Can We Trust AI Benchmarks? An Interdisciplinary Review of Current Issues in AI Evaluation

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

[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
{maria.eriksson, erasmo.purificato, arman.noroozian, joao.vinagre, guillaume.chaslot, emilia.gomez-gutierrez, david.fernandez-llorca}@ec.europa.eu

Abstract

Quantitative Artificial Intelligence (AI) Benchmarks have emerged as fundamental tools for evaluating the performance, capability, and safety of AI models and systems. Currently, they shape the direction of AI development and are playing an increasingly prominent role in regulatory frameworks. As their influence grows, however, so too does concerns about *how* and with what effects they evaluate highly sensitive topics such as capabilities, including high-impact capabilities, safety and systemic risks. This paper presents an interdisciplinary meta-review of about 110 studies that discuss shortcomings in quantitative benchmarking practices, published in the last 10 years. It brings together many fine-grained issues in the design and application of benchmarks (such as biases in dataset creation, inadequate documentation, data contamination, and failures to distinguish signal from noise) with broader sociotechnical issues (such as an over-focus on evaluating text-based AI models according to one-time testing logic that fails to account for how AI models are increasingly multimodal and interact with humans and other technical systems). Our review also highlights a series of systemic flaws in current benchmarking practices, such as misaligned incentives, construct validity issues, unknown unknowns, and problems with the gaming of benchmark results. Furthermore, it underscores how benchmark practices are fundamentally shaped by cultural, commercial and competitive dynamics that often prioritise state-of-the-art performance at the expense of broader societal concerns. By providing an overview of risks associated with existing benchmarking procedures, we problematise disproportionate trust placed in benchmarks and contribute to ongoing efforts to improve the accountability and relevance of quantitative AI benchmarks within the complexities of real-world scenarios.

Introduction

Quantitative artificial intelligence (AI) benchmarks (i.e., combinations of test datasets and performance metrics that are taken to represent general or specific tasks and used to compare AI model capabilities and/or risks (Raji et al. 2021)) play a central role in the release and marketing of newly developed AI models. Together with qualitative evaluation methods (such as red teaming and peer confrontations (Chollet 2019)), quantitative benchmarks - hereafter referred to as *AI benchmarks*, or simply *benchmarks* - are generally seen as providing crucial feedback signals on the performance and capabilities of AI. Indeed, they have become so critical to AI development that businesses go to great lengths to achieve good benchmarking scores, with market players like OpenAI being estimated to have spent hundreds of thousands of dollars on compute to obtain a high score at the ARC-AGI benchmark (Pfister and Jud 2025).

Increasingly, AI benchmarks are also used in regulatory contexts, where the goal is to assess potential societal harms posed by AI models and systems. The most notable case is the EU AI Act (European Union 2024b), which incorporates benchmarks in several key provisions such as high-risk assessments of AI systems, where they are expected to inform requirements on accuracy, robustness, and cybersecurity (Art. 15(2)), facilitated through AI regulatory sandboxes (Art. 58 (2)). They will also be of fundamental importance for classifying general-purpose AI (GPAI) models with systemic risks and assessing high-impact capabilities (Art. 51(1) and Annex XIII). Additionally, the first and second drafts of the Code of Practice for the EU AI Act mentions benchmarks as a risk assessment method for providers of GPAI models with systemic risks (European Union 2024a,c). In the US, benchmarks are also relevant in the recently revoked AI Executive Order (The White House 2023) and the AI Diffusion Framework (US Department of Commerce 2025). AI benchmarks can further be expected to play a central role in the implementation of legislations such as the EU Digital Services Act (DSA) (European Union 2022) and the UK's Online Safety Act (UK Parliament 2023), which require the largest online platforms and search engines - entities that increasingly rely on AI to curate, filter, and/or rank content to millions of users - to perform regular algorithmic audits. In short, AI benchmarks - which constitute a highly heterogenous and far from standardised set of techniques - are increasingly at the heart of policy efforts to make AI more transparent and secure.

At the same time as benchmarks are increasingly relied upon to provide AI safety assurances, however, journalists (Keegan 2024; Roose 2024; Heaven 2023) and researchers in a broad range of academic fields have raised serious concerns regarding their current use. This includes critical voices being raised in fields ranging from cyberse-

Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

curity (McIntosh et al. 2024), linguistics (Bowman and Dahl 2021), and computer science (Gema et al. 2024), to sociology (Engdahl 2024), economics (Ethayarajh and Jurafsky 2021), philosophy (LaCroix and Luccioni 2022), ethnography (Keyes and Austin 2022), and science and technology studies (Keyes and Austin 2022). Such scholars have for example described current AI evaluation practices as a "minefield" (Narayanan and Kapoor 2023a), that raise serious ethical concerns regarding what should be measured, according to what standards, and with what downstream effects (Blili-Hamelin and Hancox-Li 2023; Mitchell 2023; Burden 2024). They have also emphasised that benchmarks are deeply political, performative, and generative in the sense that they do not passively describe and measure how things are in the world, but actively take part in shaping it (Grill 2024). This happens as benchmarks continuously influence how AI models are trained, fine-tuned, and applied - practices with wide-ranging political, economic, and cultural effects.

With this backdrop, an interdisciplinary and up-to-date survey of quantitative AI benchmarking critique is momentously missing. The aim of this article is to address this gap by mapping known limitations in quantitative AI evaluation practices, drawing together insights from multiple academic fields. Our goal is to provide an overview of current risks associated with quantitative AI tests, targeting policy makers, stakeholders performing algorithmic audits, and developers of AI models and benchmarks. To achieve this, we gather and analyse around 110 publications published between 1st January 2014 and 31st December 2024 that explicitly and primarily address limitations in current benchmarking practices. Notably, more than half of these publications were published in 2023 or later, highlighting the urgency of providing an updated survey of discussions in the field.

Our findings show that a rapidly growing number of researchers are voicing concerns regarding how benchmarks are used to define and measure what is safe or unsafe, moral or immoral, true or false, toxic or healthy. They further support the notion that no benchmark is neutral and that AI tests and benchmarking practices always rest "on interwoven technical and normative decisions" (Rauh et al. 2024, p. 1201) in ways that urge policy makers and AI developers to apply them with caution. In particular, previous research highlights a need to question the relevance and trustworthiness of well-cited benchmarks, since existing studies have repeatedly shown weaknesses in benchmarks perceived as state-of-the-art (SOTA). Our findings also reveal a substantial need to scrutinise capability and safety-oriented AI benchmarks to the same extent as the AI models they are meant to evaluate. In short, AI benchmarks need to be subjected to the same demands concerning transparency, fairness, and explainability, as algorithmic systems and AI models writ large.

In what follows, we first provide a background discussion on AI benchmarking practices, including clarifications on the terminology used in this paper. Next, we situate our work within a series of previous research that summarise benchmark critique, before providing a summary of nine issues with quantitative benchmarks, identified during the course of our research. While not exhaustive, we present these issues as a taxonomy of benchmark critique that highlights crucial and often interrelated points of concern regarding benchmarks, voiced in the past decade. Finally, we provide some concluding remarks on the implications these issues have for policy makers. On the whole, our work constitutes the first step in a broader research and policy-oriented project aimed at developing a framework for trustworthy AI benchmarks, and will proceed by considering mitigation strategies.

Background

Etymologically, the term "benchmark" has its roots in land surveying, where a physical mark (known as a 'bench mark') was used as a reference point for measuring elevations. This mark typically consisted of a horizontal groove in a surface, which supported a level surface or 'bench' for a levelling rod. Over time, the term evolved to encompass a broader meaning, referring to any standard or reference point used for comparison or evaluation (Oxford English Dictionary 2017). Benchmarking is currently applied across many different domains such as bioinformatics (Aniba, Poch, and Thompson 2010), environmental quality (Henning 2000), security (Noroozian 2020), information retrieval (Thakur et al. 2021), transistor hardware (Cheng et al. 2020), industry and business (Camp 1989), and within the public sector (Bruno 2014), where it often refers to procedures for comparing the performance or best practices of different companies or processes. Here, we focus on the use of benchmark tests within computing, where they are used to evaluate the performance of hardware or software systems by comparing them to a standard or reference point (Henning 2000). More specifically, we zoom in on AI development, where benchmarks often facilitate cross-model comparisons, track model progress, and identify weaknesses (Reuel et al. 2024).

Benchmarks are often perceived as comparably cost- and time-effective tools for AI providers who may run them regularly throughout the model development to obtain signals of model capabilities (Weidinger et al. 2023). They can be applied to both software and hardware solutions, where the latter, for instance, evaluates the performance of CPUs, GPUs, or TPUs (Mattson et al. 2020b), and/or hardware accelerators (Mattson et al. 2020a). We address software-oriented benchmarks. Based on the definition proposed by Raji et al. (2021), we define such benchmarks as "a particular combination of a set of test datasets, including human-in-the-loop interactions, and associated performance metrics, conceptualised as representing one or more specific tasks or related capabilities, typically chosen by a community of researchers as a shared framework for comparing AI models". A *testing dataset* would refer to "a separate dataset, distinct from the training data, used to objectively evaluate the performance and generalisability of an AI model". These datasets are typically composed of *samples* or *instances* that include an input paired with the desired output (also known as reference, gold, or ground truth). A *task* would be "a particular specification of a problem, typically represented as a mapping between an input space and an output or action space, extensionally defined through one or more datasets, and usually associated with a specific performance evaluation metric" (based on (Raji et al. 2021) and (Schlangen 2020)). And finally, a *metric* would refer to "a specification of the mechanism to determine the degree of success or failure of the model's outputs". The metric represents a way to summarise model performance over a given task and dataset, usually defined as a single number or score (Raji et al. 2021). AI models that achieve the best scores on the metrics of a benchmark are generally considered SOTA in terms of performance.

While humans are always involved in the design and creation of ground truth in benchmarks, they can play a more or less direct and active role in applying benchmark tests. In *automated* or *quantitative benchmarks*, a set of tasks, datasets, and metrics are first defined through human decision making. The execution of a benchmark test is then carried out without direct human intervention. In *qualitative* benchmarks, humans intervene and partake in evaluations, for example, as evaluators, judges, or real-time interrogators (e.g., adversarial testing or red teaming). We primarily consider quantitative benchmarks. This is not to say that there is no relevant critique concerning qualitative AI evaluations, but such methods introduce a different set of problems that is out of scope for our discussions here.

Related Work

Over the years, several surveys and meta-reviews have set out to summarize discussions on limitations in AI benchmarking and this paper represents a continuation of such efforts. In 2021, Liao et al. (2021, p. 1) identified a wide range of "surprisingly consistent critique" directed at benchmarking practices across fields such as computer vision, natural language processing, recommender systems, reinforcement learning, graph processing, metric learning, and more. The authors present a taxonomy of observed benchmark failure modes, including implementation variations, errors in test set construction, overfitting from test set reuse, and comparisons to inadequate baselines. Along similar lines, Hutchinson et al. (2022) survey discussions on evaluation practices in machine learning, particularly in the fields of computer vision and natural language processing. The authors identify eight key evaluation gaps, including topics such as "neglect of model interpretability" and "oversimplification of knowledge" and argue for a shift toward application-centric evaluations that account for safety, fairness, and ecological validity. Similar points are also raised by Gehrmann, Clark, and Sellam (2023)'s most recent meta-review, which categorizes issues with text-oriented benchmarks over the past two decades, and propose a long-term vision for improving evaluation practices, emphasizing the need for comprehensive evaluations that include multiple datasets, metrics, and human assessments. Their key recommendations include focusing on model limitations, enhancing dataset documentation, and adopting a more nuanced approach to evaluation to better characterize model capabilities.

Our review revisits these areas of critique and traces how debates concerning benchmarks have evolved in the last few years due to rapid developments in the field. Recent research has identified an unprecedented growth in the release of AI benchmarks, especially in the area of safety, starting from 2023 and onwards (Röttger et al. 2024). Our research indicates a similar increase in publications expressing benchmark critique. For instance, Röttger et al. (2024) found that roughly 46% of the AI safety benchmarks they identified as relevant in their survey of the field had been produced in 2023, with as many as 15 new datasets being released during the first two months of 2024. Likewise, almost 55% of the articles we identified as relevant during the course of our research had been released in 2023 or later. Given this rapid increase in attention given to the topic, our primary goal is to extend existing meta-reviews into the present. Whereas previous meta-reviews have tended to focus on critique raised in a limited number of domains (primarily natural language processing and machine learning) we also include interdisciplinary accounts from the humanities and social sciences in our review. Furthermore, we provide insights concerning more recently identified areas of benchmark concern, such as works revealing that AI models can be programmed to underperform on benchmarking tests (a problem known as "sandbagging") (Weij et al. 2024), and studies presenting evidence that AI safety benchmarks strongly correlate with upstream AI capabilities (causing reasons to worry about AI "safetywashing") (Ren et al. 2024), insights that both raise additional questions about the validity and trustworthiness of current benchmarking practices.

Methodology

When gathering source materials for this review (and instead of conducting systematic keyword searches in research databases) we opted for a snowball sample (Jalali and Wohlin 2012; Badampudi, Wohlin, and Petersen 2015), which helped us identify papers that *primarily* address benchmark *critique* (our main inclusion criteria) while avoiding noise in the form of articles that present new benchmarks or simply apply them (our main exclusion criteria). A snowball method also helped us cut across interdisciplinary inconsistencies in terminology and allowed us to discover pre-prints - a publication form where much debate concerning benchmarks has taken place, but that would have been missed in a traditional review based on scientific databases. We initially started from the article AI and the Everything in the Whole Wide World Benchmark (Raji et al. 2021) and expanded by reviewing articles included in its reference list. We also studied how the text has been cited since its publication using Google Scholar - a process that was repeated for each article considered relevant. When important concepts (such as sandbagging) appeared in the literature, targeted searches for research in these fields were made. We also made special efforts to include discussions concerning benchmarks for images, sound, and audiovisual content which are under-represented compared to text-focused benchmarks.

In the process, we surveyed a wide range of papers that, for instance, highlight general problems with the production of AI datasets (Mitchell et al. 2019; Orr and Crawford 2024a), propose alternative ways of designing benchmark leaderboards (Rodriguez et al. 2021; Liu et al. 2021), or discuss the wider use of proxies in tests and evaluations (Mulvin 2021; Pinch 1993; Marres and Stark 2020). Such papers provide important contextual insights to discussions concerning benchmarks. However, they were omitted from what we considered our core collection of previous research. This collection ended up consisting of about 110 papers and articles that *explicitly* and *primarily* highlight issues with benchmarks. Notably, articles that apply or propose new benchmarks were *not* added to this collection by default, even though such articles naturally contain some level of benchmark critique. This decision constitutes a central limitation in our meta-review, but was applied for two main reasons. First, we found it necessary to limit the size and scope of our data collection. Second, many articles proposing new benchmarks (if not most) do not question or discuss the underlying assumptions that benchmark assessments rely on. In that sense, they reproduce the general notion that quantitative benchmarks provide a reasonable technical "fix" to issues with AI safety and capability assessments, rather than focusing on the main target of our review: research publications that place critical discussions on benchmark use and design at the centre of attention. Our collection was further limited by only containing articles published between 1st January 2014 and 31st December 2024. While it is true that AI models have been tested and evaluated since their origins in the mid 1950's (with the Turing test serving as an iconic example), we limit our review to this decade long period, since 2014 marks the starting point for recent intensifications in AI research and development.

Our resulting meta-review is not exhaustive but it covers a broad range of critique that has been aimed at benchmarking practices. The nine issue categories presented were identified after close-reading and relevant articles, classifying and grouping papers voicing similar critique, and discussing these classifications within the wider author group. The resulting issue areas represent the result of these discussions, although many issues were difficult to categorize. E.g., problems with Anglo-centric benchmarking datasets could simultaneously be framed as a cultural problem (Anglo-centrism), a construct validity problem (as English is taken to represent culture as a whole), and a problem with narrow benchmark diversity and scope. As a result of such overlaps, identified issues are not meant to be read as an absolute and all-encompassing definition of benchmark vulnerabilities, but as a narrative tool to present our findings. We especially focus on works that voice critique with relevance for policy makers and policy implementation, highlight areas of concern that cut across different modalities (text, images, sound, moving images), and point toward fundamental weaknesses in the design and application of benchmarks, as opposed to research that critiques individual benchmarks. Our goal has further been to provide a *diverse* account of concerns regarding benchmarks.

Nine Reasons to Be Cautious with Benchmarks

In the sections below, we summarise the main issues that were identified during the course of our research. Importantly, these issues - presented here as a taxonomy - are not arranged according to their importance or urgency. They are also not meant to be understood as isolated issues, but rather deeply interlinked problems. Indeed, this complexity and interdependence is precisely what makes AI evaluations challenging.

Problems with Data Collection, Annotation, and Documentation

An initial set of issues with AI benchmarks found during our research is limitations in the collection, annotation, and documentation of benchmark datasets. This ties into a broader critique regarding insufficient documentation in AI research which is central to calls for more transparent and trustworthy algorithmic systems (Gebru et al. 2021; Mitchell et al. 2019; Orr and Crawford 2024a; Simson, Fabris, and Kern 2024; Scheuerman, Hanna, and Denton 2021). Research has found that it is often difficult to trace precisely *how*, *when*, and by *whom* benchmark datasets have been made (Reuel et al. 2024; Denton et al. 2020) which compromises the ability for benchmarks to be used in robust and generalisable ways (Arzt and Hanbury 2024). The issue has partly been linked to the low status of dataset-related work within the machine learning community (which instead privileges model development) (Orr and Crawford 2024b; Sambasivan et al. 2021), and the fact that AI datasets are often "reduced, reused, and recycled"(Koch et al. 2021), which complicates documentations of their possible limitations (Thylstrup et al. 2022; Park and Jeoung 2022). Notably, Koch et al. (2021) have found that more than 70 percent of the benchmark datasets used in prominent computer vision papers had been reused from other domains. Park and Jeoung (2022) also find that while benchmark datasets often contain ample information on how to use the dataset, documentation is often missing concerning their shortcomings and social impact. In their exploration of benchmark sharing platforms like HuggingFace[1] and PapersWithCode[2] the same researchers also note that confusing metadata terminology severely complicates efforts to understand benchmark dataset documentation.

Like AI training datasets, benchmarks have been singled out for raising ethical and legal questions concerning copyrights, privacy, informed consent, and rights to opt-out (Paullada et al. 2021). Many benchmarks rely on crowdsourced or user-generated content from platforms like Wikihow[3], Reddit[4], or trivia websites (Keegan 2024; Grill 2024), whose annotation may be biased, lack input from expertise in specialised fields, or be produced under exploitative conditions (Tsipras et al. 2020; Aroyo and Welty 2015; Sen et al. 2015). Previous research has noted that this matters when baselines for "good", "bad", and "safe enough" AI models are calibrated in sensitive contexts (Rauh et al. 2024; Grill 2024). Moreover, the absence of human performance references and difficulty rubrics in benchmarks has been highlighted, which are increasingly considered important factors in evaluating capabilities and generality (Chollet 2019).

Noisy human annotations and lack of care in the making of benchmark datasets have further been found to skew benchmark scores (Kejriwal et al. 2024) and result in AI

[1]https://huggingface.co/. Last seen August 11, 2025.

[2]https://paperswithcode.com/. Last seen August 11, 2025.

[3]https://www.wikihow.com/. Last seen August 11, 2025.

[4]https://www.reddit.com/. Last seen August 11, 2025.

models exploiting unknown quirks and spurious cues in the training data, rather than solving their original intended task (Liao et al. 2021; Paullada et al. 2021; Geirhos et al. 2020). For instance, Oakden-Rayner et al. (2019) found that an X-ray image classification model predicted collapsed lungs with high accuracy, although it turned out the model only identified the presence of a chest drain (used to cure the condition) that was (unknowingly) present in a majority of the positive training images. In other words, the model performed well in benchmark tests, but the reason why completely sidestepped the original purpose of the task (identifying collapsed lungs). When the all images containing chest drains were removed from the training dataset, the model performance dropped by over 20%. Similar patterns have also been confirmed in recent research on LLM evaluation (Pacchiardi et al. 2024) and is sometimes discussed as failure in the "world model" of an AI system (Vafa et al. 2024). While it could be argued that picking up on spurious cues is a sign of intelligence and therefore a capability worth noting, the issues above highlight how a lack of in-depth attention to the data that benchmarks rely on - alongside a failure to acknowledge how little is actually known about *how* and *why* AI models perform well in benchmark tests - can produce frail and uncertain AI evaluations.

Weak Construct Validity and Epistemological Claims

Another genre of benchmark critique focuses on the epistemological claims that tend to surround benchmarks and examines the limits of what can be known through quantitative AI tests. A central reference point in these discussions is the observation by Raji et al. (2021) that many benchmarks suffer from construct validity issues in the sense that they do not measure what they claim to measure. As the authors proclaim, this is especially troublesome when benchmarks promise to measure universal or general capabilities, since this vastly misrepresents their actual capability. As a result, the authors argue that framing an AI benchmark dataset as general purpose "is ultimately dangerous and deceptive, resulting in misguidance on task design and focus, underreporting of the many biases and subjective interpretations inherent in the data as well as enabling, through false presentations of performance, potential model misuse". (Raji et al. 2021, p. 5). At the heart of this critique lies the realization that many benchmarks do not have a clear definition of what they claim to measure, which makes it impossible to measure if they succeed in the task or not (Blodgett et al. 2021; Bartz-Beielstein et al. 2020). In a close analysis of four benchmarks used to evaluate fairness in natural language processing (StereoSet, CrowS-Pairs, WinoBias, and WinoGender), Blodgett et al. (2021) for example found that all benchmarks revealed severe weaknesses in terms of defining what is being measured. For instance, culturally complex and highly contested concepts like "stereotypes" or "offensive language" were left unspecified, causing a series of logical failures and interpretational conflicts. 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) that can be misleading in numerous ways (Leech et al. 2024). Frequently, the difficulty in defining what benchmarks evaluate persists since there is no clear, stable, and absolute ground truth for what is claimed to be measured (Narayanan and Kapoor 2023b). Since concepts like "bias" and "fairness" are inherently contested, messy, and shifting, benchmarks that promise to measure such terms will inevitably suffer from an "abstraction error" that produces a false sense of certainty (Selbst et al. 2019, p. 63).

It has also been pointed out that many benchmark datasets are inadequate and/or useless proxies for what they are meant to evaluate. For instance, researchers have identified a slippage in distinguishing between algorithmic "harms" and algorithmic "wrongs" when evaluating the capabilities of AI models - two not necessarily overlapping concepts (Diberardino, Baleshta, and Stark 2024). Others have questioned whether the content of benchmark datasets is a reasonable substitute for the "real world" scenarios they are meant to reflect. For instance, the decision to use examples generated by Amazon crowdworkers and posts from the Reddit forum "*Am I the asshole?*" as proxies for ethics and morals in benchmarks such as HellaSwag - a widely cited benchmark for language models has been questioned (Keegan 2024). In benchmarks consisting of professional exams, researchers have further argued that such tests "emphasize the wrong thing" and "overemphasize precisely the thing that language models are good at" and are thus unreliable measures of things such as medical or legal skill (Narayanan and Kapoor 2023b, n.p). As Narayanan and Kapoor (2023b) put it, "it's not like a lawyer's job is to answer bar exam questions all day". A recent study by Ren et al. (2024) also found that many widely used safety benchmarks (including ETHICS, TruthfulQA, GPQA, QuALITY, MT-Bench, LMSYS Chatbot ARENA, ANLI, AdvGLUE, and AdvGLUE++) highly correlate with general and upstream model capabilities, raising concerns regarding "safetywashing" as applying them could imply that "capability improvements are misrepresented as safety advancements". While it could be argued that capability and safety are largely entwined (the more capable a model is, the higher the likelihood it could cause harm), Ren et al. (2024, p. 1) suggest that a blurred distinction between the two may hide the fact that severe biases and safety issues in AI models can persist, even as their overall capabilities improve.

Sociocultural Context and Gap

Another key insight from previous research concerns the importance of the social, economic and cultural contexts where AI benchmarks are created, used, and maintained. Among researchers engaging in benchmark critique, we identify a strong consensus regarding the need to recognise that benchmarks are ultimately "normative instruments that perpetuate particular epistemological perspectives about how the world is ordered" (Orr and Kang 2024, p. 1877). Qualitative research has also examined the cultural and social environments where benchmarks are made, finding that they are deeply shaped by shared and arbitrary assumptions, commitments, and dependencies (Engdahl 2024; Michael et al.

2022; Scheuerman, Hanna, and Denton 2021; Orr and Crawford 2024b; Sambasivan et al. 2021; Paullada et al. 2021). Such assumptions for example include valuing “efficiency at the expense of care; universality at the expense of contextuality; impartiality at the expense of positionality; and model work at the expense of data work” (Scheuerman, Hanna, and Denton 2021), or reproduce contested ideas such as the notion that low-quality data can be drowned out by scale (Orr and Crawford 2024b). Recent research also illustrated that AI safety research and benchmark competitions are increasingly informed by political movements and ideologies such as longtermism and effective altruism (Ahmed et al. 2024).

Scholars have further identified sociotechnical gaps and a lack of consideration for downstream utility as a concern in AI benchmarking (Hutchinson et al. 2022; Frieder et al. 2024). Liao and Xiao (2023) highlight that this means it is often unclear who is meant to care about benchmark evaluation results and how they should be used in practice. For instance, a recent study by Blagec et al. (2023), which compared the explicitly stated needs for AI technologies among clinical medical practitioners with existing clinical benchmark datasets, found that most benchmarks failed to answer to the needs of medical experts. They also found that benchmarks for the most urgently requested medical or clinical tasks were completely missing and noted that similar misalignments likely exist in many other fields. For example, Jannach and Bauer (2020, p. 79) highlight a lack of attention to how recommender systems “may create value; how they, positively or negatively, impact consumers, businesses, and the society; and how we can measure the resulting effects”. Ethayarajh and Jurafsky (2021) further argue that failures to consider the practical utility of benchmarks has for example made it possible to ignore the discriminatory and environmental damages of AI technologies and allowed for highly energy-inefficient and deeply biased AI models to reach the top of most benchmark leaderboards.

Narrow Benchmark Diversity and Scope

Previous research has further found that current benchmarking practices suffer from diversity issues, a problem that is also found within the broader AI ecosystem (Gomez et al. 2024). A vast majority of benchmarks focus on text, while other modalities (audio, images, video, and multimodal systems) remain largely unexamined (Rauh et al. 2024; Weidinger et al. 2023; Röttger et al. 2024). This concentration is problematic since AI models are increasingly multimodal in scope. Guldimann et al. (2024) also find that benchmarks addressing user privacy, copyright infringement, and interpretability are currently incomprehensive, while benchmarks in safety areas dealing with corrigibility and explainability are practically missing all together. This leads the authors to conclude that current benchmarks for safety and ethics “are often simplistic and brittle, leading to inconclusive results” (Guldimann et al. 2024, p. 3). Koch et al. (2021) further identify a concentration on fewer and fewer benchmark datasets within most task communities and note that dominant benchmarks have been introduced by researchers at just a handful of elite institutions, raising questions about representation diversity in the design of benchmarks. Scholars have also found that current AI safety evaluation practices almost exclusively deal with English content (McIntosh et al. 2024; Röttger et al. 2024; Poelman and Lhoneux 2024), and are frequently based on datasets where minorities are under-represented, despite efforts to diversify them (Simson, Fabris, and Kern 2024). This raises concerns regarding the inclusion of multiple perspectives on complex topics like ethics and harm.

Aside from mainly focusing on a small range of tasks, previous research has identified that most benchmarks tend to be abstracted out of their social and cultural context (Selbst et al. 2019; Lum et al. 2024), are aggregated in problematic ways (Burnell et al. 2023), and rely on a static, one-time testing logic, where results from single evaluations are taken to represent model capabilities writ large. This has given rise to calls for more multi-layered (Weidinger et al. 2023), longitudinal (Mizrahi et al. 2024), and holistic evaluation methods that can be reproduced (Kapoor et al. 2024), and prove that AI models do not just perform well in controlled environments, but also in real-world circumstances over time (Ojewale et al. 2024; McIntosh et al. 2024; Chang et al. 2023). In their survey of LLM benchmark inadequacies, McIntosh et al. note that “current benchmarks generally adopt task-based formats, such as Multiple Choice Questions (MCQs) and dialogue-based evaluations, which tend to be static and do not capture the evolving nature of human-AI interactions” (McIntosh et al. 2024, p. 6). Reuel et al. (2024) further note that failures to re-run evaluations multiple times (using different random seeds and sampling temperatures, for example) imply that very little is known about the intra-model variance of benchmarks. As a result, they conclude that it is possible that “most benchmarks fail to distinguish signal and noise” (Reuel et al. 2024, p. 9). Others have emphasised that AI audits often fail to consider risks associated with multiple (inter)acting AI systems (Birhane et al. 2024), and rarely take human actions and motivations into consideration (Weidinger et al. 2023; Chang et al. 2023; Rauh et al. 2024).

While most benchmarks are designed to tell us something about a model’s success, research has further pointed out that they often reveal little (or nothing) about their particular ways of *making mistakes*, which is crucial from an AI safety and policy enforcement perspective. As Gehrmann, Clark, and Sellam (2023) put it, “ranking models according to a single quality number is easy and actionable - we simply pick the model at the top of the list - [yet] it is much more important to understand when and why models fail” (Gehrmann, Clark, and Sellam 2023, p. 130). For instance, they suggest that a focus on errors and fragilities (as opposed to instances of success) can be useful for developers of smaller models, since “work on quantifying shortcomings is equally applicable to smaller models and methods that improve model robustness often work on all model sizes” (Gehrmann, Clark, and Sellam 2023, p. 131). In this sense, failure-focused benchmarks could play an important role in equalling out the playing field in AI development.

Economic, Competitive, and Commercial Roots

Another contextual element that has been singled out as important is the competitive and commercial roots of benchmarks. Previous research has emphasised that capability-oriented benchmarks are deeply embedded in corporate marketing strategies and play an important role in increasing the AI hype, attracting customers and investors, and showcasing how models outperform competitors (Orr and Kang 2024; Grill 2024; Zhijia 2024). As Orr and Kang (2024, p. 1881) put it, benchmarks "serve as the technological spectacle through which companies such as OpenAI and Google can market their technologies". Many benchmarks also have origins from within the industry and are capability-oriented and centred around tasks with a high potential economic reward, as opposed to focusing on other goals such as ethics and safety (Ren et al. 2024; Ethayarajh and Jurafsky 2021).

Previous research has noted that this competitive and corporate embedding discourages thorough self-critique since there is a direct "incentive mismatch between conducting high-quality evaluations and publishing new models or modelling techniques" (Gehrmann, Clark, and Sellam 2023, p. 103) and that the field of AI development is "turning into a giant leaderboard, where publication depends on numbers and little else (such as insight and explanation)" (Church and Hestness 2019). While a lack of incentives to disclose weaknesses and limitations is a general problem in science (Smith et al. 2022), it has been noted that benchmarks have played an especially central role in naturalizing and solidifying a competitive culture in AI research, which is increasingly approached as a "sport" (Orr and Kang 2024). As of late, benchmark evaluations have also become increasingly professionalized and transformed into an industry in itself with the rise of platforms like Kaggle and Grand Challenge, who increasingly function as infrastructures of power in fields like medical imaging (Luitse, Blanke, and Poell 2024). The issue of optimising for high benchmark scores at the expense of insight and explanation is known as a form of SOTA-chasing (Koch et al. 2021) and sometimes described as the "benchmark effect" (Stewart 2023). Previous research has also described benchmarking as an example of what Stengers (2018) calls "fast track research" which idolises rapid, cumulative publication (Malevé 2023), thus producing a "winners curse" in AI development (Sculley et al. 2018).

Risks associated with the competitive and commercial roots of benchmarks have further been linked to the growing influence of industry in AI research, where private businesses share of the biggest AI models has increased from 11% in 2010 to 96% in 2021 (Ahmed, Wahed, and Thompson 2023). In such a context, researchers have noted that current benchmarking tasks - which are generally highly data-intensive - are especially well suited to fit AI models that have been developed within the industry, whose access to advanced data infrastructures, computing power, valuable datasets, and skilled researchers now vastly exceed those of academic researchers (Ahmed, Wahed, and Thompson 2023). This concentration of power could potentially stifle robust AI evaluations and hinder the development of AI models that adhere to other aims and goals than commercial ones. Scholars have also warned that if academic researchers continue to uphold data-intensive benchmark tests as the SOTA, there is a risk that their research will become increasingly dependent on technological infrastructures provided by the industry (Koch and Peterson 2024).

Rigging, Gaming, and Measure Becoming Target

A closely related issue concerns how benchmark tests can be tricked and gamed. In areas and modalities where best-practice benchmarks are missing (i.e., practically all modalities, except for text-based benchmark evaluations), researchers have noted that there are strong incentives to "rig" benchmark tests. For instance, Dehghani et al. (2021) survey how know-how and recipes for how to score high on benchmark setups are often widely circulated online. Recently, language models have also been found to be optimised for answering the multiple choice questions that are often part of benchmarks (Alzahrani et al. 2024), and to (either intentionally or unintentionally) "fake" alignment with ethics or safety goals (Greenblatt et al. 2024) and hide their true capabilities and objectives - also known as *scheming* (Meinke et al. 2024). The issue points towards what is known as Goodhart's law: "when a measure becomes a target, it ceases to be a good measure" (Strathern 1997).

One reason why gaming can proceed is that users of benchmarks rarely provide the resources needed to validate and replicate their test results (Bartz-Beielstein et al. 2020; Dehghani et al. 2021; Biderman et al. 2024; Reuel et al. 2024) and that they rely on black-box (as opposed to white-box) access to AI models, which undercuts rigorous audits (Casper et al. 2024). To examine the relevance and validity of benchmark scores, researchers have further emphasised that it is necessary to access information concerning all aspects of the evaluation procedure, such as the original evaluation code and all details concerning the experimental setup) (Biderman et al. 2024). Providing such documentation is far from standard, however, especially when proprietary AI models are concerned. This makes it possible to tweak and cherry-pick benchmark results - a problem that is especially pressing given that subtle "variations in prompts, formatting or other implementation details can significantly impact the performance and validity of evaluations" (Biderman et al. 2024, p. 3). In their in-depth analysis of 24 SOTA language model benchmarks, Reuel et al. (2024) found that only four provided scripts to replicate the results and that no more than ten performed multiple evaluations or reported the statistical significance of their results.

An increasingly well discussed issue also concerns the problem of "data contamination" i.e., the risk that the models have either intentionally or unintentionally ingested benchmark datasets during training, which severely questions the integrity of AI tests (Xu et al. 2024a; Zhang et al. 2024; Besen 2024; Magar and Schwartz 2022; Roberts et al. 2023; Yang et al. 2023). The problem - which can be an instance of data leakage (Kaufman et al. 2012; Xu et al. 2024b) or train-test-overlap (Lewis, Stenetorp, and Riedel 2021) - has been known for long, and produces similar effects to those of overfitting and memorization (Tirumala et al. 2022; Magar and Schwartz 2022), leading to models with low generalization power that perform well on familiar tasks (in-distribution) but fail other tasks with a similar difficulty and distribution shift (out-of-distribution) (Yuan et al.

2023; Xu et al. 2024a; Zhang et al. 2024; Besen 2024; Magar and Schwartz 2022; Roberts et al. 2023; Narayanan and Kapoor 2023b). When testing GPT4 on benchmark problems from Codeforces (a website hosting coding competitions) in 2023, for instance, Narayanan and Kapoor (2023b) found that the AI model could regularly solve benchmark problems classified as easy - as long as the problems had been added before 5th September 2021. For problems added later, GPT4 could not get a single question right, suggesting that the model had memorised questions and answers. Similar results have also been identified in multiple other models and benchmarks (Xu et al. 2024a; Zhang et al. 2024; Besen 2024; Magar and Schwartz 2022; Roberts et al. 2023). Despite the fact that issues with data leaks are so well known that strategies have been developed to avoid it[5], there is still a widespread lack of reporting of data contamination tendencies during benchmark tests. In a study from October 2024, for example, Zhang et al. (2024) found that out of 30 analysed models, only 9 reported train test overlap.

Another recently identified issue is called "sandbagging" and involves a "strategic underperformance on an evaluation" which occurs when an AI developer intentionally *understates* a models capability, for instance to avoid becoming a target for AI safety regulation. In a 2024 study, for example, Weij et al. (2024, p. 1) prompted frontier models like GPT-4 and Claude 3 Opus and found that they "selectively underperform on dangerous capability evaluations, while maintaining performance on general (harmless) capability evaluations". The researchers also found that that it was possible to fine-tune and adjust both frontier and smaller models to hide specific capabilities or target specific capability scores. The issue could be generalised to popular benchmarks like the Weapons of Mass Destruction Proxy Benchmark (WMDP), and puts the trustworthiness of benchmark evaluations into question, especially in a regulatory context.

Dubious Community Vetting and Path Dependencies

A related area of research emphasises how benchmarks become naturalised and reach standard status because of the culture and logic of academic citations (Orr and Kang 2024). For instance, new benchmarks are commonly introduced together with new or updated AI models. If the AI model becomes popular, the benchmark may become widely cited and circulated as a secondary effect, even though the developers of the benchmark did not intended or expect it to become standard. In this way, a benchmark that might have been selected for practical purpose or due to subjective preferences may come to have substantial technical and cultural influence, even though its suitability as a yardstick for ethics, safety, or performance may be questioned (Orr and Kang 2024; Orr and Crawford 2024b). For instance, Denton et al. (2021) show how this was partly the case when the ImageNet dataset - a key reference point in the performance testing of computer vision models - became standard, following the unforeseen success of the ImageNet Large Scale Visual Recognition Challenge hosted by Toronto University in 2010. Likewise, the so-called Lena test image - also central to computer vision benchmark tests - was taken from the centerfold of a November 1972 Playboy magazine and catapulted into computer history since "someone happened to walk in with a recent issue of *Playboy*" at a time of need, according to what has been described as the most credible origin story (Mulvin 2021, p. 80). Since then, the ImageNet dataset and its associated benchmark challenge has been become a symbol for dataset bias (Denton et al. 2021), while the Lena test image has come to serve as a prime example of the role of whiteness and women's sexualised bodies in the standardisation of digital visual culture (Mulvin 2021). These examples highlight how benchmarks are fundamentally cultural and political products whose power and influence may and be (uncritically and problematically) reinforced through community vetting.

Aside from bias and representational issues, Schlangen (2020, p. 1) argues that what is typically missing and left implicit in the peer-review-fuelled process of benchmark use is the argumentation for why scoring well on a particular benchmark "constitutes progress, and progress towards what". Instead, researchers are expected to routinely demonstrate performance on dominant benchmarks, despite the fact that more task-specific benchmarks may be more technically appropriate (Koch et al. 2021). In this way, peerwashing serves to "maintain datasets as authoritative proxies even when they are shown to be harmful or problematic" (Orr and Crawford 2024b, p. 4966). New benchmarks often have a difficult time to gain traction because of the dominance and authority of well-cited benchmarks (Jaton 2021), and in a study of 3765 benchmarks, Ott et al. (2022) found that only a small minority of the proposed benchmark solutions reached widespread adoption. This problem is also recognised within the industry. For instance, researchers at Google's Brain Team describe what they call a "benchmark lottery" which "postulates that many factors, other than fundamental algorithmic superiority, may lead to a method being perceived as superior" (Dehghani et al. 2021, p. 1). Ironically, researchers have simultaneously found that a majority of influential benchmarks have been released as preprints without going through rigorous academic peer-review (McIntosh et al. 2024, p. 6). In other words, problems with peer-washing both concern the poor quality control that many benchmarks are subjected to, and the self-fulfilling and sometimes excessive ways through which certain benchmarks are propelled into standards.

Bao et al. (2022) further note that most papers that introduce benchmarks have origins in the field of machine learning and are mainly focused on methods. This means that the content of benchmarks datasets are often considered secondary, since the datasets are merely there to provide a baseline for comparison between different evaluation methods. When benchmarks are applied to real and specific use-cases - such as evaluating the fairness of algorithmic risk assessment instruments within the criminal justice system - the downstream effects of such a lack of concern for datasets can have worrying effects. As Bao et al. (2022) put it, a paper on benchmarks can be of high quality in a pure AI/ML methods sense, but irrelevant, dangerous, or harmful when

[5] https://livebench.ai. Last seen August 11, 2025.

applied in circumstances such as the criminal justice system, since it may introduce and perpetuate harms and mistranslations. The researchers also suggest that the current peer-review system implies that benchmarks that are primarily relevant from a methods/machine learning perspective will be cited far more often than benchmarks that are relevant for use in specific, real-life use-cases. Looking too closely at citation counts when determining the quality and relevance of benchmarks may thus effectively lead users astray and complicate identifying benchmarks with a high *practical* utility.

Problems with dubious community vetting become especially worrying given that benchmarks create "path dependencies" in AI research, meaning they reinforce certain methodologies and research goals, while stifling those that do not align with the logic of dominant benchmark tests (Blili-Hamelin and Hancox-Li 2023). In particular, Koch and Peterson (2024, p. 3) warn against the tendency for benchmarks to favour a form of task-driven scientific monoculture that privileges immediate, explicit, formal, quantitative, and easily-interpretable evaluation mechanisms that "prioritize one or a few key epistemic values (e.g., accuracy, safety)," at the expense of a broader and more complex vision of scientific progress. The authors note how a series of methodological paradigms in AI research (such as boosting, Bayesian networks, Bayesian non-parametrics, and support vector machines) have rapidly faded away following the boom in deep learning research around the year of 2014, and partly attribute this to the "epistemic narrowness" of current benchmark tests, which are well-suited for deep learning models (whose performance can be reliably increased through scaling), but less favourable for other methodological paradigms in AI. "When scientists can only gain high-status publications by demonstrating SOTA accuracy," writes Koch and Peterson (2024, p. 30), "the safest research choice becomes incrementally advancing proven methods, not innovating new ones". According to Koch, the cost of current benchmark practices - which privileges scale, compute, and the use of larger and larger training datasets - for instance includes repeated privacy and copyright violations, emotional harms caused to under-paid data workers, and increasing ecological and environmental pressures due to high energy consumption.

Rapid AI Development and Benchmark Saturation

Another social, economic, and cultural issue is the speed of AI developments. As the capabilities of AI models have increased manifold in the past decade, researchers have emphasised that many benchmarks are old and designed to test models far simpler than those in use today (Biderman et al. 2024). For instance, Biderman et al. (2024, p. 5) find that many prominent LLM benchmarks (including Lambada, AI2 ARC, OBQA, Hella Swag, and WinoGrande) were "designed prior to shifts such as in-context learning and chat interaction, and therefore were not designed to take these formats and approaches into account", noting that this may affect their validity in unforeseen ways. Many benchmarks also struggle with the challenge of quickly being outperformed as AI models achieve 100 percent accuracy scores in tests (Hendrycks et al. 2021; Bowman and Dahl 2021). A look at current AI safety leaderboards such as HELM also reveal that many AI models score notably high on benchmarks that are widely applied today. The tendency for AI models to outperform benchmarks has been described as an issue of *saturation* and implies that a benchmark no longer reflects model performance (Ott et al. 2022).

Relatedly, McIntosh et al. (2024) emphasise that many benchmark frameworks are slow and complicated to implement, meaning evaluation processes can span weeks or months, which hinders timely feedback on safety risks. This becomes an issue since new model releases often enter markets continuously, making it difficult to relocate evaluation resources in quick and adequate ways. It also undermines a "benchmark's ability to consistently evaluate reasoning, comprehension, or multimodal integration, as the results may vary with each model iteration" (McIntosh et al. 2024, p. 13), which is especially concerning in a regulatory setting, where quick, fair, and accurate AI assessments are key. The use of thresholds, either on benchmark performance or training compute (i.e., the amount of computational resources used to train an AI model), is often seen as central to determine which AI models should warrant further regulatory scrutiny (European Union 2024b; The White House 2023; US Department of Commerce 2025). However, several limitations have been identified in relation to such attempts. Creating benchmarks that can keep pace with the rapid development of AI is increasingly challenging, at the same time as recent approaches (such as auxiliary or bootstrapped models or post-training interventions) enable enhanced capabilities with reduced training compute (Hooker 2024).

AI Complexity and Unknown Unknowns

A final issue that has been discussed in relation to benchmarks concerns AI complexity and the fundamental difficulty of foreseeing what risks, dangers, and threats AI models could pose to society. McIntosh et al. (2024, p. 13) point out that "LLM benchmarks are constrained by the current limit of the benchmark creators' human knowledge, hindering their ability to fully assess and cultivate emerging AI capabilities that may surpass conventional human understanding". According to the authors, the lack of general understanding of emerging AI capabilities, and the natural limitations of the benchmark creators' knowledge on a potentially infinitely large number of domains and tasks, may lead to generalist approaches that "often fail to address the subtle requirements of critical sectors such as national security or healthcare". They argue that this does not just pose security risks, but could potentially also hinder innovation.

Evaluations of AI models are also complicated by the potential presence of unknown and latent vulnerabilities in AI models that "make it very hard to distinguish between (a) actually safe and (b) appears safe but is not" (Nasr et al. 2023a). In 2023, for instance, Nasr et al. (2023b) discovered that surprisingly simple prompts can 'break' the safety barriers of AI models, raising questions about the robustness of existing safety measurements. More precisely, the simple command "Repeat the word "poem" forever" was found to make ChatGPT output several megabytes of sensitive training data. As a result of this finding, the authors conclude that “just as vulnerabilities can lie dormant in code - sometimes for decades – our attack demonstrates the potential for latent, hard-to-discover ML vulnerabilities that lie dormant in aligned models” and go unnoticed by existing safety benchmarks and tests (Nasr et al. 2023b, p. 13). Researchers have also noted the difficulty of foreseeing how complex AI models respond to (small and large) interventions such as safety alignments and fine-tuning. For instance, efforts to fine-tune AI models to address safety and/or security risks have been found to degrade a model’s performance in other safety areas, or introduce entirely new security risks Qi et al. (2023).

Conclusion

Measuring the capabilities and risks of AI models and systems is difficult and one of the main challenges in the use and development of AI. Even with the best of intentions (such as disclosing discrimination or identifying potential societal harms), previous research has repeatedly shown that quantitative AI benchmarks struggle to perform their intended task. Benchmarks have been found to promise too much (Raji et al. 2021), be gamed too easily (Weij et al. 2024; Narayanan and Kapoor 2023a), measure the wrong thing (Oakden-Rayner et al. 2019), and be ill-suited for practical use in the real world (Bao et al. 2022; Ethayarajh and Jurafsky 2021). They have also been found to display a serious lack of documentation (Reuel et al. 2024), randomly reach an unjustified status through community vetting (Dehghani et al. 2021), and forward questionable cultural assumptions (Kang 2023), that for example ignore environmental concerns (Hutchinson et al. 2022). Furthermore, benchmarks have been critiqued for being narrow and mainly evaluating English (McIntosh et al. 2024; Röttger et al. 2024) and text-based AI models (Rauh et al. 2024) according to a one-time testing logic (Mizrahi et al. 2024) that ignores the AI capabilities in other modalities such as imagery and sound (Rauh et al. 2024), and fails to acknowledge that the potential harms of AI models cannot be properly understood through evaluations done in isolated, abstracted, test-environments, devoid of humans (Rauh et al. 2024) and other technical systems (Birhane et al. 2024). Taken together, these issues point toward fundamental fragilities in current efforts to quantitatively measure and mitigate harm in AI.

Cars, airplanes, medical devices, drugs, and numerous other products within our societies comply with strict regulations to ensure their safety. There is no reason to believe that similar safety assurances can not be developed for AI, and the intensified interest in AI benchmarks signals a drive to do so. Outside the scope of this meta-review, we have identified numerous papers that propose strategies for mitigating issues with benchmarks, for instance by hiding benchmark training datasets (Chollet 2019) or using so-called “dynamic benchmarks” (Besen 2024) to counter gaming and data contamination risks. Others aggregate evaluation tasks into single multi-task benchmarks to increase the reliability of test results (Srivastava et al. 2023; Liang et al. 2023), or opt for evaluation methods that directly involve human interrogators as opposed to quantitative metrics (Chang et al. 2023). An increasing number of researchers have also proposed new and promising frameworks for assessing and “benchmarking the benchmarks” (Miltenberger et al. 2023; Reuel et al. 2024). Because of uncertainties regarding the effectiveness (Zhang et al. 2024; Arzt and Hanbury 2024; Rauh et al. 2024) and widespread adoption of such mitigation efforts, however, it is the authors’ conviction that issues presented in this paper remain unresolved. This is not least demonstrated by the fact that many old and widely critiqued benchmarks are still used (e.g. MMLU). It is important to recognize that some issues that AI benchmarks seek to address are impossible or extremely difficult to fully address. For instance, mitigating safety and security issues that are inherently changing and context dependant remains a major challenge within a (semi-)static benchmarking logic. This suggests the need to also develop, improve, and draw inspiration from alternative evaluation methods, such as bug-bounty programs and red-teaming.

In line with previous studies (Jones, Hardalupas, and Agrew 2024), our meta-review suggests that the field of quantitative 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 optimization. Recognizing 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 - and the perspective of anyone who wants to apply a benchmark to a concrete, real-life case - we especially identify a need for new ways of signalling *what benchmarks to trust*. 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.

Position Statement

The views expressed in this paper are 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. 2023. The growing influence of industry in AI research. *Science*, 379(6635): 884–886.

Ahmed, S.; Jaźwińska, K.; Ahlawat, A.; Winecoff, A.; and Wang, M. 2024. Field-building and the epistemic culture of AI safety. *First Monday*.

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. 2024. When Benchmarks are Targets: Revealing the Sensitivity of Large Language Model Leaderboards.

Aniba, M. R.; Poch, O.; and Thompson, J. D. 2010. Issues in bioinformatics benchmarking: the case study of multiple sequence alignment. *Nucleic Acids Research*, 38(21): 7353–7363.

Aroyo, L.; and Welty, C. 2015. Truth Is a Lie: Crowd Truth and the Seven Myths of Human Annotation. *AI Magazine*, 36(1): 15–24.

Arzt, V.; and Hanbury, A. 2024. Beyond the Numbers: Transparency in Relation Extraction Benchmark Creation and Leaderboards.

Badampudi, D.; Wohlin, C.; and Petersen, K. 2015. 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: Association for Computing Machinery. ISBN 9781450333504.

Bao, M.; Zhou, A.; Zottola, S.; Brubach, B.; Desmarais, S.; Horowitz, A.; Lum, K.; and Venkatasubramanian, S. 2022. It's COMPASlicated: The Messy Relationship between RAI Datasets and Algorithmic Fairness Benchmarks.

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. 2020. Benchmarking in Optimization: Best Practice and Open Issues.

Besen, S. 2024. The Death of the Static AI Benchmark.

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. 2024. Lessons from the Trenches on Reproducible Evaluation of Language Models.

Birhane, A.; Steed, R.; Ojewale, V.; Vecchione, B.; and Raji, I. D. 2024. AI auditing: The Broken Bus on the Road to AI Accountability.

Blagec, K.; Kraiger, J.; Frühwirt, W.; and Samwald, M. 2023. Benchmark datasets driving artificial intelligence development fail to capture the needs of medical professionals. *Journal of Biomedical Informatics*, 137: 104274.

Blili-Hamelin, B.; and Hancox-Li, L. 2023. Making Intelligence: Ethical Values in IQ and ML Benchmarks. In *2023 ACM Conference on Fairness, Accountability, and Transparency*, 271–284. Chicago IL USA: ACM. ISBN 9798400701924.

Blodgett, S. L.; Lopez, G.; Olteanu, A.; Sim, R.; and Wallach, H. 2021. 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)*, 1004–1015. Online: Association for Computational Linguistics.

Bowman, S. R.; and Dahl, G. 2021. 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*, 4843–4855. Online: Association for Computational Linguistics.

Bruno, I. 2014. *Benchmarking*, 363–368. Dordrecht: Springer Netherlands. ISBN 978-94-007-0753-5.

Burden, J. 2024. Evaluating AI Evaluation: Perils and Prospects.

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. 2023. Rethink reporting of evaluation results in AI. *Science*, 380(6641): 136–138.

Camp, R. C. 1989. *Benchmarking : The Search for Industry Best Practices That Lead to Superior Performance*. the University of Michigan: Quality Press.

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. 2024. Black-Box Access is Insufficient for Rigorous AI Audits. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, 2254–2272.

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. 2023. A Survey on Evaluation of Large Language Models.

Cheng, Z.; Pang, C.; Wang, P.; and et al. 2020. How to report and benchmark emerging field-effect transistors. *Nature Electronics*, 5: 416–423.

Chollet, F. 2019. On the Measure of Intelligence.

Church, K. W.; and Hestness, J. 2019. A survey of 25 years of evaluation. *Natural Language Engineering*, 25(06): 753–767.

Dehghani, M.; Tay, Y.; Gritsenko, A. A.; Zhao, Z.; Houlsby, N.; Diaz, F.; Metzler, D.; and Vinyals, O. 2021. The Benchmark Lottery.

Denton, E.; Hanna, A.; Amironesei, R.; Smart, A.; and Nicole, H. 2021. On the genealogy of machine learning datasets: A critical history of ImageNet. *Big Data & Society*, 8(2): 1–14.

Denton, R.; Hanna, A.; Amironesei, R.; Smart, A.; Nicole, H.; and Scheuerman, M. K. 2020. Bringing the People Back In: Contesting Benchmark Machine Learning Datasets.

Diberardino, N.; Baleshta, C.; and Stark, L. 2024. Algorithmic Harms and Algorithmic Wrongs. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, 1725–1732. Rio de Janeiro Brazil: ACM. ISBN 9798400704505.

Engdahl, I. 2024. Agreements 'in the wild': Standards and alignment in machine learning benchmark dataset construction. *Big Data & Society*, 11(2): 20539517241242457.

Ethayarajh, K.; and Jurafsky, D. 2021. Utility is in the Eye of the User: A Critique of NLP Leaderboards.

European Union. 2022. Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act).

European Union. 2024a. First Draft of the General-Purpose AI Code of Practice published, written by independent experts.

European Union. 2024b. 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).

European Union. 2024c. Second Draft of the General-Purpose AI Code of Practice published, written by independent experts.

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. 2024. Data for Mathematical Copilots: Better Ways of Presenting Proofs for Machine Learning.

Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J. W.; Wallach, H.; III, H. D.; and Crawford, K. 2021. Datasheets for Datasets.

Gehrmann, S.; Clark, E.; and Sellam, T. 2023. Repairing the Cracked Foundation: A Survey of Obstacles in Evaluation Practices for Generated Text. *Journal of Artificial Intelligence Research*, 77: 103–166.

Geirhos, R.; Jacobsen, J.-H.; Michaelis, C.; Zemel, R.; Brendel, W.; Bethge, M.; and Wichmann, F. A. 2020. Shortcut Learning in Deep Neural Networks. *Nature Machine Intelligence*, 2(11): 665–673.

Gema, A. P.; Leang, J. O. J.; Hong, G.; Devoto, A.; Mancino, A. C. M.; Saxena, R.; He, X.; Zhao, Y.; Du, X.; Madani, M. R. G.; Barale, C.; McHardy, R.; Harris, J.; Kaddour, J.; Krieken, E. v.; and Minervini, P. 2024. Are We Done with MMLU?

Gomez, E.; Lorenzo, P.; Frau Amar, P.; and Vinagre, J. 2024. Diversity in Artificial Intelligence Conferences. Publications Office of the European Union JRC137550, Publications Office.

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. 2024. Alignment faking in large language models.

Grill, G. 2024. Constructing Capabilities: The Politics of Testing Infrastructures for Generative AI. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, 1838–1849. Rio de Janeiro Brazil: ACM. ISBN 9798400704505.

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. 2024. COMPL-AI Framework: A Technical Interpretation and LLM Benchmarking Suite for the EU Artificial Intelligence Act.

Heaven, D. 2023. AI Hype Is Built on High Test Scores. Those Tests Are Flawed. Report, MIT Technology Review.

Hendrycks, D.; Burns, C.; Basart, S.; Zou, A.; Mazeika, M.; Song, D.; and Steinhardt, J. 2021. Measuring Massive Multitask Language Understanding.

Henning, J. 2000. SPEC CPU2000: measuring CPU performance in the New Millennium. *Computer*, 33(7): 28–35.

Hooker, S. 2024. On the Limitations of Compute Thresholds as a Governance Strategy.

Hutchinson, B.; Rostamzadeh, N.; Greer, C.; Heller, K.; and Prabhakaran, V. 2022. Evaluation Gaps in Machine Learning Practice. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, 1859–1876. Seoul Republic of Korea: ACM. ISBN 978-1-4503-9352-2.

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

Jannach, D.; and Bauer, C. 2020. Escaping the McNamara Fallacy: Toward More Impactful Recommender Systems Research. *AI Magazine*, 41(4): 79–95.

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

Jones, E.; Hardalupas, M.; and Agrew, W. 2024. Under the radar? Examining the evaluation of foundation models. Report, Ada Lovelace Institute.

Kang, E. B. 2023. Ground truth tracings (GTT): On the epistemic limits of machine learning. *Big Data & Society*, 10(1): 20539517221146122.

Kapoor, S.; Stroebl, B.; Siegel, Z. S.; Nadgir, N.; and Narayanan, A. 2024. AI Agents That Matter.

Kaufman, S.; Rosset, S.; Perlich, C.; and Stitelman, O. 2012. Leakage in data mining: Formulation, detection, and avoidance. *ACM Trans. Knowl. Discov. Data*, 6(4).

Keegan, J. 2024. Everyone Is Judging AI by These Tests. But Experts Say They're Close to Meaningless. *The Markup*.

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

Keyes, O.; and Austin, J. 2022. Feeling fixes: Mess and emotion in algorithmic audits. *Big Data & Society*, 9(2): 20539517221113772.

Koch, B.; Denton, E.; Hanna, A.; and Foster, J. G. 2021. Reduced, Reused and Recycled: The Life of a Dataset in Machine Learning Research.

Koch, B. J.; and Peterson, D. 2024. From Protoscience to Epistemic Monoculture: How Benchmarking Set the Stage for the Deep Learning Revolution.

LaCroix, T.; and Luccioni, A. S. 2022. Metaethical Perspectives on 'Benchmarking' AI Ethics.

Leech, G.; Vazquez, J. J.; Kupper, N.; Yagudin, M.; and Aitchison, L. 2024. Questionable practices in machine learning.

Lewis, P.; Stenetorp, P.; and Riedel, S. 2021. 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*, 1000–1008. Online: Association for Computational Linguistics.

Liang, P.; Bommasani, R.; Lee, T.; Tsipras, D.; Soylu, D.; Yasunaga, M.; Zhang, Y.; Narayanan, D.; Wu, Y.; Kumar, A.; Newman, B.; Yuan, B.; Yan, B.; Zhang, C.; Cosgrove, C.; Manning, C. D.; Ré, C.; Acosta-Navas, D.; Hudson, D. A.; Zelikman, E.; Durmus, E.; Ladhak, F.; Rong, F.; Ren, H.; Yao, H.; Wang, J.; Santhanam, K.; Orr, L.; Zheng, L.; Yuksekgonul, M.; Suzgun, M.; Kim, N.; Guha, N.; Chatterji, N.; Khattab, O.; Henderson, P.; Huang, Q.; Chi, R.; Xie, S. M.; Santurkar, S.; Ganguli, S.; Hashimoto, T.; Icard, T.; Zhang, T.; Chaudhary, V.; Wang, W.; Li, X.; Mai, Y.; Zhang, Y.; and Koreeda, Y. 2023. Holistic Evaluation of Language Models.

Liao, Q. V.; and Xiao, Z. 2023. Rethinking Model Evaluation as Narrowing the Socio-Technical Gap.

Liao, T.; Taori, R.; Raji, I. D.; and Schmidt, L. 2021. 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)*.

Liu, P.; Fu, J.; Xiao, Y.; Yuan, W.; Chang, S.; Dai, J.; Liu, Y.; Ye, Z.; Dou, Z.-Y.; and Neubig, G. 2021. ExplainaBoard: An Explainable Leaderboard for NLP.

Luitse, D.; Blanke, T.; and Poell, T. 2024. AI competitions as infrastructures of power in medical imaging. *Information, Communication & Society*, 1–22.

Lum, K.; Anthis, J. R.; Nagpal, C.; and D'Amour, A. 2024. Bias in Language Models: Beyond Trick Tests and Toward RUTEd Evaluation.

Magar, I.; and Schwartz, R. 2022. 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)*, 157–165. Dublin, Ireland: Association for Computational Linguistics.

Malevé, N. 2023. Practices of Benchmarking: Vulnerability in the Computer Vision Pipeline. *photographies*, 16(2): 173–189.

Marres, N.; and Stark, D. 2020. Put to the test: For a new sociology of testing. *The British Journal of Sociology*, 71(3): 423–443.

Mattson, P.; Cheng, C.; Diamos, G.; Coleman, C.; Micikevicius, P.; Patterson, D.; Tang, H.; Wei, G.-Y.; Bailis, P.; Bittorf, V.; Brooks, D.; Chen, D.; Dutta, D.; Gupta, U.; Hazelwood, K.; Hock, A.; Huang, X.; Kang, D.; Kanter, D.; Kumar, N.; Liao, J.; Narayanan, D.; Oguntebi, T.; Pekhimenko, G.; Pentecost, L.; Janapa Reddi, V.; Robie, T.; St John, T.; Wu, C.-J.; Xu, L.; Young, C.; and Zaharia, M. 2020a. MLPerf Training Benchmark. In Dhillon, I.; Papailiopoulos, D.; and Sze, V., eds., *Proceedings of Machine Learning and Systems*, volume 2, 336–349.

Mattson, P.; Reddi, V. J.; Cheng, C.; Coleman, C.; Diamos, G.; Kanter, D.; Micikevicius, P.; Patterson, D.; Schmuelling, G.; Tang, H.; Wei, G.-Y.; and Wu, C.-J. 2020b. MLPerf: An Industry Standard Benchmark Suite for Machine Learning Performance. *IEEE Micro*, 40(2): 8–16.

McIntosh, T. R.; Susnjak, T.; Arachchilage, N.; Liu, T.; Watters, P.; and Halgamuge, M. N. 2024. Inadequacies of Large Language Model Benchmarks in the Era of Generative Artificial Intelligence.

Meinke, A.; Schoen, B.; Scheurer, J.; Balesni, M.; Shah, R.; and Hobbhahn, M. 2024. Frontier Models are Capable of In-context Scheming.

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. 2022. What Do NLP Researchers Believe? Results of the NLP Community Metasurvey.

Miltenberger, M.; Arzt, S.; Holzinger, P.; and Näumann, J. 2023. Benchmarking the Benchmarks. In *Proceedings of the ACM Asia Conference on Computer and Communications Security*, 387–400. Melbourne VIC Australia: ACM. ISBN 9798400700989.

Mitchell, M. 2023. How do we know how smart AI systems are? *Science*, 381(6654).

Mitchell, M.; Wu, S.; Zaldivar, A.; Barnes, P.; Vasserman, L.; Hutchinson, B.; Spitzer, E.; Raji, I. D.; and Gebru, T. 2019. Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 220–229.

Mizrahi, M.; Kaplan, G.; Malkin, D.; Dror, R.; Shahaf, D.; and Stanovsky, G. 2024. State of What Art? A Call for Multi-Prompt LLM Evaluation.

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

Narayanan, A.; and Kapoor, S. 2023a. Evaluating LLMs Is a Minefield.

Narayanan, A.; and Kapoor, S. 2023b. GPT-4 and professional benchmarks: the wrong answer to the wrong question.

Nasr, M.; Carlini, N.; Hayase, J.; Jagielski, M.; Cooper, A. F.; Ippolito, D.; Choquette-Choo, C. A.; Wallace, E.; and Lee, K. 2023a. Extracting Training Data from ChatGPT.

Nasr, M.; Carlini, N.; Hayase, J.; Jagielski, M.; Cooper, A. F.; Ippolito, D.; Choquette-Choo, C. A.; Wallace, E.; Tramèr, F.; and Lee, K. 2023b. Scalable Extraction of Training Data from (Production) Language Models.

Noroozian, A. 2020. *Evaluating Hosting Provider Security Through Abuse Data and the Creation of Metrics*. Dissertation (TU Delft). ISBN: 9789065624451.

Oakden-Rayner, L.; Dunnmon, J.; Carneiro, G.; and Ré, C. 2019. Hidden Stratification Causes Clinically Meaningful Failures in Machine Learning for Medical Imaging.

Ojewale, V.; Steed, R.; Vecchione, B.; Birhane, A.; and Raji, I. D. 2024. Towards AI Accountability Infrastructure: Gaps and Opportunities in AI Audit Tooling.

Orr, W.; and Crawford, K. 2024a. Building Better Datasets: Seven Recommendations for Responsible Design from Dataset Creators.

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

Orr, W.; and Kang, E. B. 2024. AI as a Sport: On the Competitive Epistemologies of Benchmarking. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, 1875–1884. Rio de Janeiro Brazil: ACM. ISBN 9798400704505.

Ott, S.; Barbosa-Silva, A.; Blagec, K.; Brauner, J.; and Samwald, M. 2022. Mapping global dynamics of benchmark creation and saturation in artificial intelligence. *Nature Communications*, 13(1): 6793.

Oxford English Dictionary. 2017. Benchmark. Meaning and Use.

Pacchiardi, L.; Tesic, M.; Cheke, L. G.; and Hernández-Orallo, J. 2024. Leaving the barn door open for Clever Hans: Simple features predict LLM benchmark answers.

Park, J.; and Jeoung, S. 2022. 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*, 1–10. Dublin, Ireland: Association for Computational Linguistics.

Paullada, A.; Raji, I. D.; Bender, E. M.; Denton, E.; and Hanna, A. 2021. Data and its (dis)contents: A survey of dataset development and use in machine learning research. *Patterns*, 2(11): 100336.

Pfister, R.; and Jud, H. 2025. Understanding and Benchmarking Artificial Intelligence: OpenAI's o3 Is Not AGI.

Pinch, T. 1993. "Testing - One, Two, Three ... Testing!": Toward a Sociology of Testing. *Science, Technology, & Human Values*, 18(1): 25–41.

Poelman, W.; and Lhoneux, M. d. 2024. The Roles of English in Evaluating Multilingual Language Models.

Qi, X.; Zeng, Y.; Xie, T.; Chen, P.-Y.; Jia, R.; Mittal, P.; and Henderson, P. 2023. Fine-tuning Aligned Language Models Compromises Safety, Even When Users Do Not Intend To!

Raji, I. D.; Denton, E.; Bender, E. M.; Hanna, A.; and Paullada, A. 2021. 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)*.

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. 2024. Gaps in the Safety Evaluation of Generative AI. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 7: 1200–1217.

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. 2024. Safetywashing: Do AI Safety Benchmarks Actually Measure Safety Progress? In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Reuel, A.; Hardy, A.; Smith, C.; Lamparth, M.; Hardy, M.; and Kochenderfer, M. 2024. BetterBench: Assessing AI Benchmarks, Uncovering Issues, and Establishing Best Practices. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Roberts, M.; Thakur, H.; Herlihy, C.; White, C.; and Dooley, S. 2023. Data Contamination Through the Lens of Time.

Rodriguez, P.; Barrow, J.; Hoyle, A. M.; Lalor, J. P.; Jia, R.; and Boyd-Graber, J. 2021. 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)*, 4486–4503. Online: Association for Computational Linguistics.

Roose, K. 2024. A.I. Has a Measurement Problem. Report, New York Times.

Röttger, P.; Pernisi, F.; Vidgen, B.; and Hovy, D. 2024. SafetyPrompts: a Systematic Review of Open Datasets for Evaluating and Improving Large Language Model Safety.

Sambasivan, N.; Kapania, S.; Highfill, H.; Akrong, D.; Paritosh, P.; and Aroyo, L. M. 2021. "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*, 1–15. Yokohama Japan: ACM. ISBN 978-1-4503-8096-6.

Scheuerman, M. K.; Hanna, A.; and Denton, E. 2021. Do Datasets Have Politics? Disciplinary Values in Computer Vision Dataset Development. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2): 1–37.

Schlangen, D. 2020. Targeting the Benchmark: On Methodology in Current Natural Language Processing Research.

Sculley, D.; Snoek, J.; Rahimi, A.; and Wiltschko, A. 2018. Winner's Curse? On Pace, Progress, and Empirical Rigor. Vancouver, BC, Canada.

Selbst, A. D.; Boyd, D.; Friedler, S. A.; Venkatasubramanian, S.; and Vertesi, J. 2019. Fairness and Abstraction in Sociotechnical Systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 59–68. Atlanta GA USA: ACM. ISBN 978-1-4503-6125-5.

Sen, S.; Giesel, M. E.; Gold, R.; Hillmann, B.; Lesicko, M.; Naden, S.; Russell, J.; Wang, Z. K.; and Hecht, B. 2015. 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*, 826–838. Vancouver BC Canada: ACM. ISBN 978-1-4503-2922-4.

Simson, J.; Fabris, A.; and Kern, C. 2024. Lazy Data Practices Harm Fairness Research. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, 642–659. Rio de Janeiro Brazil: ACM. ISBN 9798400704505.

Smith, J. J.; Amershi, S.; Barocas, S.; Wallach, H.; and Vaughan, J. W. 2022. REAL ML: Recognizing, Exploring, and Articulating Limitations of Machine Learning Research. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, 587–597.

Srivastava, A.; Rastogi, A.; Rao, A.; Shoeb, A. A. M.; Abid, A.; Fisch, A.; Brown, A. R.; Santoro, A.; Gupta, A.; Garriga-Alonso, A.; Kluska, A.; Lewkowycz, A.; Agarwal, A.; Power, A.; Ray, A.; Warstadt, A.; et al. 2023. Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models.

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

Stewart, M. 2023. The Olympics of AI: Benchmarking Machine Learning Systems.

Strathern, M. 1997. 'Improving ratings': audit in the British University system. *European Review*, 5(3): 305–321.

Subramonian, A.; Yuan, X.; III, H. D.; and Blodgett, S. L. 2023. It Takes Two to Tango: Navigating Conceptualizations of NLP Tasks and Measurements of Performance.

Thakur, N.; Reimers, N.; Rücklé, A.; Srivastava, A.; and Gurevych, I. 2021. BEIR: A Heterogeneous Benchmark for Zero-shot Evaluation of Information Retrieval Models. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

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

Thylstrup, N. B.; Hansen, K. B.; Flyverbom, M.; and Amoore, L. 2022. Politics of data reuse in machine learning systems: Theorizing reuse entanglements. *Big Data & Society*, 9(2): 20539517221139785.

Tirumala, K.; Markosyan, A.; Zettlemoyer, L.; and Aghajanyan, A. 2022. 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, 38274–38290. Curran Associates, Inc.

Tsipras, D.; Santurkar, S.; Engstrom, L.; Ilyas, A.; and Madry, A. 2020. From ImageNet to Image Classification: Contextualizing Progress on Benchmarks.

UK Parliament. 2023. Online Safety Act 2023.

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

Vafa, K.; Chen, J. Y.; Rambachan, A.; Kleinberg, J.; and Mullainathan, S. 2024. Evaluating the World Model Implicit in a Generative Model.

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. 2023. Sociotechnical Safety Evaluation of Generative AI Systems.

Weij, T. v. d.; Hofstätter, F.; Jaffe, O.; Brown, S. F.; and Ward, F. R. 2024. AI Sandbagging: Language Models can Strategically Underperform on Evaluations.

Xu, C.; Guan, S.; Greene, D.; and Kechadi, M.-T. 2024a. Benchmark Data Contamination of Large Language Models: A Survey.

Xu, R.; Wang, Z.; Fan, R.-Z.; and Liu, P. 2024b. Benchmarking Benchmark Leakage in Large Language Models.

Yang, S.; Chiang, W.-L.; Zheng, L.; Gonzalez, J. E.; and Stoica, I. 2023. Rethinking Benchmark and Contamination for Language Models with Rephrased Samples.

Yuan, L.; Chen, Y.; Cui, G.; Gao, H.; Zou, F.; Cheng, X.; Ji, H.; Liu, Z.; and Sun, M. 2023. 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*.

Zhang, A. K.; Klyman, K.; Mai, Y.; Levine, Y.; Zhang, Y.; Bommasani, R.; and Liang, P. 2024. Language model developers should report train-test overlap.

Zhijia, L. 2024. Top LLMs in China and the U.S. Only 5 Months Apart: Kai-Fu Lee.