AI Benchmarks: Interdisciplinary Issues and Policy Considerations

Maria Eriksson [1] **Erasmo Purificato** [2] **Arman Noroozian** [3] **João Vinagre** [1]
Guillaume Chaslot [3] **Emilia Gomez** [1] **David Fernandez-Llorca** [1]

Abstract

Artificial Intelligence (AI) benchmarks have emerged as essential for evaluating AI performance, capabilities, and risks. However, as their influence grows, concerns arise about their limitations and side effects when assessing sensitive topics such as high-impact capabilities, safety and systemic risks. In this work we summarise the results of an interdisciplinary meta-review of approximately 110 studies over the last decade (Eriksson et al., 2025), which identify key shortcomings in AI benchmarking practices, including issues in the design and application (e.g., biases, inadequate documentation, data contamination, and failures to distinguish signal from noise) and broader sociotechnical issues (e.g., over-focus on text-based and one-time evaluation logic, neglecting multimodality and interactions). We also highlight systemic flaws, such as misaligned incentives, construct validity issues, unknown unknowns, and the gaming of benchmark results. We underscore how benchmark practices are shaped by cultural, commercial and competitive dynamics that often prioritise performance at the expense of broader societal concerns. As a result, AI benchmarking may be ill-suited to provide the assurances required by policymakers. To address these challenges, it is crucial to consider key policy aspects that can help mitigate the shortcomings of current AI benchmarking practices.

1. Introduction

AI benchmarks play a central role in AI development (Raji et al., 2021) and regulation (European Union, 2024), but researchers have raised concerns about their use. Benchmarks are seen as deeply political, performative and generative, shaping the world rather than passively describing it (Grill, 2024). This paper summarises the results of an interdisciplinary meta-review of around 110 publications during the last decade (Eriksson et al., 2025), aiming to address the gap in research on AI benchmarking critique by mapping and discussing known limitations.

In computing, benchmarks are used to evaluate the performance of hardware or software systems by comparing them to a standard or reference point (Henning, 2000). In AI development, benchmarks are used to facilitate cross-model comparisons, measure performance, and track model progress (Reuel et al., 2024). We focus on software-oriented benchmarks, which are defined as a combination of test datasets and associated performance metrics, representing one or more specific tasks or capabilities (Raji et al., 2021). We primarily consider quantitative benchmarks, which are executed without direct human intervention, as opposed to qualitative benchmarks, which involve human evaluators.

We used a snowball sampling method (Jalali & Wohlin, 2012; Badampudi et al., 2015) to gather source materials, starting from the article "AI and the Everything in the Whole Wide World Benchmark" (Raji et al., 2021) and expanding through reference lists and Google Scholar citations. A conceptual diagram illustrating both the snowball sampling approach and the paper selection criteria is shown in Fig. 1. We targeted papers that primarily address benchmark critique, excluding those that present new benchmarks or simply apply them. Our collection consists of around 110 papers published between 2014 and 2024, which explicitly and primarily highlight issues with benchmarks (Mitchell et al., 2019; Orr & Crawford, 2024b; Rodriguez et al., 2021; Liu et al., 2021; Mulvin, 2021; Pinch, 1993; Marres & Stark, 2020). The number of papers per year and the cumulative trend are shown in Fig. 2. We excluded papers that propose new benchmarks, as they often reproduce assumptions about quantitative benchmarks providing a technical "fix" to AI safety and capability assessments. Our meta-review is not exhaustive, but it covers a broad range of critique aimed at benchmarking practices. We identified nine issue categories after close-reading and classifying papers and discussing these classifications within the author group. The resulting

[1]European Commission, Joint Research Centre (JRC), Seville, Spain. [2]European Commission, Joint Research Centre (JRC), Ispra, Italy. [3]European Commission, Joint Research Centre (JRC), Brussels, Belgium. Correspondence to: Maria Eriksson <maria.eriksson@ec.europa.eu>, David Fernández-Llorca <david.fernandez-llorca@ec.europa.eu>.

Workshop on Technical AI Governance (TAIG) at ICML 2025, Vancouver, Canada.

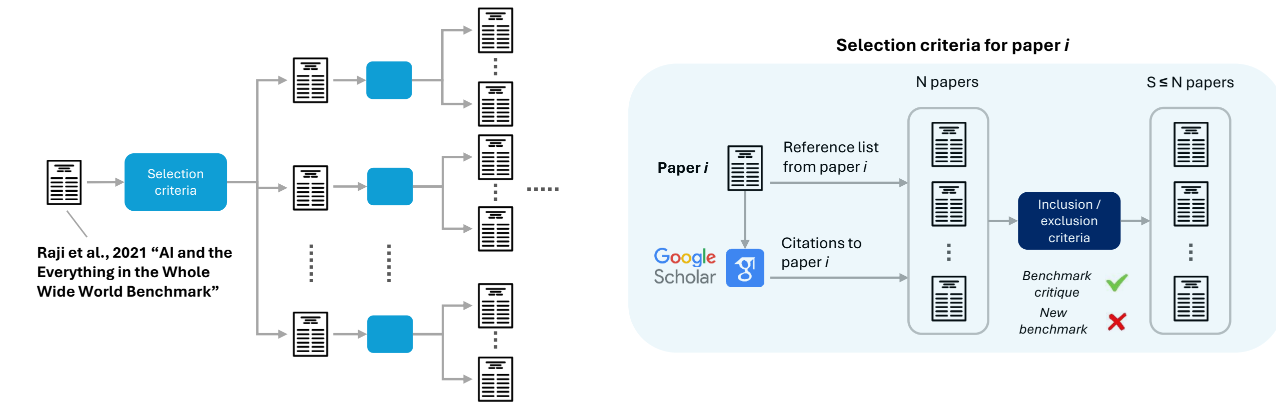

Figure 1. (Left) Illustration of snowball sampling. We start with one relevant paper and expand our search according to (Right) a set of selection criteria, using reference lists from the paper and Google Scholar citations to the paper, and applying the inclusion and exclusion criteria. This process was repeated iteratively until no additional relevant references are found.

issue areas represent the result of these discussions, acknowledging that many issues overlap and are not absolute or all-encompassing. We focus on works that voice critique relevant to policy makers, highlight areas of concern across different modalities, and point to fundamental weaknesses in benchmark design and application. Our goal is to provide a diverse account of concerns regarding benchmarks, that can be relevant to both AI developers and policymakers.

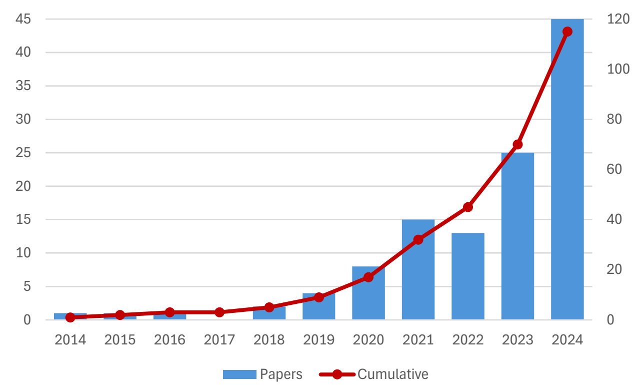

Figure 2. Publications per year for the period 2014-2024 and cumulative number of publications trend.

2. Current AI Benchmarking Issues

We summarise the main issues identified in our research, presented as a taxonomy of nine reason to be cautious with AI benchmarks. These interlinked problems, depicted in Fig. 3, are not ranked by importance or urgency, and their complexity and interdependence pose a challenge for AI evaluations.

2.1. Data Collection, Annotation, and Documentation

Limitations in collecting, annotating, and documenting AI benchmark are a significant issue in AI research, tied to broader critiques of insufficient documentation and transparency (Gebru et al., 2021; Mitchell et al., 2019; Orr & Crawford, 2024b; Simson et al., 2024; Scheuerman et al., 2021). It is often difficult to trace the origin and creation of benchmark datasets (Reuel et al., 2024; Denton et al., 2020), compromising their robustness and generalisability (Arzt & Hanbury, 2024). This issue is partly due to the low status of dataset-related work (Orr & Crawford, 2024a; Sambasivan et al., 2021) and the reuse of datasets (Koch et al., 2021), which complicates documentation of their limitations and social impact (Thylstrup et al., 2022; Park & Jeoung, 2022). Moreover, benchmarks raise ethical and legal concerns regarding copyrights, privacy, informed consent and opt-out rights (Paullada et al., 2021). The use of crowd-sourced or user-generated content from platforms such as Wikihow, Reddit or trivia websites, can lead to noisy and biased annotations (Keegan, 2024; Grill, 2024; Tsipras et al., 2020; Aroyo & Welty, 2015; Sen et al., 2015), and the absence of human performance references and difficulty rubrics can hinder evaluation of capabilities and generality (Chollet, 2019). A lack of care in creating benchmark datasets can result in AI models exploiting quirks and spurious cues rather than solving the intended task (Liao et al., 2021; Paullada et al., 2021; Geirhos et al., 2020), as seen in examples such as X-ray image classification (Oakden-Rayner et al., 2019) and LLM evaluation (Pacchiardi et al., 2024).

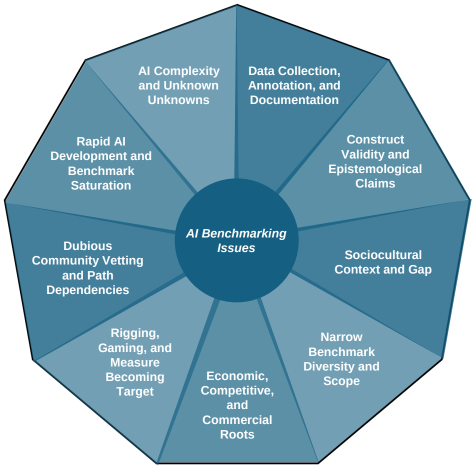

Figure 3. Proposed categorisation of current interlinked AI benchmarking issues.

2.2. Construct Validity and Epistemological Claims

Another critique of benchmarks focuses on their epistemological claims and the limits of quantitative AI tests. Many benchmarks suffer from construct validity issues, failing to measure what they claim to measure (Raji et al., 2021). This is particularly problematic when benchmarks promise to measure universal or general capabilities, as it misrepresents their actual capability. A central issue is that many benchmarks lack a clear definition of what they claim to measure, making it impossible to evaluate their success (Blodgett et al., 2021; Bartz-Beielstein et al., 2020). For example, benchmarks evaluating fairness in natural language processing have been found to have severe weaknesses in defining what is being measured (Blodgett et al., 2021). Elsewhere, research has shown strong disagreements in how benchmark tasks are *conceptualised* and *operationalised* (Subramonian et al., 2023), and found that benchmarks are applied in highly idiosyncratic ways (Röttger et al., 2024). The difficulty in defining what benchmarks evaluate persists due to the lack of a clear, stable, and absolute ground truth (Narayanan & Kapoor, 2023a). Concepts like ”bias” and ”fairness” are inherently contested and messy, leading to an ”abstraction error” that produces a false sense of certainty (Selbst et al., 2019). Many benchmark datasets have also been found to be inadequate or unuseful proxies for what they are meant to evaluate. For instance, there is a slippage in distinguishing between algorithmic ”harms” and ”wrongs” (Diberardino et al., 2024), and the content of benchmark datasets may not be reasonable substitutes for real-world scenarios (Keegan, 2024). Benchmarks consisting of professional exams have been argued to be unreliable measures of skills like medical or legal skills (Narayanan & Kapoor, 2023a), and many widely used “safety” benchmarks highly correlate with general model capabilities, raising concerns about ”safetywashing” (Ren et al., 2024). This highlights the need for a clear distinction between capabilities and risks in AI models, as severe biases and safety issues can persist even as overall capabilities improve.

2.3. Sociocultural Context and Gap

Research highlights the importance of social, economic, and cultural contexts in AI benchmark creation, use, and maintenance. There is a consensus in benchmark critique that benchmarks are ”normative instruments” that perpetuate particular epistemological perspectives (Orr & Kang, 2024). Qualitative research shows that benchmarks are shaped by shared assumptions, commitments, and dependencies (Engdahl, 2024; Michael et al., 2022; Scheuerman et al., 2021; Orr & Crawford, 2024a; Sambasivan et al., 2021; Paullada et al., 2021), such as prioritising efficiency over care, universality over contextuality or impartiality over positionality (Scheuerman et al., 2021). AI safety research and benchmark competitions are also influenced by political movements and ideologies (Ahmed et al., 2024). A key concern is the sociotechnical gap and lack of consideration for downstream real-world utility in AI benchmarking (Hutchinson et al., 2022). It is often unclear who should care about benchmark results and how they should be used in practice (Liao & Xiao, 2023). Studies have found that benchmarks often fail to address the needs of practitioners, such as medical experts (Blagec et al., 2023), and that there is a lack of attention to the practical utility of benchmarks (Jannach & Bauer, 2020). This has been argued to have led to the development of biased and energy-inefficient AI models that ignore discriminatory and environmental damages (Ethayarajh & Jurafsky, 2021).

2.4. Narrow Benchmark Diversity and Scope

Current benchmarking practices suffer from diversity issues, with a majority of benchmarks focusing on text and neglecting other modalities (Rauh et al., 2024; Weidinger et al., 2023; Röttger et al., 2024). Benchmarks for safety, risks and ethics are also lacking, with a concentration on simplistic and brittle evaluation practices (Guldimann et al., 2024). The design of benchmarks is often dominated by elite institutions, raising concerns about representation diversity (Koch et al., 2021). Additionally, AI safety evaluation practices are mostly limited to English content and datasets with under-represented minorities, neglecting multiple perspectives on complex topics like ethics and harm (McIntosh et al., 2024; Simson et al., 2024). Most benchmarks are also abstracted from their social and cultural context, relying on static, one-time testing logic (Selbst et al., 2019). This has led to calls for more multi-layered, longitudinal, and holistic evaluation methods that capture AI model performance in real-world circumstances (Weidinger et al., 2023; Mizrahi et al., 2024; Ojewale et al., 2024). Current benchmarks often fail to distinguish signal and noise, and rarely consider risks associated with multiple interacting AI systems or human actions and motivations (Reuel et al., 2024; Birhane et al., 2024). Furthermore, benchmarks often reveal little about ways of making mistakes, which is crucial for AI safety and policy enforcement (Gehrmann et al., 2023). A focus on errors and fragilities, rather than instances of success, could be useful for developers and help equalise the playing field in AI development (Gehrmann et al., 2023).

2.5. Economic, Competitive, and Commercial Roots

The competitive and commercial roots of benchmark tests have been identified as a significant contextual element in AI research. Capability-oriented benchmarks are often embedded in corporate marketing strategies, increasing AI hype, and showcasing model performance to attract customers and investors (Orr & Kang, 2024; Grill, 2024; Zhijia, 2024). Many benchmarks originate from industry and focus on tasks with high economic reward, rather than ethics and safety (Ren et al., 2024; Ethayarajh & Jurafsky, 2021). This competitive culture discourages thorough self-critique, as there is an incentive mismatch between conducting high-quality evaluations and publishing new models (Gehrmann et al., 2023). The field of AI development has become a ”giant leaderboard” where publication depends on numbers, rather than insight and explanation (Church & Hestness, 2019). The professionalisation of benchmark evaluations has transformed into an industry, with platforms like Kaggle and Grand Challenge providing support to AI competitions (Luitse et al., 2024). This has led to the issue of optimising for high benchmark scores, known as SOTA-chasing (Koch et al., 2021) or the ”benchmark effect” (Stewart, 2023), and the “fast track research” issue (Stengers, 2018) linked with the “winners curse” in AI development (Sculley et al., 2018). The growing influence of industry in AI research, with private businesses now dominating the development of large AI models, has raised concerns about the concentration of power and the potential stifling of robust AI evaluations (Ahmed et al., 2023). Scholars have warned that upholding data-intensive benchmark tests as the standard could make academic research increasingly dependent on industry-provided technological infrastructures (Koch & Peterson, 2024).

2.6. Rigging, Gaming, and Measure Becoming Target

Benchmark tests can be tricked and gamed, particularly in areas where best-practice benchmarks are missing. Researchers have noted that there are strong incentives to ”rig” benchmark tests, and that know-how for scoring high on benchmarks is often widely circulated online (Dehghani et al., 2021), including optimisation for answering multiple-choice questions, and ”fake” alignment with ethics or safety goals (Alzahrani et al., 2024; Greenblatt et al., 2024). This issue is related to Goodhart’s law: ”when a measure becomes a target, it ceases to be a good measure” (Strathern, 1997). The lack of transparency and validation in benchmark tests facilitates gaming, with many models optimised for specific benchmarks rather than general performance (Bartz-Beielstein et al., 2020; Dehghani et al., 2021; Biderman et al., 2024; Reuel et al., 2024). Data contamination, where models ingest benchmark datasets during training, is another issue that questions the integrity of AI tests (Xu et al., 2024a; Zhang et al., 2024; Besen, 2024; Magar & Schwartz, 2022; Roberts et al., 2023). Despite the known risks of data leaks, there is still a lack of reporting on data contamination tendencies during benchmark tests (Zhang et al., 2024). Additionally, ”sandbagging” involves intentionally understate a model’s capability to avoid regulation (Weij et al., 2024), further undermining the trustworthiness of benchmark evaluations, especially in a regulatory context.

2.7. Dubious Community Vetting and Path Dependencies

Benchmarks can become naturalized and reach standard status due to the culture and logic of academic citations, even if they were not intended to be widely adopted (Orr & Kang, 2024). This can happen when a new benchmark is introduced with a popular AI model, and the benchmark becomes widely cited as a side effect (Orr & Kang, 2024; Orr & Crawford, 2024a). The peer-review process can perpetuate the dominance of certain benchmarks, making it difficult for new benchmarks to gain traction (Jaton, 2021; Ott et al., 2022). This can lead to a ”benchmark lottery” where the perceived superiority of a method is influenced by factors other than algorithmic improvements (Dehghani et al., 2021). Furthermore, many influential benchmarks have been released as preprints without rigorous peer-review (McIntosh et al., 2024). The focus on methods over datasets in benchmark papers can have worrying effects when benchmarks are applied to real-world use cases (Bao et al., 2022). The current peer-review system prioritises benchmarks that are relevant from a methods perspective, rather than those with practical utility (Bao et al., 2022). This can create ”path dependencies” in AI research, reinforcing certain methodologies and research goals while stifling others (Blili-Hamelin & Hancox-Li, 2023). The dominance of certain benchmarks can also lead to a form of ”task-driven scientific monoculture” that prioritises narrow epistemic values over broader scientific progress (Koch & Peterson, 2024).

2.8. Rapid AI Development and Benchmark Saturation

The rapid development of AI models has created a challenge for benchmarks, as many are old and designed to test simpler models (Keegan, 2024; Biderman et al., 2024). For instance, prominent LLM benchmarks were designed before shifts in AI capabilities, which may affect their validity

(Biderman et al., 2024). Many benchmarks also struggle with AI models achieving very high accuracy scores, leading to saturation and rendering the benchmark ineffective (Hendrycks et al., 2021; Bowman & Dahl, 2021; Ott et al., 2022). The slow and complicated implementation of benchmark frameworks can hinder timely feedback on AI model risks, as evaluation processes can take weeks or months (McIntosh et al., 2024). This is concerning in a regulatory setting, where quick and accurate assessments are crucial. The use of thresholds to determine which AI models warrant regulatory scrutiny is also limited, as creating benchmarks that keep pace with AI development is challenging (European Union, 2024; The White House, 2023; US Department of Commerce, 2025). Recent approaches can enhance AI capabilities with reduced training compute, further complicating benchmarking efforts (Hooker, 2024).

2.9. AI Complexity and Unknown Unknowns

The complexity of AI models and the difficulty of foreseeing potential risks pose a significant challenge for benchmarks (McIntosh et al., 2024). Benchmark creators' limited knowledge and understanding of emerging AI capabilities can lead to generalist approaches that fail to address critical sector requirements, posing safety and security risks and hindering innovation (McIntosh et al., 2024). The presence of unknown and latent vulnerabilities in AI models can also make it difficult to distinguish between safe and unsafe models (Nasr et al., 2023a). Simple prompts can "break" safety barriers, revealing sensitive training data and highlighting the potential for latent vulnerabilities (Nasr et al., 2023b). Furthermore, fine-tuning AI models to address safety and security risks can degrade performance in other areas or introduce new risks (Qi et al., 2023). These challenges underscore the need for more comprehensive and nuanced benchmarking approaches that can account for the complexities and uncertainties of AI development.

3. Conclusions and Policy Considerations

Measuring AI capabilities and risks is a challenge, and benchmarks have been found to promise too much (Raji et al., 2021), be easily gamed (Weij et al., 2024; Narayanan & Kapoor, 2023b), and measure the wrong thing (Oakden-Rayner et al., 2019). They also lack documentation (Reuel et al., 2024), perpetuate cultural assumptions (Kang, 2023; Keegan, 2024), and are narrow, focusing on English and text-based models (McIntosh et al., 2024; Röttger et al., 2024; Rauh et al., 2024). These issues highlight fundamental fragilities in current efforts to quantitatively measure and mitigate harm in AI. Cars, airplanes, medical devices, and drugs are strictly regulated to ensure safety. Similarly, AI can be subject to safety assurances, and the growing interest in AI benchmarks reflects a drive to develop such regulations.

In line with previous studies (Jones et al., 2024), our meta-review suggests that the field of quantitative AI benchmarking is currently ill-suited to single-handedly (or primarily) provide the safety and capability assurances requested by policy makers. Our review also shows that from a policy perspective, relying on indicators such as citation counts to determine what benchmarks to trust is insufficient. We identify a strong incentive gap in the use of benchmarks between academic researchers (who may for example primarily be interested in methods development), corporations (who are driven by economic incentives in their use and development of benchmarks), and regulators (who have a particular responsibility to consider practical utility and potential downstream effects).

Future policymakers need to ensure that applied and trusted benchmarks are well-documented and transparent; include clearly defined tasks, metrics, and performance evaluation mechanisms to prevent capabilities misrepresentation; evaluate diversity and inclusivity in benchmark design, accounting for various perspectives and cultural contexts; apply benchmarks that target multimodal and real-world capabilities, rather than narrow tasks; continuously assess potential misuse while integrating dynamic benchmarks to prevent gaming, sandbagging, and data contamination; establish rigorous evaluation protocols to validate and update benchmark results in line with rapid model improvements; and apply benchmarks that evaluate errors and unintended consequences alongside performance and capabilities.

As our review has shown, evaluation frameworks repeatedly influence downstream AI development by becoming targets for model optimisation. Recognising the power of such a downstream influence, we stress that policymakers have a unique opportunity to shape AI evaluation, benchmark design, and ultimately AI development by setting the bar high and demanding robust benchmark practices. From a regulatory perspective, we especially identify a need for new ways of signalling *what benchmarks to trust* (i.e., trustworthy benchmarks). We do not necessarily need standardised benchmark metrics and methods. But we do need standardised methods for assessing the trustworthiness of benchmarks from an applied and regulatory perspective.

Disclaimer

The views expressed in this paper are purely those of the authors and may not, under any circumstances, be regarded as an official position of the European Commission.

References

Ahmed, N., Wahed, M., and Thompson, N. C. The growing influence of industry in AI research. *Science*, 379(6635):884–886, March 2023. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.ade2420. URL https://www.science.org/doi/10.1126/science.ade2420.

Ahmed, S., Jaźwińska, K., Ahlawat, A., Winecoff, A., and Wang, M. Field-building and the epistemic culture of AI safety. *First Monday*, April 2024. ISSN 1396-0466. doi: 10.5210/fm.v29i4.13626. URL https://firstmonday.org/ojs/index.php/fm/article/view/13626.

Alzahrani, N., Alyahya, H. A., Alnumay, Y., Alrashed, S., Alsubaie, S., Almushaykeh, Y., Mirza, F., Alotaibi, N., Altwairesh, N., Alowisheq, A., Bari, M. S., and Khan, H. When Benchmarks are Targets: Revealing the Sensitivity of Large Language Model Leaderboards, July 2024. URL http://arxiv.org/abs/2402.01781.

Aroyo, L. and Welty, C. Truth Is a Lie: Crowd Truth and the Seven Myths of Human Annotation. *AI Magazine*, 36(1):15–24, March 2015. ISSN 0738-4602, 2371-9621. doi: 10.1609/aimag.v36i1.2564. URL https://onlinelibrary.wiley.com/doi/10.1609/aimag.v36i1.2564.

Arzt, V. and Hanbury, A. Beyond the Numbers: Transparency in Relation Extraction Benchmark Creation and Leaderboards, November 2024. URL http://arxiv.org/abs/2411.05224.

Badampudi, D., Wohlin, C., and Petersen, K. Experiences from using snowballing and database searches in systematic literature studies. In *Proceedings of the 19th International Conference on Evaluation and Assessment in Software Engineering*, EASE '15, New York, NY, USA, 2015. Association for Computing Machinery. ISBN 9781450333504. doi: 10.1145/2745802.2745818. URL https://doi.org/10.1145/2745802.2745818.

Bao, M., Zhou, A., Zottola, S., Brubach, B., Desmarais, S., Horowitz, A., Lum, K., and Venkatasubramanian, S. It's COMPASlicated: The Messy Relationship between RAI Datasets and Algorithmic Fairness Benchmarks, April 2022. URL http://arxiv.org/abs/2106.05498.

Bartz-Beielstein, T., Doerr, C., Berg, D. v. d., Bossek, J., Chandrasekaran, S., Eftimov, T., Fischbach, A., Kerschke, P., Cava, W. L., Lopez-Ibanez, M., Malan, K. M., Moore, J. H., Naujoks, B., Orzechowski, P., Volz, V., Wagner, M., and Weise, T. Benchmarking in Optimization: Best Practice and Open Issues, December 2020. URL http://arxiv.org/abs/2007.03488.

Besen, S. The Death of the Static AI Benchmark, March 2024. URL https://towardsdatascience.com/the-death-of-the-static-ai-benchmark-88b5ff437086.

Biderman, S., Schoelkopf, H., Sutawika, L., Gao, L., Tow, J., Abbasi, B., Aji, A. F., Ammanamanchi, P. S., Black, S., Clive, J., DiPofi, A., Etxaniz, J., Fattori, B., Forde, J. Z., Foster, C., Hsu, J., Jaiswal, M., Lee, W. Y., Li, H., Lovering, C., Muennighoff, N., Pavlick, E., Phang, J., Skowron, A., Tan, S., Tang, X., Wang, K. A., Winata, G. I., Yvon, F., and Zou, A. Lessons from the Trenches on Reproducible Evaluation of Language Models, May 2024. URL http://arxiv.org/abs/2405.14782.

Birhane, A., Steed, R., Ojewale, V., Vecchione, B., and Raji, I. D. AI auditing: The Broken Bus on the Road to AI Accountability, January 2024. URL http://arxiv.org/abs/2401.14462.

Blagec, K., Kraiger, J., Frühwirt, W., and Samwald, M. Benchmark datasets driving artificial intelligence development fail to capture the needs of medical professionals. *Journal of Biomedical Informatics*, 137:104274, January 2023. ISSN 15320464. doi: 10.1016/j.jbi.2022.104274. URL https://linkinghub.elsevier.com/retrieve/pii/S1532046422002799.

Blili-Hamelin, B. and Hancox-Li, L. Making Intelligence: Ethical Values in IQ and ML Benchmarks. In *2023 ACM Conference on Fairness, Accountability, and Transparency*, pp. 271–284, Chicago IL USA, June 2023. ACM. ISBN 9798400701924. doi: 10.1145/3593013.3593996. URL https://dl.acm.org/doi/10.1145/3593013.3593996.

Blodgett, S. L., Lopez, G., Olteanu, A., Sim, R., and Wallach, H. Stereotyping Norwegian Salmon: An Inventory of Pitfalls in Fairness Benchmark Datasets. In *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, 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 *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4843–4855, Online, 2021. Association for Computational Linguistics. doi:

10.18653/v1/2021.naacl-main.385. URL https://aclanthology.org/2021.naacl-main.385.

Burnell, R., Schellaert, W., Burden, J., Ullman, T. D., Martinez-Plumed, F., Tenenbaum, J. B., Rutar, D., Cheke, L. G., Sohl-Dickstein, J., Mitchell, M., Kiela, D., Shanahan, M., Voorhees, E. M., Cohn, A. G., Leibo, J. Z., and Hernandez-Orallo, J. Rethink reporting of evaluation results in AI. *Science*, 380(6641):136–138, April 2023. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.adf6369. URL https://www.science.org/doi/10.1126/science.adf6369.

Casper, S., Ezell, C., Siegmann, C., Kolt, N., Curtis, T. L., Bucknall, B., Haupt, A., Wei, K., Scheurer, J., Hobbhahn, M., Sharkey, L., Krishna, S., Hagen, M. V., Alberti, S., Chan, A., Sun, Q., Gerovitch, M., Bau, D., Tegmark, M., Krueger, D., and Hadfield-Menell, D. Black-Box Access is Insufficient for Rigorous AI Audits. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pp. 2254–2272, June 2024. doi: 10.1145/3630106.3659037. URL http://arxiv.org/abs/2401.14446.

Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., Chen, H., Yi, X., Wang, C., Wang, Y., Ye, W., Zhang, Y., Chang, Y., Yu, P. S., Yang, Q., and Xie, X. A Survey on Evaluation of Large Language Models, December 2023. URL http://arxiv.org/abs/2307.03109.

Chollet, F. On the Measure of Intelligence, November 2019. URL http://arxiv.org/abs/1911.01547.

Church, K. W. and Hestness, J. A survey of 25 years of evaluation. *Natural Language Engineering*, 25 (06):753–767, November 2019. ISSN 1351-3249, 1469-8110. doi: 10.1017/S1351324919000275. URL https://www.cambridge.org/core/product/identifier/S1351324919000275/type/journal_article.

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

Denton, E., Hanna, A., Amironesei, R., Smart, A., and Nicole, H. On the genealogy of machine learning datasets: A critical history of ImageNet. *Big Data & Society*, 8(2):1–14, July 2021. ISSN 2053-9517, 2053-9517. doi: 10.1177/20539517211035955. URL http://journals.sagepub.com/doi/10.1177/20539517211035955.

Denton, R., Hanna, A., Amironesei, R., Smart, A., Nicole, H., and Scheuerman, M. K. Bringing the People Back In: Contesting Benchmark Machine Learning Datasets, July 2020. URL http://arxiv.org/abs/2007.07399.

Diberardino, N., Baleshta, C., and Stark, L. Algorithmic Harms and Algorithmic Wrongs. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1725–1732, Rio de Janeiro Brazil, June 2024. ACM. ISBN 9798400704505. doi: 10.1145/3630106.3659001. URL https://dl.acm.org/doi/10.1145/3630106.3659001.

Engdahl, I. Agreements 'in the wild': Standards and alignment in machine learning benchmark dataset construction. *Big Data & Society*, 11 (2):20539517241242457, June 2024. ISSN 2053-9517, 2053-9517. doi: 10.1177/20539517241242457. URL https://journals.sagepub.com/doi/10.1177/20539517241242457.

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, 2025. URL https://arxiv.org/abs/2502.06559.

Ethayarajh, K. and Jurafsky, D. Utility is in the Eye of the User: A Critique of NLP Leaderboards, March 2021. URL http://arxiv.org/abs/2009.13888.

European Union. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (Artificial Intelligence Act), 2024.

Frieder, S., Bayer, J., Collins, K. M., Berner, J., Loader, J., Juhász, A., Ruehle, F., Welleck, S., Poesia, G., Griffiths, R.-R., Weller, A., Goyal, A., Lukasiewicz, T., and Gowers, T. Data for Mathematical Copilots: Better Ways of Presenting Proofs for Machine Learning, December 2024. URL http://arxiv.org/abs/2412.15184.

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., III, H. D., and Crawford, K. Datasheets for Datasets, December 2021. URL http://arxiv.org/abs/1803.09010.

Gehrmann, S., Clark, E., and Sellam, T. Repairing the Cracked Foundation: A Survey of Obstacles in Evaluation Practices for Generated Text. *Journal of Artificial Intelligence Research*, 77:103–166, May 2023. ISSN 1076-9757. doi: 10.1613/jair.1.13715. URL https://www.jair.org/index.php/jair/article/view/13715.

Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., and Wichmann, F. A. Shortcut

Learning in Deep Neural Networks. *Nature Machine Intelligence*, 2(11):665–673, November 2020. ISSN 2522-5839. doi: 10.1038/s42256-020-00257-z. URL http://arxiv.org/abs/2004.07780.

Gomez, E., Lorenzo, P., Frau Amar, P., and Vinagre, J. Diversity in artificial intelligence conferences. Publications Office of the European Union JRC137550, Publications Office, 2024. URL https://data.europa.eu/doi/10.2760/796551.

Greenblatt, R., Denison, C., Wright, B., Roger, F., MacDiarmid, M., Marks, S., Treutlein, J., Belonax, T., Chen, J., Duvenaud, D., Khan, A., Michael, J., Mindermann, S., Perez, E., Petrini, L., Uesato, J., Kaplan, J., Shlegeris, B., Bowman, S. R., and Hubinger, E. Alignment faking in large language models, December 2024. URL http://arxiv.org/abs/2412.14093.

Grill, G. Constructing Capabilities: The Politics of Testing Infrastructures for Generative AI. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1838–1849, Rio de Janeiro Brazil, June 2024. ACM. ISBN 9798400704505. doi: 10.1145/3630106.3659009. URL https://dl.acm.org/doi/10.1145/3630106.3659009.

Guldimann, P., Spiridonov, A., Staab, R., Jovanović, N., Vero, M., Vechev, V., Gueorguieva, A., Balunović, M., Konstantinov, N., Bielik, P., Tsankov, P., and Vechev, M. COMPL-AI Framework: A Technical Interpretation and LLM Benchmarking Suite for the EU Artificial Intelligence Act, October 2024. URL http://arxiv.org/abs/2410.07959.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J. Measuring Massive Multitask Language Understanding, January 2021. URL http://arxiv.org/abs/2009.03300.

Henning, J. Spec cpu2000: measuring cpu performance in the new millennium. *Computer*, 33(7):28–35, 2000. doi: 10.1109/2.869367.

Hooker, S. On the Limitations of Compute Thresholds as a Governance Strategy, July 2024. URL http://arxiv.org/abs/2407.05694.

Hutchinson, B., Rostamzadeh, N., Greer, C., Heller, K., and Prabhakaran, V. Evaluation Gaps in Machine Learning Practice. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1859–1876, Seoul Republic of Korea, June 2022. ACM. ISBN 978-1-4503-9352-2. doi: 10.1145/3531146.3533233. URL https://dl.acm.org/doi/10.1145/3531146.3533233.

Jalali, S. and Wohlin, C. Systematic literature studies: database searches vs. backward snowballing. In *Proceedings of the ACM-IEEE International Symposium on Empirical Software Engineering and Measurement*, ESEM '12, pp. 29–38, New York, NY, USA, 2012. Association for Computing Machinery. ISBN 9781450310567. doi: 10.1145/2372251.2372257. URL https://doi.org/10.1145/2372251.2372257.

Jannach, D. and Bauer, C. Escaping the McNamara Fallacy: Toward More Impactful Recommender Systems Research. *AI Magazine*, 41(4):79–95, December 2020. ISSN 0738-4602, 2371-9621. doi: 10.1609/aimag.v41i4.5312. URL https://onlinelibrary.wiley.com/doi/10.1609/aimag.v41i4.5312.

Jaton, F. *The Constitution of Algorithms: Ground-Truthing, Programming, Formulating*. Inside Technology. The MIT Press, Cambridge, 2021. ISBN 978-0-262-54214-2 978-0-262-36323-5.

Jones, E., Hardalupas, M., and Agrew, W. Under the radar? examining the evaluation of foundation models. Report, Ada Lovelace Institute, 2024. URL https://www.adalovelaceinstitute.org/report/under-the-radar/.

Kang, E. B. Ground truth tracings (GTT): On the epistemic limits of machine learning. *Big Data & Society*, 10(1):20539517221146122, January 2023. ISSN 2053-9517, 2053-9517. doi: 10.1177/20539517221146122. URL https://journals.sagepub.com/doi/10.1177/20539517221146122.

Kapoor, S., Stroebl, B., Siegel, Z. S., Nadgir, N., and Narayanan, A. AI Agents That Matter, July 2024. URL http://arxiv.org/abs/2407.01502.

Kaufman, S., Rosset, S., Perlich, C., and Stitelman, O. Leakage in data mining: Formulation, detection, and avoidance. *ACM Trans. Knowl. Discov. Data*, 6(4), December 2012. ISSN 1556-4681. doi: 10.1145/2382577.2382579. URL https://doi.org/10.1145/2382577.2382579.

Keegan, J. Everyone Is Judging AI by These Tests. But Experts Say They're Close to Meaningless. *The Markup*, July 2024. URL https://themarkup.org/artificial-intelligence/2024/07/17/everyone-is-judging-ai-by-these-tests-but-experts-say-theyre-close-to-meaningless.

Kejriwal, M., Santos, H., Shen, K., Mulvehill, A. M., and McGuinness, D. L. A noise audit of human-labeled benchmarks for machine commonsense reasoning. *Scientific Reports*, 14(1):8609, April 2024.

ISSN 2045-2322. doi: 10.1038/s41598-024-58937-4. URL https://www.nature.com/articles/s41598-024-58937-4.

Koch, B., Denton, E., Hanna, A., and Foster, J. G. Reduced, Reused and Recycled: The Life of a Dataset in Machine Learning Research, December 2021. URL http://arxiv.org/abs/2112.01716.

Koch, B. J. and Peterson, D. From Protoscience to Epistemic Monoculture: How Benchmarking Set the Stage for the Deep Learning Revolution, April 2024. URL http://arxiv.org/abs/2404.06647.

Leech, G., Vazquez, J. J., Kupper, N., Yagudin, M., and Aitchison, L. Questionable practices in machine learning, July 2024. URL https://arxiv.org/abs/2407.12220v2.

Lewis, P., Stenetorp, P., and Riedel, S. Question and answer test-train overlap in open-domain question answering datasets. In Merlo, P., Tiedemann, J., and Tsarfaty, R. (eds.), *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 1000–1008, Online, April 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.86. URL https://aclanthology.org/2021.eacl-main.86.

Liao, Q. V. and Xiao, Z. Rethinking Model Evaluation as Narrowing the Socio-Technical Gap, June 2023. URL http://arxiv.org/abs/2306.03100.

Liao, T., Taori, R., Raji, I. D., and Schmidt, L. Are we learning yet? a meta review of evaluation failures across machine learning. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. URL https://openreview.net/forum?id=mPducS1MsEK.

Liu, P., Fu, J., Xiao, Y., Yuan, W., Chang, S., Dai, J., Liu, Y., Ye, Z., Dou, Z.-Y., and Neubig, G. ExplainaBoard: An Explainable Leaderboard for NLP, July 2021. URL http://arxiv.org/abs/2104.06387.

Luitse, D., Blanke, T., and Poell, T. AI competitions as infrastructures of power in medical imaging. *Information, Communication & Society*, pp. 1–22, March 2024. ISSN 1369-118X, 1468-4462. doi: 10.1080/1369118X.2024.2334393. URL https://www.tandfonline.com/doi/full/10.1080/1369118X.2024.2334393.

Lum, K., Anthis, J. R., Nagpal, C., and D'Amour, A. Bias in Language Models: Beyond Trick Tests and Toward RUTEd Evaluation, February 2024. URL https://arxiv.org/abs/2402.12649v1.

Magar, I. and Schwartz, R. Data contamination: From memorization to exploitation. In Muresan, S., Nakov, P., and Villavicencio, A. (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 157–165, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-short.18. URL https://aclanthology.org/2022.acl-short.18.

Malevé, N. Practices of Benchmarking: Vulnerability in the Computer Vision Pipeline. *photographies*, 16 (2):173–189, May 2023. ISSN 1754-0763, 1754-0771. doi: 10.1080/17540763.2023.2189159. URL https://www.tandfonline.com/doi/full/10.1080/17540763.2023.2189159.

Marres, N. and Stark, D. Put to the test: For a new sociology of testing. *The British Journal of Sociology*, 71(3):423–443, June 2020. ISSN 0007-1315, 1468-4446. doi: 10.1111/1468-4446.12746. URL https://onlinelibrary.wiley.com/doi/10.1111/1468-4446.12746.

McIntosh, T. R., Susnjak, T., Arachchilage, N., Liu, T., Watters, P., and Halgamuge, M. N. Inadequacies of Large Language Model Benchmarks in the Era of Generative Artificial Intelligence, October 2024. URL http://arxiv.org/abs/2402.09880.

Meinke, A., Schoen, B., Scheurer, J., Balesni, M., Shah, R., and Hobbhahn, M. Frontier Models are Capable of In-context Scheming, December 2024. URL https://arxiv.org/abs/2412.04984v1.

Michael, J., Holtzman, A., Parrish, A., Mueller, A., Wang, A., Chen, A., Madaan, D., Nangia, N., Pang, R. Y., Phang, J., and Bowman, S. R. What Do NLP Researchers Believe? Results of the NLP Community Metasurvey, August 2022. URL http://arxiv.org/abs/2208.12852.

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., and Gebru, T. Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 220–229, January 2019. doi: 10.1145/3287560.3287596. URL http://arxiv.org/abs/1810.03993.

Mizrahi, M., Kaplan, G., Malkin, D., Dror, R., Shahaf, D., and Stanovsky, G. State of What Art? A Call for Multi-Prompt LLM Evaluation, May 2024. URL http://arxiv.org/abs/2401.00595.

Mulvin, D. *Proxies: The Cultural Work of Standing In*. Infrastructures. The MIT Press, Cambridge, 2021. ISBN 978-0-262-04514-8 978-0-262-36624-3.

Narayanan, A. and Kapoor, S. GPT-4 and professional benchmarks: the wrong answer to the wrong question, March 2023a. URL https://www.aisnakeoil.com/p/gpt-4-and-professional-benchmarks.

Narayanan, A. and Kapoor, S. Evaluating LLMs is a minefield, 2023b. URL https://www.cs.princeton.edu/~arvindn/talks/evaluating_llms_minefield/.

Nasr, M., Carlini, N., Hayase, J., Jagielski, M., Cooper, A. F., Ippolito, D., Choquette-Choo, C. A., Wallace, E., and Lee, K. Extracting Training Data from ChatGPT, November 2023a. URL https://not-just-memorization.github.io/extracting-training-data-from-chatgpt.html?ref=404media.co.

Nasr, M., Carlini, N., Hayase, J., Jagielski, M., Cooper, A. F., Ippolito, D., Choquette-Choo, C. A., Wallace, E., Tramèr, F., and Lee, K. Scalable Extraction of Training Data from (Production) Language Models, November 2023b. URL http://arxiv.org/abs/2311.17035.

Oakden-Rayner, L., Dunnmon, J., Carneiro, G., and Ré, C. Hidden Stratification Causes Clinically Meaningful Failures in Machine Learning for Medical Imaging, November 2019. URL http://arxiv.org/abs/1909.12475.

Ojewale, V., Steed, R., Vecchione, B., Birhane, A., and Raji, I. D. Towards AI Accountability Infrastructure: Gaps and Opportunities in AI Audit Tooling, March 2024. URL http://arxiv.org/abs/2402.17861.

Orr, W. and Crawford, K. The social construction of datasets: On the practices, processes, and challenges of dataset creation for machine learning. *New Media & Society*, 26 (9):4955–4972, 2024a.

Orr, W. and Crawford, K. Building Better Datasets: Seven Recommendations for Responsible Design from Dataset Creators, August 2024b. URL http://arxiv.org/abs/2409.00252.

Orr, W. and Kang, E. B. AI as a Sport: On the Competitive Epistemologies of Benchmarking. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1875–1884, Rio de Janeiro Brazil, June 2024. ACM. ISBN 9798400704505. doi: 10.1145/3630106.3659012. URL https://dl.acm.org/doi/10.1145/3630106.3659012.

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, November 2022. ISSN 2041-1723. doi: 10.1038/s41467-022-34591-0. URL https://www.nature.com/articles/s41467-022-34591-0.

Pacchiardi, L., Tesic, M., Cheke, L. G., and Hernández-Orallo, J. Leaving the barn door open for Clever Hans: Simple features predict LLM benchmark answers, October 2024. URL http://arxiv.org/abs/2410.11672.

Park, J. and Jeoung, S. Raison d'être of the benchmark dataset: A Survey of Current Practices of Benchmark Dataset Sharing Platforms. In *Proceedings of NLP Power! The First Workshop on Efficient Benchmarking in NLP*, pp. 1–10, Dublin, Ireland, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.nlppower-1.1. URL https://aclanthology.org/2022.nlppower-1.1.

Paullada, A., Raji, I. D., Bender, E. M., Denton, E., and Hanna, A. Data and its (dis)contents: A survey of dataset development and use in machine learning research. *Patterns*, 2(11):100336, November 2021. ISSN 26663899. doi: 10.1016/j.patter.2021.100336. URL http://arxiv.org/abs/2012.05345.

Pinch, T. "Testing - One, Two, Three ... Testing!": Toward a Sociology of Testing. *Science, Technology, & Human Values*, 18(1):25–41, January 1993. ISSN 0162-2439, 1552-8251. doi: 10.1177/016224399301800103. URL https://journals.sagepub.com/doi/10.1177/016224399301800103.

Poelman, W. and Lhoneux, M. d. The Roles of English in Evaluating Multilingual Language Models, December 2024. URL http://arxiv.org/abs/2412.08392.

Qi, X., Zeng, Y., Xie, T., Chen, P.-Y., Jia, R., Mittal, P., and Henderson, P. Fine-tuning Aligned Language Models Compromises Safety, Even When Users Do Not Intend To!, October 2023. URL http://arxiv.org/abs/2310.03693.

Raji, I. D., Denton, E., Bender, E. M., Hanna, A., and Paullada, A. AI and the everything in the whole wide world benchmark. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. URL https://openreview.net/forum?id=j6NxpQbREA1.

Rauh, M., Marchal, N., Manzini, A., Hendricks, L. A., Comanescu, R., Akbulut, C., Stepleton, T., Mateos-Garcia, J., Bergman, S., Kay, J., Griffin, C., Bariach, B., Gabriel, I., Rieser, V., Isaac, W., and Weidinger, L. Gaps in the Safety Evaluation of Generative AI. *Proceedings* *of the AAAI/ACM Conference on AI, Ethics, and Society*, 7:1200–1217, October 2024. ISSN 3065-8365. doi: 10.1609/aies.v7i1.31717. URL https://ojs.aaai.org/index.php/AIES/article/view/31717.

Ren, R., Basart, S., Khoja, A., Gatti, A., Phan, L., Yin, X., Mazeika, M., Pan, A., Mukobi, G., Kim, R. H., Fitz, S., and Hendrycks, D. Safetywashing: Do AI safety benchmarks actually measure safety progress? In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. URL https://openreview.net/forum?id=YagfTP3RK6.

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.

Roberts, M., Thakur, H., Herlihy, C., White, C., and Dooley, S. Data Contamination Through the Lens of Time, October 2023. URL http://arxiv.org/abs/2310.10628.

Rodriguez, P., Barrow, J., Hoyle, A. M., Lalor, J. P., Jia, R., and Boyd-Graber, J. Evaluation Examples are not Equally Informative: How should that change NLP Leaderboards? In *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. 4486–4503, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.346. URL https://aclanthology.org/2021.acl-long.346.

Röttger, P., Pernisi, F., Vidgen, B., and Hovy, D. SafetyPrompts: a Systematic Review of Open Datasets for Evaluating and Improving Large Language Model Safety, April 2024. URL http://arxiv.org/abs/2404.05399.

Sambasivan, N., Kapania, S., Highfill, H., Akrong, D., Paritosh, P., and Aroyo, L. M. "Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pp. 1–15, Yokohama Japan, May 2021. ACM. ISBN 978-1-4503-8096-6. doi: 10.1145/3411764.3445518. URL https://dl.acm.org/doi/10.1145/3411764.3445518.

Scheuerman, M. K., Hanna, A., and Denton, E. Do Datasets Have Politics? Disciplinary Values in Computer Vision Dataset Development. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–37, October 2021. ISSN 2573-0142. doi: 10.1145/3476058. URL https://dl.acm.org/doi/10.1145/3476058.

Schlangen, D. Targeting the Benchmark: On Methodology in Current Natural Language Processing Research, July 2020. URL http://arxiv.org/abs/2007.04792.

Sculley, D., Snoek, J., Rahimi, A., and Wiltschko, A. Winner's Curse? On Pace, Progress, and Empirical Rigor. Vancouver, BC, Canada, 2018. URL https://openreview.net/pdf?id=rJWF0Fywf.

Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., and Vertesi, J. Fairness and Abstraction in Sociotechnical Systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 59–68, Atlanta GA USA, January 2019. ACM. ISBN 978-1-4503-6125-5. doi: 10.1145/3287560.3287598. URL https://dl.acm.org/doi/10.1145/3287560.3287598.

Sen, S., Giesel, M. E., Gold, R., Hillmann, B., Lesicko, M., Naden, S., Russell, J., Wang, Z. K., and Hecht, B. Turkers, Scholars, "Arafat" and "Peace": Cultural Communities and Algorithmic Gold Standards. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, pp. 826–838, Vancouver BC Canada, February 2015. ACM. ISBN 978-1-4503-2922-4. doi: 10.1145/2675133.2675285. URL https://dl.acm.org/doi/10.1145/2675133.2675285.

Simson, J., Fabris, A., and Kern, C. Lazy Data Practices Harm Fairness Research. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pp. 642–659, Rio de Janeiro Brazil, June 2024. ACM. ISBN 9798400704505. doi: 10.1145/3630106.3658931. URL https://dl.acm.org/doi/10.1145/3630106.3658931.

Smith, J. J., Amershi, S., Barocas, S., Wallach, H., and Vaughan, J. W. REAL ML: Recognizing, Exploring, and Articulating Limitations of Machine Learning Research. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 587–597, June 2022. doi: 10.1145/3531146.3533122. URL http://arxiv.org/abs/2205.08363.

Stengers, I. *Another science is possible: a manifesto for slow science*. Polity press, Cambridge, 2018. ISBN 978-1-5095-2180-7.

Stewart, M. The olympics of ai: Benchmarking machine learning systems, 2023. URL https://towardsdatascience.com/the-olympics-of-ai-benchmarking-machine-learning-systems-c4b2051fbd2b.

Strathern, M. 'Improving ratings': audit in the British University system. *European Review*, 5(3):305–321, July 1997. ISSN 10627987, 1234981X. doi: 10.1002/(SICI)1234-981X(199707)5:3⟨305::AID-EURO184⟩3.0.CO;2-4. URL https://www.cambridge.org/core/product/identifier/S1062798700002660/type/journal_article.

Subramonian, A., Yuan, X., III, H. D., and Blodgett, S. L. It Takes Two to Tango: Navigating Conceptualizations of NLP Tasks and Measurements of Performance, May 2023. URL http://arxiv.org/abs/2305.09022.

The White House. Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence, 2023.

Thylstrup, N. B., Hansen, K. B., Flyverbom, M., and Amoore, L. Politics of data reuse in machine learning systems: Theorizing reuse entanglements. *Big Data & Society*, 9(2):20539517221139785, July 2022. ISSN 2053-9517, 2053-9517. doi: 10.1177/20539517221139785. URL https://journals.sagepub.com/doi/10.1177/20539517221139785.

Tirumala, K., Markosyan, A., Zettlemoyer, L., and Aghajanyan, A. Memorization Without Overfitting: Analyzing the Training Dynamics of Large Language Models. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A. (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 38274–38290. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/fa0509f4dab6807e2cb465715bf2d249-Paper-Conference.pdf.

Tsipras, D., Santurkar, S., Engstrom, L., Ilyas, A., and Madry, A. From ImageNet to Image Classification: Contextualizing Progress on Benchmarks, May 2020. URL http://arxiv.org/abs/2005.11295.

US Department of Commerce. Framework for Artificial Intelligence Diffusion, 2025.

Vafa, K., Chen, J. Y., Rambachan, A., Kleinberg, J., and Mullainathan, S. Evaluating the World Model Implicit in a Generative Model, November 2024. URL http://arxiv.org/abs/2406.03689.

Weidinger, L., Rauh, M., Marchal, N., Manzini, A., Hendricks, L. A., Mateos-Garcia, J., Bergman, S., Kay, J., Griffin, C., Bariach, B., Gabriel, I., Rieser, V., and Isaac, W. Sociotechnical Safety Evaluation of Generative AI Systems, October 2023. URL http://arxiv.org/abs/2310.11986.

Weij, T. v. d., Hofstätter, F., Jaffe, O., Brown, S. F., and Ward, F. R. AI Sandbagging: Language Models can Strategically Underperform on Evaluations, June 2024. URL http://arxiv.org/abs/2406.07358.

Xu, C., Guan, S., Greene, D., and Kechadi, M.-T. Benchmark Data Contamination of Large Language Models: A Survey, June 2024a. URL http://arxiv.org/abs/2406.04244.

Xu, R., Wang, Z., Fan, R.-Z., and Liu, P. Benchmarking Benchmark Leakage in Large Language Models, April 2024b. URL http://arxiv.org/abs/2404.18824.

Yang, S., Chiang, W.-L., Zheng, L., Gonzalez, J. E., and Stoica, I. Rethinking Benchmark and Contamination for Language Models with Rephrased Samples, November 2023. URL http://arxiv.org/abs/2311.04850.

Yuan, L., Chen, Y., Cui, G., Gao, H., Zou, F., Cheng, X., Ji, H., Liu, Z., and Sun, M. Revisiting Out-of-distribution Robustness in NLP: Benchmark, Analysis, and LLMs Evaluations. *37th Conference on Neural Information Processing Systems (NeurIPS 2023) Track on Datasets and Benchmarks*, 2023.

Zhang, A. K., Klyman, K., Mai, Y., Levine, Y., Zhang, Y., Bommasani, R., and Liang, P. Language model developers should report train-test overlap, October 2024. URL http://arxiv.org/abs/2410.08385.

Zhijia, L. Top llms in china and the u.s. only 5 months apart: Kai-fu lee, 2024. URL https://en.tmtpost.com/post/7289212.

A. Appendix. Summary of AI benchmarking issues

Table 1. Distribution of AI benchmarking shortcomings and related references, as they appear in the original text (Eriksson et al., 2025).

CATEGORY	REFERENCES
PROBLEMS WITH DATA COLLECTION, ANNOTATION, AND DOCUMENTATION	(GEBRU ET AL., 2021; MITCHELL ET AL., 2019; ORR & CRAWFORD, 2024B; SIMSON ET AL., 2024; SCHEUERMAN ET AL., 2021; REUEL ET AL., 2024; DENTON ET AL., 2020; ARZT & HANBURY, 2024; ORR & CRAWFORD, 2024A; SAMBASIVAN ET AL., 2021; KOCH ET AL., 2021; THYLSTRUP ET AL., 2022; PARK & JEOUNG, 2022; PAULLADA ET AL., 2021; KEEGAN, 2024; GRILL, 2024; TSIPRAS ET AL., 2020; AROYO & WELTY, 2015; SEN ET AL., 2015; RAUH ET AL., 2024; CHOLLET, 2019; LIAO ET AL., 2021; GEIRHOS ET AL., 2020; OAKDEN-RAYNER ET AL., 2019; PACCHIARDI ET AL., 2024; VAFA ET AL., 2024; KEJRIWAL ET AL., 2024)
WEAK CONSTRUCT VALIDITY AND EPISTEMOLOGICAL CLAIMS	(RAJI ET AL., 2021; BLODGETT ET AL., 2021; BARTZ-BEIELSTEIN ET AL., 2020; SUBRAMONIAN ET AL., 2023; RÖTTGER ET AL., 2024; NARAYANAN & KAPOOR, 2023A; SELBST ET AL., 2019; DIBERARDINO ET AL., 2024; KEEGAN, 2024; REN ET AL., 2024; LEECH ET AL., 2024)
SOCIOCULTURAL CONTEXT AND GAP	(ORR & KANG, 2024; ENGDAHL, 2024; MICHAEL ET AL., 2022; SCHEUERMAN ET AL., 2021; ORR & CRAWFORD, 2024A; SAMBASIVAN ET AL., 2021; PAULLADA ET AL., 2021; AHMED ET AL., 2024; HUTCHINSON ET AL., 2022; LIAO & XIAO, 2023; BLAGEC ET AL., 2023; JANNACH & BAUER, 2020; ETHAYARAJH & JURAFSKY, 2021; FRIEDER ET AL., 2024)
NARROW BENCHMARK DIVERSITY AND SCOPE	(GOMEZ ET AL., 2024; RAUH ET AL., 2024; WEIDINGER ET AL., 2023; RÖTTGER ET AL., 2024; GULDIMANN ET AL., 2024; KOCH ET AL., 2021; MCINTOSH ET AL., 2024; SELBST ET AL., 2019; SIMSON ET AL., 2024; MIZRAHI ET AL., 2024; OJEWALE ET AL., 2024; CHANG ET AL., 2023; REUEL ET AL., 2024; BIRHANE ET AL., 2024; GEHRMANN ET AL., 2023; POELMAN & LHONEUX, 2024; BURNELL ET AL., 2023; KAPOOR ET AL., 2024; LUM ET AL., 2024)
ECONOMIC, COMPETITIVE, AND COMMERCIAL ROOTS	(ORR & KANG, 2024; GRILL, 2024; ZHIJIA, 2024; REN ET AL., 2024; ETHAYARAJH & JURAFSKY, 2021; GEHRMANN ET AL., 2023; CHURCH & HESTNESS, 2019; SMITH ET AL., 2022; LUITSE ET AL., 2024; KOCH ET AL., 2021; STEWART, 2023; MALEVÉ, 2023; STENGERS, 2018; SCULLEY ET AL., 2018; AHMED ET AL., 2023; KOCH & PETERSON, 2024)
RIGGING, GAMING, AND MEASURE BECOMING TARGET	(DEHGHANI ET AL., 2021; ALZAHRANI ET AL., 2024; GREENBLATT ET AL., 2024; STRATHERN, 1997; BARTZ-BEIELSTEIN ET AL., 2020; BIDERMAN ET AL., 2024; REUEL ET AL., 2024; XU ET AL., 2024A; ZHANG ET AL., 2024; BESEN, 2024; MAGAR & SCHWARTZ, 2022; ROBERTS ET AL., 2023; TIRUMALA ET AL., 2022; LEWIS ET AL., 2021; KAUFMAN ET AL., 2012; YUAN ET AL., 2023; NARAYANAN & KAPOOR, 2023A; WEIJ ET AL., 2024; CASPER ET AL., 2024; MEINKE ET AL., 2024; YANG ET AL., 2023; XU ET AL., 2024B)
DUBIOUS COMMUNITY VETTING AND PATH DEPENDENCIES	(ORR & KANG, 2024; ORR & CRAWFORD, 2024A; DENTON ET AL., 2021; MULVIN, 2021; SCHLANGEN, 2020; KOCH ET AL., 2021; JATON, 2021; OTT ET AL., 2022; DEHGHANI ET AL., 2021; MCINTOSH ET AL., 2024; BAO ET AL., 2022; BLILI-HAMELIN & HANCOX-LI, 2023; KOCH & PETERSON, 2024)
RAPID AI DEVELOPMENT AND BENCHMARK SATURATION	(KEEGAN, 2024; BIDERMAN ET AL., 2024; HENDRYCKS ET AL., 2021; BOWMAN & DAHL, 2021; OTT ET AL., 2022; MCINTOSH ET AL., 2024; EUROPEAN UNION, 2024; THE WHITE HOUSE, 2023; US DEPARTMENT OF COMMERCE, 2025; HOOKER, 2024)
AI COMPLEXITY AND UNKNOWN UNKNOWNS	(MCINTOSH ET AL., 2024; NASR ET AL., 2023A;B; QI ET AL., 2023)