Stanford University
Human-Centered
Artificial Intelligence

Policy Brief
HAI Policy & Society
December 2024

What Makes a Good AI Benchmark?

Anka Reuel, Amelia Hardy, Chandler Smith, Max Lamparth, Malcolm Hardy, and Mykel J. Kochenderfer

THE RAPID ADVANCEMENT AND PROLIFERATION OF AI SYSTEMS, and in particular foundation models (FMs), has made AI evaluation crucial for assessing model capabilities and risks. AI model evaluations currently include both internal approaches—such as privately testing models on proprietary data—and external approaches—such as scoring models on public benchmarks. Researchers and practitioners alike have adopted AI benchmarks as a standard practice for facilitating comparisons between, measuring the performance of, tracking progress in, and identifying weaknesses in different models.

Yet, no studies to date have assessed the quality of AI benchmarks in general in a structured manner, including both FM and non-FM benchmarks. Further, no comparative analyses have assessed the quality differences across the benchmark life cycle between widely used AI benchmarks. This leaves a significant gap for practitioners who may be relying on these benchmarks to select models for downstream tasks and policymakers who are increasingly integrating benchmarking in their AI policy apparatuses.

Our paper, "BetterBench: Assessing AI Benchmarks, Uncovering Issues, and Establishing Best Practices," develops an assessment framework that considers 46 best practices across a benchmark's life cycle, drawing on expert interviews and domain literature. We evaluate 24 AI benchmarks—16 FM and 8 non-FM benchmarks—against this framework, noting quality differences

Key Takeaways

The rapid advancement and proliferation of AI systems, including foundation models, has catalyzed the widespread adoption of AI benchmarks—yet only very limited research to date has evaluated the quality of AI benchmarks in a structured manner.

We reviewed benchmarking literature and interviewed expert stakeholders to define what makes a high-quality benchmark, and developed a novel assessment framework for evaluating AI benchmarks based on 46 criteria across five benchmark life-cycle phases.

In scoring 24 AI benchmarks, we found large quality differences between them, including those widely relied on by developers and policymakers. Most benchmarks are highest quality at the design stage and lowest quality at the implementation stage.

Policymakers should encourage developers, companies, civil society groups, and government organizations to articulate benchmark quality when conducting or relying on AI model evaluations and consult best practices for minimum quality assurance.

across the two types of benchmarks. Looking forward, we propose a minimum quality assurance checklist to support test developers seeking to adopt best practices. We further make publicly available a living repository of benchmark assessments at betterbench.stanford.edu.

This research aims to help make AI evaluations more transparent and empower benchmark developers to improve benchmark quality. We hope to inspire developers, companies, civil society groups, and policymakers to actively consider benchmark quality differences, articulate best practices, and collectively move toward standardizing benchmark development and reporting.

> *We define a high-quality AI benchmark as one that is interpretable, clear about its intended purpose and scope, and usable.*

Introduction

Benchmarks are used in a variety of fields—from environmental quality to bioinformatics—to test and compare the performance of different systems or tools. In the context of AI, we adopt the definition of a benchmark as “a particular combination of a dataset or sets […], and a metric, conceptualized as representing one or more specific tasks or sets of abilities, picked up by a community of researchers as a shared framework for the comparison of methods.” Despite AI benchmarks having become standard practice, there are still vast inconsistencies when it comes to what these benchmarks measure and *how* the measurements are used. Because past work has already focused on the limitations of existing benchmarks, as well as specific proposals for data curation and documentation for AI benchmarks, our work aims to offer practical insights and proposes a rigorous framework that empowers developers to assess and enhance benchmark quality.

To understand what makes a high-quality, effective benchmark, we extracted core themes from benchmarking literature in fields beyond AI and conducted unstructured interviews with representatives from five stakeholder groups, including more than 20 policymakers, model developers, benchmark developers, model users, and AI researchers. The core themes include:

- Designing benchmarks for downstream utility, for example, by making benchmarks situation- and use-case-specific.
- Ensuring validity, for example, by outlining how to collect and interpret evidence.
- Prioritizing score interpretability, for example, by stating evaluation goals and presenting results as inputs for decision-making, not absolutes.
- Guaranteeing accessibility, for example, by providing data and scripts for others to reproduce results.

Based on this review, we define a *high-quality* AI benchmark as one that is interpretable, clear about its intended purpose and scope, and usable. We also identified a five-stage benchmark life cycle and paired each benchmark stage with criteria we could use for our quality assessments:

1. Design: 14 criteria (e.g., Have domain experts been involved in the development?)
2. Implementation: 11 criteria (e.g., Is the evaluation script available?)
3. Documentation: 19 criteria (e.g., Is the applicable license specified?)
4. Maintenance: 3 criteria (e.g., Is a feedback channel available for users?)
5. Retirement: no criteria (only suggested best practices in our paper, since we cannot evaluate the retirement of active benchmarks)

We used this scoring system to assess 16 FM benchmarks (including MMLU, HellaSwag, GSM8K, ARC Challenge, BOLD, WinoGrande, and TruthfulQA) and 8 non-FM benchmarks (including Procgen, WordCraft, FinRL-Meta, and MedMNIST v2) according to each criterion, assigning 15 (fully meeting criterion), 10 (partially meeting), 5 (mentioning without fulfilling), or 0 (neither referencing nor satisfying). At least two authors independently reviewed each benchmark and reached consensus on all final scores.

AI model benchmarks—including ones that are commonly used—vary significantly in their quality.

Research Outcomes

Our research highlights that AI model benchmarks—including ones that are commonly used—vary significantly in their quality. For example, the widely used MMLU benchmark scored the lowest on usability (5.0) among all 24 benchmarks we evaluated, while another commonly used benchmark, GPQA, scored much higher (10.9). Yet it is common for developers to report results on both MMLU and GPQA without articulating their limitations or quality differences—for example, when introducing major models such as GPT-4, Claude-3, and Gemini. Similarly, the UK's AI Safety Institute has developed a framework for evaluating LLMs that includes both MMLU and GPQA, while the EU AI Act specifically mentions the use of such benchmarks. This means policymakers and other actors often rely on conflicting and even misleading evaluations.

Most benchmarks we evaluated also fail to distinguish between signals and noise. Developers may test two models with one benchmark but struggle to understand if different results reflect genuine performance differences or merely noisy outputs.

Implementation remains another major weakness of AI benchmarks. Both FM and non-FM benchmarks, on average, achieve their highest scores at the design stage (10.6 and 11.1 on average, respectively) and their lowest scores at the implementation stage (5.5 and 7.4 on average, respectively), in line with previously reported implementation challenges.

Of note, both FM and non-FM benchmarks are particularly weak on the reproducibility and interpretation of results: 17 of 24 benchmarks do not provide easy-

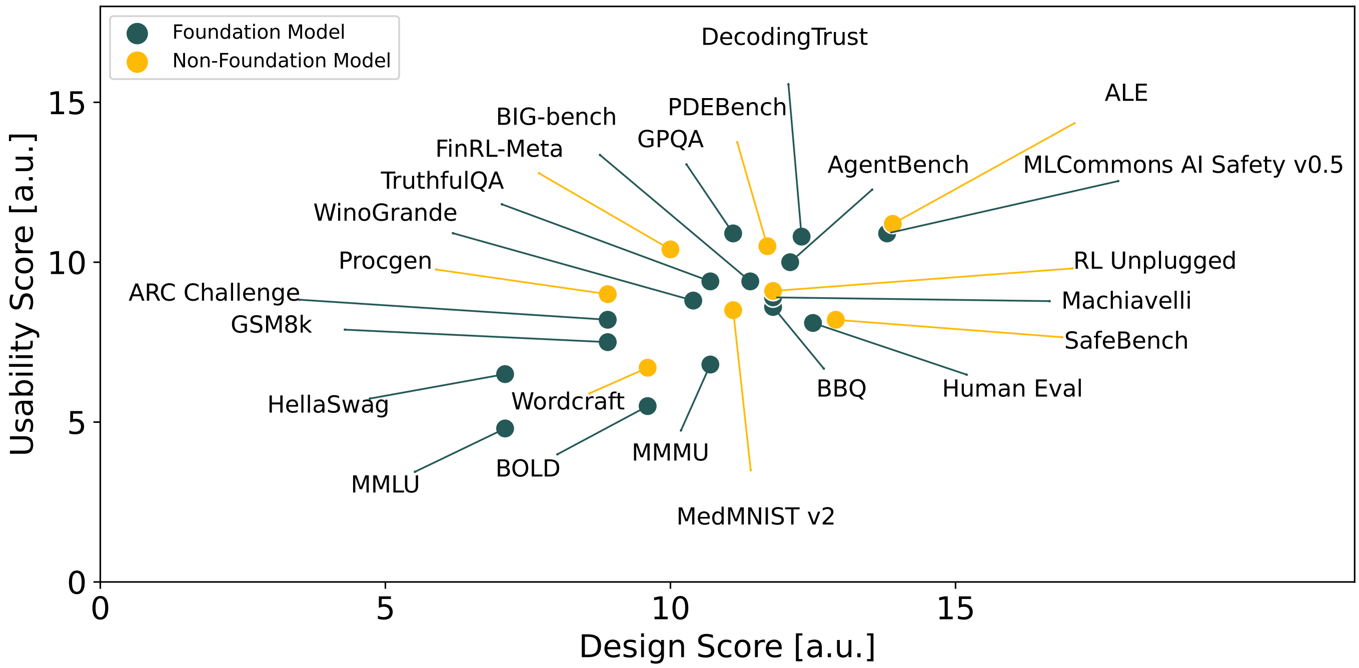

Figure 1: Design and usability scores for all 24 assessed benchmarks, whereby the usability score is the weighted average of the implementation, documentation, and maintenance scores.

to-run scripts to replicate the results from initial papers, and only 4 of 24 benchmarks provide scripts to replicate some of the results. This is a problem. Reproducibility is important for validating benchmarks, but there are clear gaps when it comes to empowering developers, companies, civil society groups, and policymakers to evaluate and replicate results.

We also found statistically significant correlations between design and usability scores for FM and non-FM benchmarks, suggesting that poorly designed benchmarks tend to be less usable.

Finally, the strong discrepancies we found in AI benchmark quality highlight the urgent need for the development of best practices that can help ensure

Reproducibility is important for validating benchmarks, but there are clear gaps when it comes to empowering developers, companies, civil society groups, and policymakers to evaluate and replicate results.

a minimum quality standard for AI benchmarks, especially given their increasing popularity and use in governance contexts. We developed a checklist of 46 best practices and encourage developers to adopt these during their process of creating benchmarks. These include, for example, making sure that the benchmark clearly describes how scores should be interpreted, makes its evaluation code publicly available, documents its limitations, and includes a feedback channel for users.

In addition to identifying best practices, we outlined a variety of design considerations that benchmark developers should take into account when developing high-quality benchmarks but that were either context-dependent or harder to operationalize as concrete criteria. These include considering whether to prioritize broad concept benchmarks or those focused on specific AI contexts and domains; how to assess multimodal models across their multiple modalities; whether to prioritize dynamic versus static benchmarks in different situations; and how to prevent cheating and to ensure evaluations accurately reflect model performance.

Policy Discussion

AI benchmarks have already become a widely accepted tool for developers to compare model performance and in some cases inform decisions regarding downstream tasks. Policymakers, too, are increasingly working to understand AI benchmarks, promoting their use across companies and relying on their outcomes for policy decisions. Making AI benchmarks more practicable, transparent, and comparable is therefore crucial.

> *Small changes can lead to significant improvements in overall benchmark practices.*

Our research underscores that policymakers should go one step further to articulate what makes AI benchmarks high-quality and what their limitations are—stating clearly in guidance documents that benchmarks vary in quality and approach, and that developers should strive to articulate the quality of their benchmark evaluations. These statements alone should clarify misconceptions about different benchmarks' applicability and encourage industry to strengthen benchmarks' interpretability and usability. Articulating quality metrics across benchmark life cycle stages (e.g., when they are designed vs. implemented) can strengthen evaluations.

Our research also shows that small changes can lead to significant improvements in overall benchmark practices. Many criteria we laid out over the benchmarking life cycle's five phases are relatively easy to implement, even for existing AI benchmarks. For example, adding code documentation and a point of contact to a benchmark are not time-consuming, but they can significantly enhance a benchmark's usability, transparency, and accountability. Developers and civil society groups should make these kinds of measures an explicit best practice, and policymakers should integrate such recommendations into their AI evaluation guidance.

Policymakers should additionally encourage research on open challenges in AI benchmarking. These include quick saturation (benchmarks becoming quickly outdated because model capabilities advance so quickly that models achieve near-perfect scores), contamination (model developers training on benchmark data, such as when scraping the web), poor construct validity (not designing a test such that it accurately measures the concept it is intended to measure), and standardization of benchmark reporting. Future research on these topics could build on our concept of measuring the quality of benchmarks and focus further on empowering developers and evaluators to produce systematic, repeatable, and interpretable results for different AI applications.

The greater adoption of AI benchmarks helps with systematic model evaluation, transparency, and, ideally, accountability. By adopting this framework and checklist to generate higher-quality benchmarks, we hope that developers, policymakers, and other stakeholders can make better-informed model selections and decisions for downstream tasks—while potentially reducing risks and improving outcomes in high-stakes applications.

Reference: The original article is accessible at Anka Reuel, Amelia Hardy et al., "**BetterBench: Assessing AI Benchmarks, Uncovering Issues, and Establishing Best Practices**," arxiv.org, November 20, 2024, https://arxiv.org/abs/2411.12990. The article was accepted as a Spotlight Paper at NeurIPS 2024. A living repository of benchmark assessments is available at http://betterbench.stanford.edu/.

Stanford University's Institute for Human-Centered Artificial Intelligence (HAI) applies rigorous analysis and research to pressing policy questions on artificial intelligence. A pillar of HAI is to inform policymakers, industry leaders, and civil society by disseminating scholarship to a wide audience. HAI is a nonpartisan research institute, representing a range of voices. The views expressed in this policy brief reflect the views of the authors. For further information, please contact **HAI-Policy@stanford.edu**.

Anka Reuel is a PhD student in computer science at the Stanford Trustworthy AI Research Lab and the Stanford Intelligent Systems Lab at Stanford University.

Amelia Hardy is a researcher with the Stanford NLP Group at Stanford University.

Chandler Smith is a research scholar at ML Alignment & Theory Scholars.

Max Lamparth is a postdoctoral fellow at the Center for International Security and Cooperation at Stanford University and the Stanford Center for AI Safety.

Malcolm Hardy is a researcher at the Stanford Intelligent Systems Labs at Stanford University.

Mykel J. Kochenderfer is an associate professor of aeronautics and astronautics at Stanford University and a faculty affiliate of the Stanford Institute for Human-Centered Artificial Intelligence.

Stanford University
Human-Centered
Artificial Intelligence

Stanford HAI: 353 Jane Stanford Way, Stanford CA 94305-5008
T 650.725.4537 **F** 650.123.4567 **E** HAI-Policy@stanford.edu **hai.stanford.edu**