trevor Darrell, Alane Suhr, and Angjoo Kanazawa. 2025. Visually prompted benchmarks are surprisingly fragile. *arXiv preprint arXiv:2512.17875*.

Charles A. E. Goodhart. 1984. Problems of monetary management: The UK experience. In *Monetary Theory and Practice*, pages 91–121. Macmillan.

Nika Haghtalab, Ariel D. Procaccia, Han Shao, Serena Lutong Wang, and Kunhe Yang. 2026. Pluralistic leaderboards. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, pages 39356–39375. PMLR.

Hadi Khalaf, Serena Lutong Wang, Daniel Halpern, Itai Shapira, Flavio Calmon, and Ariel D. Procaccia. 2026. Robust AI evaluation through maximal lotteries. In *Proceedings of the 43rd International Conference on Machine Learning*.

Mitesh M. Khapra and Ananya B. Sai. 2021. A tutorial on evaluation metrics used in natural language generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Tutorials (NAACL-HLT)*, pages 15–19.

John P. Lalor, Pedro Rodriguez, João Sedoc, and Jose Hernandez-Orallo. 2024. Item response theory for natural language processing. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Tutorial Abstracts (EACL)*, pages 9–13.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, and 31 others. 2023. Holistic evaluation of language models. *Transactions on Machine Learning Research (TMLR)*.

Fengshuo Liu, Ying Liu, Ruize Sun, Lie Luo, and Siyuan Guo. 2026. Coding agents have converged: Why the swe-bench leaderboard can no longer order its top entries, and what to measure instead. *arXiv preprint arXiv:2609.17394*.

Frederic M. Lord. 1980. *Applications of Item Response Theory to Practical Testing Problems*. Lawrence Erlbaum Associates.

Ziqiao Ma, Michael Saxon, and Xiang Yue. 2025. The science of benchmarking: What's measured, what's missing, what's next. Tutorial at the 39th Conference on Neural Information Processing Systems (NeurIPS).

David Manheim and Scott Garrabrant. 2018. Categorizing variants of goodhart's law. *arXiv preprint arXiv:1803.04585*.

Samuel Messick. 1995. Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50(9):741–749.

Moran Mizrahi, Guy Kaplan, Dan Malkin, Rotem Dror, Dafna Shahaf, and Gabriel Stanovsky. 2024. State of what art? a call for multi-prompt LLM evaluation.

Transactions of the Association for Computational Linguistics, 12:933–949.

Ariel Procaccia, Ben Schiffer, Serena Wang, and Shirley Zhang. 2025. Metritocracy: representative metrics for lite benchmarks. *Advances in Neural Information Processing Systems*, 38.

Inioluwa Deborah Raji, Emily M. Bender, Amandalynne Paullada, Emily Denton, and Alex Hanna. 2021. AI and the everything in the whole wide world benchmark. In *NeurIPS Datasets and Benchmarks Track*.

Richard Ren, Steven Basart, Adam Khoja, Alice Gatti, Long Phan, Xuwang Yin, Mantas Mazeika, Alexander Pan, Gabriel Mukobi, Ryan H. Kim, Stephen Fitz, and Dan Hendrycks. 2024. Safetywashing: Do AI safety benchmarks actually measure safety progress? In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David Ifeoluwa Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, Raymond Ng, Shayne Longpre, Sebastian Ruder, Wei-Yin Ko, Antoine Bosselut, Alice Oh, Andre Martins, Leshem Choshen, Daphne Ippolito, and 4 others. 2025. Global MMLU: Understanding and addressing cultural and linguistic biases in multilingual evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18761–18799, Vienna, Austria. Association for Computational Linguistics.

Lintang Sutawika and Hailey Schoelkopf. 2024. Challenges in language model evaluations. Tutorial at the 41st International Conference on Machine Learning (ICML).

Atnafu Lambebo Tonja, Srija Anand, Emilio Villa-Cueva, Israel Abebe Azime, Jesujoba Oluwadara Alabi, Muhidin A. Mohamed, Debela Desalegn Yadeta, Negasi Haile Abadi, Abigail Oppong, Nnaemeka Casmir Obiefuna, Idris Abdulmumin, Naome A Etori, Eric Peter Wairagala, Kanda Patrick Tshinu, Imanigirimbabazi Emmanuel, Gabofetswe Malema, Alham Fikri Aji, David Ifeoluwa Adelani, and Thamar Solorio. 2026. Afri-MCQA: Multimodal cultural question answering for African languages. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 40249–40282, San Diego, California, United States. Association for Computational Linguistics.

Sang Truong, Yuheng Tu, Percy Liang, Bo Li, and Sanmi Koyejo. 2025a. Reliable and efficient amortized model-based evaluation. *Preprint*, arXiv:2503.13335.

Sang Truong, Serena Wang, Nick Haber, and Sanmi Koyejo. 2026. Strategic evaluation: Incentivizing AI capability coverage with private benchmarks.

Sang T. Truong, Yuheng Tu, Michael Hardy, Anka Reuel, Zeyu Tang, Jirayu Burapacheep, Jonathan Perera, Chibuike Uwakwe, Benjamin W. Domingue, Nick Haber, and Sanmi Koyejo. 2025b. Fantastic bugs and where to find them in AI benchmarks. In *Conference on Neural Information Processing Systems, Datasets and Benchmarks Track*.

Serena Wang, Stephen Bates, P Aronow, and Michael Jordan. 2024. On counterfactual metrics for social welfare: Incentives, ranking, and information asymmetry. In *International Conference on Artificial Intelligence and Statistics*, pages 1522–1530. PMLR.

Serena Wang, Michael Jordan, Katrina Ligett, and R Preston McAfee. 2025. Relying on the metrics of evaluated agents. In *Proceedings of the ACM on Web Conference 2025*, pages 1468–1487.

Parker Whitfill, Cheryl Wu, Joel Becker, and Nate Rush. 2026. Many swe-bench-passing prs would not be merged into main. METR research note.

Max Zhang, Ameen Patel, Sang T. Truong, and Sanmi Koyejo. 2026. Why do safety guardrails degrade across languages? In *Conference on Language Modeling*.