Benchmarks

Joaquin Vanschoren JOAQUIN.VANSCHOREN@GMAIL.COM
Eindhoven University of Technology

Luis Oala LUIS@BRICKROADAPP.COM
Brickroad, USA

Simon Frieder SIMON.FRIEDER@CS.OX.AC.UK
University of Oxford, UK
Benchmarks & Baselines, Austria

Anka Ruel ANKA@CS.STANFORD.EDU
Stanford University, USA

Isabelle Guyon GUYON@CHALEARN.ORG
Université Paris-Saclay, France, ChaLearn, USA, and Google, USA

Adrien Pavão PAVAO@MLCHALLENGES.COM
Université Paris-Saclay, France

Ihsan Ullah IHSAN2131@GMAIL.COM
ChaLearn, USA

Evelyne Viegas EVELYNEVS@OUTLOOK.COM
Microsoft, USA

Reviewed on OpenReview: *https://openreview.net/forum?id=XXXX*

Abstract

This chapter explores the essential topic of benchmarks within the realm of AI and machine learning. We commence with a historical perspective, tracing the origins and evolution of benchmarking practices over the years. The narrative then shifts to the specific applications of benchmarks in machine learning, discussing various modalities, fields, and techniques. A subsequent section addresses the pertinent issues encountered in benchmarking machine learning algorithms, emphasizing challenges in ensuring reproducibility in scientific experiments. The chapter concludes by considering the critical infrastructure required to facilitate robust and effective benchmarking. Through this exploration, readers will gain a comprehensive understanding of the importance, applications, and complexities of benchmarking in the AI domain.

Keywords: data, benchmark

1 Introduction

In the evolving landscape of machine learning, benchmarks play a critical role in advancing reproducible science by providing a standardized framework to evaluate, compare, and refine algorithms. Traditionally, benchmarks consist of publicly available datasets — such as ImageNet — and an evaluation metric — such as accuracy or F1 scores — that allow for repeated experimentation across different methods, creating a shared reference point for assessing model performance. These benchmarks are essential for building consensus within the community, offering a common ground where developers and researchers can measure their work against established standards.

But, although benchmarks offer a seemingly controlled environment for comparison, the validity of results can still be compromised for a variety of reasons, including biases in data or evaluation, task over-simplification, benchmark over-fitting or habituation, and the dynamic nature of technology progress and real-world problems that may render them rapidly obsolete.

Benchmarks come in various forms, including performance-based benchmarks that focus on a single or small set of tasks, meta-datasets that aggregate multiple datasets to test model generalization, and leaderboards, which provide ongoing rankings of models based on their performance on a particular dataset or task. Performance-based benchmarks typically emphasize task completion metrics such as accuracy or speed, while meta-datasets push models to generalize across multiple domains. Leaderboards, on the other hand, introduce a competitive dynamic, highlighting the top-performing models within a specific context. Despite their differences, all benchmarks aim to provide a structured environment for evaluating models, but the choice of metrics and tasks profoundly influences the conclusions that can be drawn from them.

Benchmarking is closely related to the concepts of evaluation and measurement. While a benchmark provides the structure —a set of tasks or datasets— evaluation refers to the actual process of testing a model against these tasks, and measurement involves the metrics used to quantify the model's performance. The diversity of metrics in modern AI benchmarks reflects the complexity of today's models. Accuracy alone is no longer sufficient; robustness, fairness, efficiency, and interpretability are becoming equally important in assessing a model's real-world utility. This broadening of evaluation criteria highlights the need for more sophisticated benchmarking methods that go beyond traditional single-metric approaches.

With the advent of large-scale models like transformers and generative AI systems, the landscape of benchmarking has dramatically changed, introducing new challenges and reshaping the way models are evaluated. One of the most significant shifts has been the increased computational requirements for benchmarking large models. Where once it was feasible for researchers to train models from scratch on public platforms, today's benchmarks often focus on inference due to the high resource demands of training large models. This shift has limited participation to institutions with access to vast computational resources, reducing the inclusivity of the benchmarking process and raising concerns about accessibility and environmental impact.

Human-in-the-loop evaluations may be essential in certain cases where automated metrics alone are insufficient to capture the nuances of model performance. For tasks that involve subjective judgments—such as assessing aesthetics, detecting harmful content, or evaluating fairness—human input provides critical context that machines cannot fully replicate. While automated evaluations, often performed by models trained to mimic human judgments, can approximate these evaluations to some extent, they are not always reliable substitutes for human oversight. In fact, relying solely on automated evaluations risks overlooking subtleties in language, cultural context, or ethical implications that human evaluators are better equipped to identify. As models grow more complex and are deployed in real-world scenarios, integrating human assessments alongside automated metrics can offer a more comprehensive evaluation, combining the speed of automation with the depth of human judgment. This hybrid approach ensures that models are not only optimized for performance but also align with societal values and expectations.

In a book focused on challenges and benchmarks, it is important to clarify the distinction between the two. Challenges typically introduce a novel problem, accompanied by a unique dataset in a competitive setting, operating within a set time frame, and selecting winners based on a single specific metric. This may lead to complex engineered solutions making it difficult to evaluate the

role of the various algorithmic components. In contrast, benchmarks aim to evaluate algorithms across multiple metrics, fostering broader research and development rather than prioritizing a single outcome. While challenges often focus on identifying the best-performing team under particular conditions, they can evolve into benchmarks to compare multiple algorithms with multiple metrics beyond the challenge deadline. A common difference is also that challenges hide the test set from participants, following the Common Task Framework Donoho (2017b), while benchmarks usually rely on the user's honor not to peek at or overfit to the test data .

As benchmarks evolve, there is increasing interest in dynamic benchmarking, which continuously adapts to new tasks and challenges, preventing models from saturating static benchmarks. However, maintaining dynamic benchmarks introduces challenges in ensuring stability, fairness, and long-term relevance. The need for ongoing maintenance and community involvement is critical to keep benchmarks up to date and reflective of the evolving nature of machine learning.

This chapter explores these issues in depth. First, we examine the challenges of establishing and maintaining benchmarks, particularly in the context of large models and dynamic benchmarks. We then consider what makes a great benchmark, identifying key factors like fairness, reproducibility, and comprehensiveness. A closer look at lessons learned from the NeurIPS Datasets and Benchmarks Track provides insights into the practicalities of designing benchmarks. We also review applications of benchmarks in advancing machine learning research, followed by a discussion of emerging best practices to ensure benchmarks remain robust and inclusive. Finally, we offer historical context on the evolution of benchmarking, tracing how it has adapted to the changing demands of machine learning research.

1.1 Historical overview

Benchmarks have played a crucial role in the evolution of AI and predictive modeling (Liberman, 2010; Donoho, 2017a). Indeed, the empirical nature of the field naturally calls for quantitative and comparative benchmarks of AI models (Langley, 1988). In her NeurIPS 2022 keynote talk[1], Isabelle Guyon presents the rich history of benchmarks in machine learning, serving as a primary inspiration for the historical overview that follows.

The birth of the field of machine learning occurs in the twentieth century, hence we'll take this as a starting point. Before the 1950s, researchers had discovered and refined statistical techniques, establishing the foundation for future developments. In 1950, Alan Turing proposed the *Turing Test*, a measure of a machine's ability to exhibit intelligent behavior indistinguishable from that of a human. While not a machine learning challenge in the modern sense, it set an early standard for evaluating artificial intelligence (Turing, 1950). In the 1950s, pioneering machine learning research was carried out using light algorithms, and by the 1960s, Bayesian methods were introduced for probabilistic inference within the field. However, the 1970s marked an "AI winter", characterized by pessimism regarding the effectiveness of machine learning. This period of stagnation was followed by a revival in the 1980s, triggered by the discovery of the modern backpropagation algorithm.

Before the 1990s, datasets were exceedingly scarce, often consisting of what we now refer to as "toy data". It was common to primarily demonstrate new algorithms using synthetic data or small datasets. One such dataset remains commonly used in many introductory machine learning tutorials. This dataset is the Iris dataset (Fisher, 1936), which presents a 3-class classification challenge involving three distinct types of Iris flowers. Remarkably, these classes can be effectively separated

1. https://nips.cc/virtual/2022/invited-talk/56158

using a linear discriminant classifier, commonly known as Fisher's linear discriminant, employing just four features: petal length and width, as well as sepal length and width. Each class in the dataset comprises 50 examples.

The 1990's marked a shift from knowledge-driven to data-driven approaches in machine learning, with a focus on analyzing large datasets. This era witnessed the rise of support-vector machines (SVMs), recurrent neural networks (RNNs), and an increase in computational complexity through neural networks. An important landmark in ML dataset availability was the creation of the UCI ML repository in 1987 by David Aha and his students. The initial datasets were also rather small in size, with the number of examples ranging from 50 to a few hundred and the number of features not exceeding 100. The initial datasets were all tabular datasets, that is tables with samples in rows and features in columns. Raw data, like images, sound, text, video, appeared only later. These datasets were widely used in NeurIPS papers during the 1990's and as you can imagine people started overfitting them. Even more concerning, people started reporting results only on the subset of datasets that made their algorithms shine. This was denounced with humor in a joke paper submitted to NeurIPS in 2002, "Data Set Selection", pretending to give a theoretical backing to the bad habit of dataset selection (LaLoudouana and Tarare, 2002). The 1990's were marked also by the appearance of systematic benchmarks like the EU project Statlog (King et al., 1995). A consortium of 13 institutions worked together to compare a large number of methods on a large number of datasets, mostly coming from the UCI ML repository. They produced a ranking of algorithms for each dataset. The algorithms included classical statistics (linear and quadratic discriminant, logistic regression, KNN, BN), machine learning (decision tree, rule-based) and neural networks (MLP).

Interestingly, they made a methodology statement:

> *"The Project laid down strict guidelines for the testing procedure. First an agreed data format was established, algorithms were "deposited" at one site, with appropriate instructions [...]. Each dataset was then divided into a training set and a testing set, and any parameters in an algorithm could be "tuned" or estimated only by reference to the training set. Once a rule had been determined, it was then applied to the test data. This procedure was validated at another site by another (more naive) user. This ensured that the guidelines for parameter selection were not violated, and also gave some information on the ease-of-use for a non-expert in the domain."* – Michie et al. (1994)

The methodology described by the Statlog authors encapsulates crucial principles of machine learning. They separate the data into training and testing sets, which allows for the evaluation of a model's performance on unseen data. They ensure that model parameters are tuned only with reference to the training set, preventing overfitting. They also apply replication and validation procedures, enhancing the robustness of their findings and preventing over-optimization. By storing algorithms at a single site, they enable standardization and comparison, contributing to the development of robust and reproducible machine learning research and practices.

As the field entered the 2000's, we observed a rise in popularity of techniques like support-vector machines, kernel methods, and unsupervised learning. In these years, people from the NeurIPS community started focusing more on raw data, as opposed to data preprocessed as nice tables. In the realm of image recognition, many datasets relating to Optical Character Recognition (OCR) were collected. OCR is the technology used to convert different types of documents, such as scanned paper documents, PDF files or images captured by a digital camera, into editable and

searchable data. There was also an increased interest in datasets for face and object recognition, along with video datasets designed to recognize human actions. In the field of speech recognition, the Linguistic Data Consortium popularized various datasets (Cieri and Liberman, 2000; Maeda and Strassel, 2004). This organization also provided several extensive text corpora for language study, significantly contributing to language research and natural language processing. Researchers also started to explore other forms of sensor data, including Electroencephalogram (EEG) data, which captures electrical activity in the brain. Additionally, the first datasets representing graph data began to emerge, opening new possibilities for research and algorithm development in the field of graph theory and network analysis.

The 2010's marked the rise of deep learning, which made machine learning an essential component of various software services and applications used widely across industries. At the dawn of our millennium, the euphoria of "big data" led industry leaders to believe that all problems could eventually be solved by adding in more data. Peter Norvig, Director of Research at Google, is often quoted for having said in 2011 that "We don't have better algorithms, we just have more data". However, this simplistic idea has back-fired, with several embarrassing failures of algorithms making racist or sexist decisions (Zou and Schiebinger, 2018). This prompted Peter Norvig to revise his claim in 2017 to "More data beats clever algorithms, but better data beats more data". These are also the years where the first systematic competitions platform appeared, as *Kaggle* (Goldbloom and Hamner, 2010) was launched in 2010 and *CodaLab Competitions* (Pavao et al., 2023) in 2013. There was a noticeable trend of an increasing number of challenges being organized in conjunction with scientific conferences. ChaLearn, a non-profit organization focused on organizing machine learning challenge, was founded in 2011. At that time, the researchers behind ChaLearn had already organized many challenges, for instance, the Feature Selection Challenge at NIPS 2003 (Guyon et al., 2004), the Performance Prediction Challenge at WCCI 2006 (Guyon et al., 2006), and the Active Learning Challenge at AISTATS 2010 (Guyon et al., 2011). They pursue these efforts by hosting a variety of challenges, covering diverse areas like computer vision through the "Looking at People" series, neurology, causal discovery, automated machine learning and physics.[2]

This period also saw a rise in interest in the area of ethics in AI and algorithmic fairness, as the awareness of the societal implications of AI grew. More recently, there has been a growing emphasis on the need for explainable and interpretable AI, particularly given the opaque nature of many deep learning models. Despite the challenges, the progress made in these years set the stage for many exciting developments in machine learning that we are witnessing today.

And there we are now in the 2020's, hopefully at the beginning of maturity in Machine Learning. We envision the centerpiece to be peer review of datasets and benchmarks, which should become a standard in the field, that NeurIPS is strongly encouraging by establishing the NeurIPS D&B track. This decade also brings with it a stronger focus on the robustness and reproducibility of machine learning models, underlining the importance of understanding and documenting every step of the machine learning pipeline from data collection to model deployment. David Donoho coined in (Donoho, 2024) the term "frictionless reproducibility", which denotes the efforts (sharing data, sharing code that processes the data, and competitive testing) that were the drivers of the spectacular progress the field of machine learning has seen. The growing popularity of initiatives like PapersWithCode and HuggingFace underscores this point. These initiatives facilitate frictionless reproducibility by providing unified platforms that embody the principles underlying frictionless

2. `http://www.chalearn.org/challenges.html`

reproducibility – as well as numerous other features, such as social network mechanisms of "likes" or "stars" to increase structured user interaction. Competitions continue to demonstrate their efficiency in driving innovation and solving long-standing problems, as illustrated by the Critical Assessment of Structure Prediction 14 (CASP14) competition,[3] where DeepMind's AlphaFold 2 achieved breakthrough accuracy in protein structure prediction (Jumper et al., 2021b,a).

Large language models gave rise to a new generation of benchmark suites designed to probe their capabilities across many dimensions simultaneously. BIG-Bench (Gur-Ari et al., 2022) assembled over 200 tasks contributed by the research community, covering reasoning, language understanding, coding, and domain knowledge, with the explicit goal of identifying where language models fail. At the same time, Dehghani et al. (2021) warned of the "benchmark lottery": the choice of which benchmark to report often determines which model appears to win, independent of genuine algorithmic superiority, underscoring the need for diverse and carefully curated evaluation suites.

As language models rapidly saturated existing benchmarks, new evaluations raised the bar. The Abstraction and Reasoning Corpus (Chollet, 2019) tests fluid intelligence through visual puzzles that humans handle effortlessly but that have resisted machine learning approaches for years; competitions built around it confirmed that state-of-the-art models score significantly below human performance (Chollet et al., 2024). Humanity's Last Exam (Phan et al., 2025) assembled expert-level questions across hundreds of academic disciplines, where frontier models score poorly despite performing well on earlier academic benchmarks. FrontierMath (Glazer et al., 2024) targets research-level mathematics requiring genuine mathematical insight, with all known models scoring in the low single digits at time of publication. These increasingly hard benchmarks illustrate a recurring dynamic: once models approach saturation on one evaluation, a harder one emerges to re-establish the gap between machine and human performance.

2 Fundamental problems in establishing and maintaining benchmarks

The field of machine learning benchmarking faces fundamental difficulties that affect the development, implementation, and interpretation of benchmarks, highlighting deeper systemic issues within the discipline.

2.1 Longevity Versus Saturation

A fundamental tension in machine learning benchmarking lies in balancing the maintenance of relevance with the acknowledgement of progress. As the field rapidly advances, benchmarks that once posed significant challenges can quickly become saturated, with top-performing models achieving near-perfect scores. This phenomenon, observed in benchmarks such as GLUE (Wang et al., 2018) and ImageNet (Russakovsky et al., 2015), raises critical questions about the continued utility of these benchmarks and the necessity for more complex evaluation tasks (Zellers et al., 2019).
We note that saturation is distinct from biased model performance estimation on a benchmark. Saturation reflects whether a benchmark *is still useful for the community*, while biased model performance estimation is related to the question of how often one should use the same test data (which may be expensive to create) for model evaluation. While statistical learning theory presents a pessimistic view on the long-term validity of any benchmark, it has been shown that static benchmarks are valid for much longer than one would expect naively. Best practices originating from the statis-

3. `https://predictioncenter.org/casp14/`

tical learning theory advise that a hold-out dataset be used only *once*, and then never again, as the empirical risk is an unbiased estimator of the population risk (Hardt and Recht, 2022). Yet, this is completely opposite to the machine-learning practice, where benchmarks are used many times over, which leads to what has been called "training on the test error" as early as 1973 (Duda et al., 1973). While empirically, the important benchmark dataset on which models have been evaluated has been used much more often than once, this has not harmed progress. How could this apparent paradox be explained? The absolute performance values might naturally be skewed once the test data is used multiple times, as with each new model that one uses again on the same test data to obtain a new evaluation, information is leaked; while it is possible to put this into a formal framework that explains precisely how bits of information degrades performance (Hardt and Recht, 2022), newer work shows that not the absolute numbers of a benchmark is what matters, but rather the robustness of the model rankings is what characterizes a good benchmark (Salaudeen and Hardt, 2024). By reproducing the ImageNet dataset and creating the ImageN*o*t dataset, it has been shown that the rankings of the models are preserved, while the absolute performance values are not. This partly explains why benchmarks such as ImageNet, which exhibited a strong case of "training on the test data", still resulted in driving the performance of vision models over several years.

Another solution to the challenge of benchmark saturation has emerged in the form of evolving, dynamic benchmarks. These benchmarks aim to automatically adjust their difficulty or introduce new tasks based on the current state of the art, as exemplified by the BIG-bench project (Gur-Ari et al., 2022) or SWE-bench (Jimenez et al., 2024), which relies on real-world GitHub data and is designed to be easily updated by adding new data at successive time steps. However, implementing such adaptive systems presents its own set of technical and methodological challenges, including the need for continuous curation and the risk of introducing biases through automated task generation. Preliminary works have shown how such updates can be done within a formal framework, where datapoints are added via an API to an existing dataset, and models are evaluated selectively on the newly added datapoints, thus saving the cost that occurs if a full re-evaluation, on the full dataset, were performed (Prabhu et al., 2024). This approach, called "lifelong benchmarks", highlights how the scientific community can easily update the dataset underlying a benchmark, thereby extending its validity. While the previous updates were performed by hand, automatic variants of these approaches have been proposed where that the datasets underlying the benchmarks can be automatically updated (Ying et al., 2024), which drives the costs down to keep any particular benchmark up-to-date, and extends the longevity of the benchmark.

The concept of an evolving benchmark, raises the problem of dataset version control. The history of benchmarks shows clearly that they serve as a standard. Hence, similar to code, platforms such as Zenodo[4] have been created which assign a version number to each dataset release, thereby allowing researchers to tie the performance of their models to a specific version of a benchmark.

Lastly, we mention there is ongoing debate within the research community regarding the interpretation of benchmark saturation. There is a trend within the community that it represents meaningful progress in the field (Chollet, 2019). The sunsetting of truly saturated benchmarks, such as MNIST, while potentially disruptive to established research practices, may be necessary to drive innovation and prevent the field from overfitting to specific evaluation metrics.

To avoid disruption, the field has evolved to accommodate flexible benchmarking approaches. We mention three approaches and detail one well-known representative example below: Periodic dataset

4. `https://zenodo.org`

refreshment (the CASP challenge), adversarial human-in-the- loop (as represented by the DynaBench approach), and the testing breadth (as represented by HELM benchmark).

The Critical Assessment of Structure Prediction (CASP) is a long-running series of challenges (Moult, 2005), starting in 1994,[5] to evaluate prediction models for protein models. The AlphaFold models 1-3 (Senior et al., 2020; Jumper et al., 2021c; Abramson et al., 2024), reached prominence by the large improvements in this challenge, and. The CASP challenges managed to maintain the evergreen status, by keeping the structure of the challenge similar, while refreshing the data for each round.

A different approach to protect against benchmark saturation is taken by DynaBench (Kiela et al., 2021), which uses a human-in-the-loop approach and adversarial data collection. The core mechanism is as folows: annotators interact with a live target model and try to craft examples the model will misclassify but that another human would label correctly. As models improve, it becomes harder to fool them, which itself becomes a meaningful progress signal. This continual, adversarial loop both exposes current blind spots and yields fresh data.

Lastly, benchmarks that test a large number of skills instead of a single one, offer a larger saturation buffer: Although individual subtasks may easily be saturated, we should saturate them all with a single model. This is the idea behind the Holistic Evaluation of Language Models (HELM) benchmark (Liang et al., 2022). Several other benchmarks of these types exist, such as BIG-bench (Gur-Ari et al., 2022), or MMLU (Hendrycks et al., 2020), as well several variations of each of these including HELM, such as AHELM (Lee et al., 2025) which focuses on audio language models, or MedHELM (Bedi et al., 2025), which focuses on the medical domain.

2.2 Incentive Design Under Resource Constraints

The creation and maintenance of high-quality benchmarks require significant resources, both in terms of computational power and human expertise. The increasing demand for computational resources, particularly for training and evaluating large models, has created a significant barrier to entry for many researchers and institutions (Strubell et al., 2019). This disparity in access to high-performance computing resources can create an uneven playing field, potentially skewing benchmark results and limiting participation from resource-constrained institutions (Schwartz et al., 2020). Furthermore, the incentives for creating and maintaining benchmarks are not always aligned with the academic reward system, which often prioritizes novel algorithms over benchmark development. This misalignment can lead to a shortage of well-designed, comprehensive benchmarks. The challenge lies in engineering a benchmark ecosystem that is not only technically sound but also sustainable in the long term, balancing the needs of the research community with the available resources. In academia, the dominant incentive is to publish a benchmark dataset once—typically as a contribution to a venue such as the NeurIPS Datasets and Benchmarks Track—and move on. Long-term platform maintenance is far more resource-intensive and yields fewer citations. This also creates a conflict of interest: benchmark datasets distributed openly are more widely adopted, but also more susceptible to manipulation (selective reporting of metrics, overfitting across many unreported runs) than benchmarks running in controlled environments.

Several alternative incentive models exist. Competition platforms such as Kaggle demonstrate that corporate prize money can attract large-scale participation for short-term problems, but do not

5. `https://predictioncenter.org/index.cgi`

automatically generate durable benchmarks. Consortium-based models, such as MLCommons,[6] pool resources from multiple industry and academic partners to fund and maintain shared benchmarks, offering one path toward long-term sustainability. Government-sponsored evaluations (e.g., NIST benchmarks for speech and language) provide another model, where public funding ensures continuity independent of commercial interests. Non-monetary incentives—including leaderboard recognition, co-authorship on benchmark papers, and community reputation—also play a role in crowdsourced and volunteer-driven benchmark efforts.

2.3 Transparency versus integrity

The integrity of benchmarks is constantly threatened by various forms of information leakage. Over time, as more researchers interact with a benchmark, information about the test set gradually permeates the research community, compromising its ability to fairly evaluate new approaches. This "benchmark leakage" is particularly acute in natural language processing, where large language models can inadvertently memorize test set content encountered during pre-training on web-scale corpora (Bender et al., 2021).

A closely related phenomenon is leaderboard leakage: even when the test data is not memorized verbatim, accumulated knowledge from many experiments run against the same held-out set gives later entrants an implicit advantage, making the leaderboard ranking a cumulative artifact rather than a true measure of generalization (Hardt and Recht, 2022). This is exacerbated by the non-i.i.d. nature of benchmark participation: early submissions are evaluated on a genuinely unseen test set, whereas later submissions benefit from knowing which approaches have already failed.

The tension between openness and integrity also manifests in the academic–industrial divide. Open academic benchmarks promote reproducibility and broad participation but are more vulnerable to leakage and saturation. Industrial benchmarks, maintained by organizations with dedicated engineering resources, can sustain secret test sets more reliably and update them at lower cost; however, they face legitimate criticism over limited transparency and the influence of commercial interests on evaluation design. Neither model is universally preferable, and hybrid approaches—for instance, a public development set paired with a private held-out test administered through a controlled platform—offer a practical compromise (Liang et al., 2022).

2.4 Community Building and Maintenance Overhead

Perhaps one of the most overlooked difficulties in benchmark development is the significant effort required for community building and maintenance. A benchmark's success is not solely determined by its technical merits but also by its adoption and sustained use by the research community. This requires ongoing effort in documentation, support, and community engagement (Kiela et al., 2021).

The challenge of maintaining momentum goes beyond technical considerations. It involves managing a diverse user community, addressing evolving needs, and continuously demonstrating the benchmark's relevance. This overhead is particularly daunting for academic researchers who may lack the resources or institutional incentives to engage in long-term community management.

Moreover, there is a question of whether the failure of a benchmark to gain traction reflects a natural selection dynamic or a structural gap in available support. Not all benchmarks should succeed, but ensuring that high-quality ones receive adequate support remains a significant challenge (Etha-

6. `https://mlcommons.org`

yarajh and Jurafsky, 2020). Sustainable funding models—such as consortium arrangements (e.g., MLCommons[7]), government-sponsored evaluations, and platform revenue sharing—are necessary to maintain benchmarks beyond the initial publication effort.

Benchmark design also faces a tension between the breadth and interpretability of evaluation. A benchmark with too few metrics risks overlooking important capabilities and biasing comparisons toward what it does measure; one with too many metrics allows each system to claim superiority on some dimension, making overall conclusions difficult. Principled aggregation methods—such as normalized average ranks, Pareto frontiers, or composite scores with documented weighting—can make trade-offs explicit and help the community reach actionable conclusions.

Standardizing how benchmark datasets are described is one concrete step toward reducing maintenance overhead and improving interoperability. The Croissant format (Akhtar et al., 2023) proposes a machine-readable metadata schema for ML-ready datasets, covering data structure, splits, preprocessing steps, and licensing. Adopting such standards makes benchmark datasets easier to discover, reuse, and integrate across platforms, lowering the per-benchmark cost of participation and long-term upkeep.

In conclusion, addressing these challenges requires a concerted effort from the machine learning community: innovative approaches to benchmark design, sustainable resource allocation, robust methods to maintain integrity, and recognition of community-building as a first-class scientific contribution. As the field continues to evolve, so must our approaches to benchmarking.

2.5 The benchmark wish list

While we acknowledge that, due to systemic constraints, “perfect benchmarks" may not exist, we highlight in this sections ingredients of good machine learning benchmarks that we identified as practitioners.

- **Purposeful benchmark:** Well defined need, (impactful) objective pursued, and well-defined tasks.
- **Problem coverage:** Plurality of datasets and tasks as well as plurality of metrics to evaluate the various aspects of the problem at hand.
- **Bias avoidance:** Good data and task design, following best practices, as opposed to recycling of data collected under unknown or partially known circumstances.
- **Good ethical practices:** Fair compensation of crowd workers, review of design by ethical committee when human subjects are involved.
- **Large datasets:** Probing of statistical significant of results by computing error bars or test statistics using SOTA methods before releasing the benchmark.
- **Avoidance of dataset compromisation:** Keeping datasets “hidden" (revealed only to trustworthy judges) or introducing “canaries” to track use of test data for training.
- **Benchmark freshness:** Regularly renew benchmark and retire order version.

7. `https://mlcommons.org`

- **Sound aggregation of metrics:** Aggregate results in a principled way to avoid that certain aspects dominate the evaluation.
- **Code submission:** Prefer benchmarks in which the code is submitted and run in fair identical conditions on data unknown to the participants, when possible.
- **Community Self-Policing & Regulation:** Community oversight to prevent gaming or manipulation of benchmarks and adapt to evolving standards.

3 Review of recent benchmarks in machine learning

Papers with Code (PwC, `paperswithcode.com`) is a platform that connects machine learning research with practical implementation by cataloging papers, code, datasets, and benchmarks. It provides leaderboards for standardized comparisons, fostering transparency and accelerating progress in the field. As of December 2024, PwC indexes more than 128,461 evaluation records from 33,647 models across 11,005 datasets and 3,167 machine learning tasks. PwC entries are sourced through automated extraction from arXiv papers and via community submissions, where researchers directly link their papers, code repositories, and evaluation results through the platform's web interface; the underlying data is published under a CC-BY-SA license and available for download.

PwC organization is dataset-centric. PwC datasets are organized in 39 "modalities", that loosely describe the type of data as seen in Figure 1. PwC tasks are subdivided in categories as in Figure 2

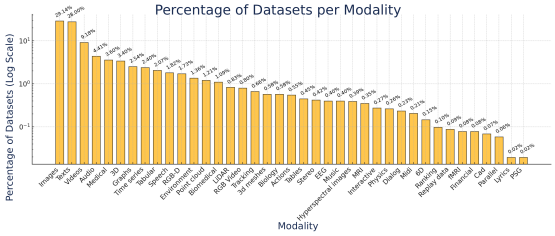

Figure 1: Percentage of datasets per modality (PwC datasets.json)

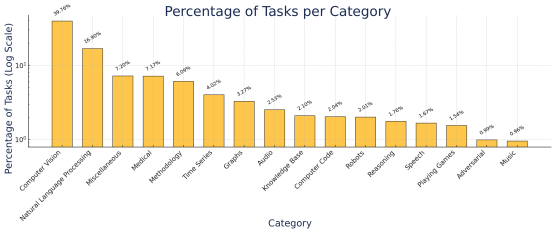

Figure 2: Percentage of tasks per category (PwC evaluation-tables.json)

The fact that both datasets modalities and tasks categories have a catchall "medical" instance is evidence of the lack of resolution around disciplines that are not focus of the machine learning community, and potentially selection bias on what gets indexed in PwC, leaving better semantic tagging of both datasets and tasks as an opportunity. That said, PwC is one of the most widely used platforms for benchmark and State-of-the-Art (SOTA) tracking in machine learning.

In PwC, benchmarks are defined as the combination of a dataset, a task, and an evaluation metric. A task in PwC corresponds to a specific problem type (e.g., image classification, question answering, named entity recognition) that can be evaluated across multiple datasets and with multiple metrics. The definition of PwC of a benchmark may not match that of a machine learning researcher. For example, BIG-Bench, a benchmark defined in paper Srivastava et al. (2022) and is composed by multiple datasets and tasks, is associated in PwC with more than 100 "benchmarks" tasks under PwC (`https://paperswithcode.com/dataset/big-bench`) as seen in Figure 3.

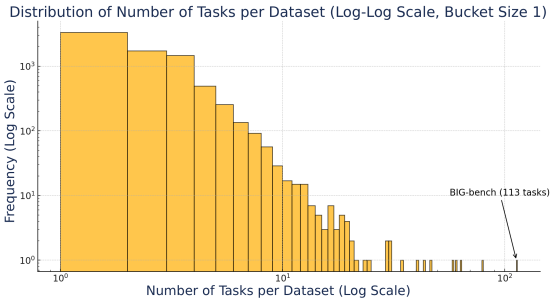

Figure 3: Distribution of number of tasks associated with each dataset (log scale)

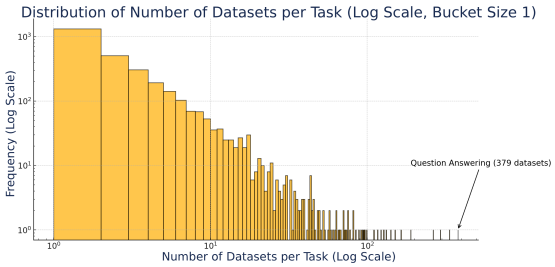

Figure 4: Distribution of number of datasets associated with each task (log scale)

Benchmarks datasets have seen a meteoric rise in early 2020 as seen in Figure 5. Benchmarks that have served the community the most with the total amount of model evaluations are aging out (e.g., ImageNet & CIFAR-10) while other for more modern categories (e.g., Reasoning) are rising quickly 6

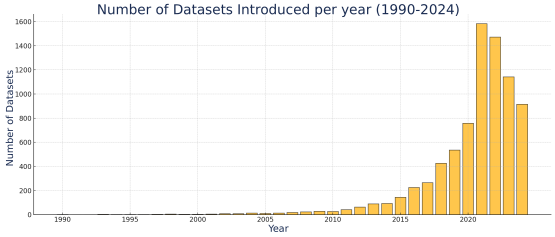

Figure 5: Number of datasets introduced in each year

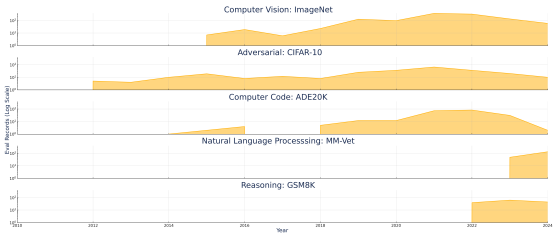

Figure 6: Number of model evaluations over time for datasets with most evaluation in each category, top-5 categories.

Recently, evaluations have been proposed that differ qualitatively from previous benchmarks in that they do not evaluate a given model on a given task but collect several distinct tasks into a single, general benchmark, where a given model (agent) has to solve all of them. For this, collections of Kaggle datasets are typically used, that are centered on a single task. Examples are AIDE [8] and AutoKaggle (Li et al., 2024). These are, in a sense, *meta*-benchmark, as they assume agents that generate code that, in turn, makes the predictions towards the individual benchmark. This, these types of benchmarks need to go beyond a single static dataset and define formal competition frameworks that agents can then operate on.

8. https://www.weco.ai/blog/technical-report

4 Towards best practices

4.1 Open access to resources

Benchmarking often faces limitations due to restricted access to key resources. A major issue is the inaccessibility of **private code** used in experiments, preventing external validation and comparison. Additionally, **private data** limits the reproducibility of results and prevents others from evaluating algorithms on the same datasets. **Trained models are often proprietary**, especially in industrial settings, further limiting comparisons. Finally, the **cost of running large-scale experiments**, particularly in terms of compute power, becomes restricting for many researchers and prevent replication.

These issues can be answered by promoting open science initiatives, such as open code and data repositories, and by crowdsourcing benchmarks. Public model zoos and public benchmarking systems, such as Hugging Face or Codabench, where data remains local but models are compared globally, could help reduce costs while maintaining reproducibility. Benchmarking platforms are critical in maintaining standards and ensuring the longevity of benchmarks. These platforms must not only provide reliable tools for comparing algorithms, but also **ensure that the data, code, models, and computing environments** remain accessible over time. Without proper information dissemination, retrieval, and maintenance, the utility of a benchmark is greatly lowered. Ensuring the perennity of benchmarking results is crucial for the long-term development of the field. Refer to Chapter 10 for a deeper dive into the role of platforms, and Chapter 13 for a detailed discussion on the importance of information retrieval and maintenance.

Developing widely accepted standards for benchmarks, including format, metrics, and protocols, can enhance interoperability across platforms. Creating centralized repositories for datasets, models, and code with persistent identifiers (e.g., DOIs) can improve long-term accessibility and reproducibility.

4.2 Avoiding bias

Benchmarking is susceptible to various forms of bias. Datasets used in evaluations may not be representative of the problem space, leading to biased results. In addition, the choice of tasks, metrics, and evaluation protocols can introduce unintentional biases. For example, choosing specific metrics might favor certain types of models over others. A significant issue is leaderboard overfitting, where models are specifically tuned to outperform on a public benchmark, but generalize poorly in real-world applications. Moreover, the tendency to report only positive results, a form of survival bias, overlooks failures, skews the perceived performance of algorithms.

Encouraging the use of diverse, representative datasets and metrics can reduce bias. This is discussed in Chapter 2 and Chapter 4. Benchmark platforms should promote a balanced evaluation with multiple tasks and objectives. Additionally, reporting negative results can reduce selective reporting.

A distinct form of contamination is benchmark poisoning, where benchmark examples appear in the pre-training corpus of a model, inflating its apparent performance. Large models trained on web-scale data are particularly susceptible, since benchmark questions and solutions are often published online. This form of contamination is difficult to detect after the fact and argues for maintaining private or regularly rotated test sets.

While benchmarks present themselves as impartial measures of progress, they are not free from bias, especially the bias introduced by their own creators. The creation of the NeurIPS Dataset and Benchmark track in 2021 introduced a peer-reviewed design of benchmarks, essential to limit bias, fostering fairer comparisons and leading to more robust, generalizable advancements.

One of the most pervasive forms of bias in science is the “inventor-as-evaluator" bias, where the organizers of a benchmark influence the outcomes by favoring certain methods or frameworks with which they are familiar. Historically, in early in the development of machine learning competitions and benchmarks, certain methods like Support Vector Machines (SVMs) or Random Forest (RF) were often favored. This happened not only because they were considered state-of-the-art but also because many competition organizers had a personal stake in promoting these techniques. This creates a skewed environment where newer, potentially better methods may not receive the attention or validation they deserve. It can be argued that the ImageNet benchmark, where neural networks (previously underutilized) gained prominence, was designed in a fair way. Had the organizers imposed constraints based on their own preferences or expertise, this monumental shift might never have occurred.

Organizers often select datasets that are easy to access or align with their own research, inadvertently giving some participants an advantage. This “selection bias" is compounded by issues such as “data leakage,” where certain features in the data inadvertently reveal information that should remain hidden during the training process. This underscores the need for competition organizers to carefully curate datasets and design their benchmarks with peer input to avoid inadvertently rewarding trivial shortcuts rather than true scientific advancement.

To address these biases, it is crucial that benchmarks be subjected to peer review. Just as scientific papers are reviewed by multiple experts to ensure accuracy, rigor, and fairness, benchmarks should undergo a similar process. Peer review can help identify potential sources of bias in task design, dataset selection, and evaluation criteria before the competition begins, ensuring that the challenge is structured in a way that promotes unbiased and meaningful progress. Moreover, peer-reviewed benchmarks ensure that no single organization, research group, or sponsor dominates the competition’s direction. Many competitions are sponsored by private companies or government agencies with their own research agendas. This financial support can influence the themes and objectives of challenges, potentially skewing them toward commercial interests rather than pure scientific discovery. Involving a broader community in the design of benchmarks, through a transparent, peer-reviewed process, can help to mitigate these external pressures and keep the focus on scientific rigor and innovation.

5 Conclusion

5.1 Summary of key points

This chapter reviewed the role of benchmarks as a foundation of reproducible machine learning. From their historical roots to their modern forms, benchmarks have evolved from simple datasets into structured scientific instruments. They serve not only to measure progress but also to frame questions, reveal limitations, and align the community around shared standards. The following points summarize the main insights discussed in this chapter.

Benchmarks are most valuable when they help answer concrete scientific questions rather than simply improving a leaderboard position. Useful benchmarks make their purpose explicit: the target capability, the assumptions being tested, and the failure modes they aim to expose (e.g., data effi-

ciency, robustness under shift, or compositional generalization). They thereby support cumulative knowledge rather than one-off records.

Responsible AI is part of the objective, not an afterthought. Beyond headline accuracy, meaningful reporting includes fairness, safety, privacy, robustness, security, and environmental impact. Provenance, demographic coverage, licensing, and compute/energy disclosure should be standard. Trade-offs must be visible: for instance, accuracy versus robustness, or latency versus energy.

Good practice keeps results trustworthy and durable. Clarity in task and metric definitions, multiple datasets or splits, strong baselines, and statistically principled aggregation with uncertainty are essential. Code submission under controlled, comparable environments improves reproducibility. Community review, explicit limitations, versioning, and deprecation policies deter gaming and help maintain relevance over time.

At the same time, the field must navigate persistent tensions: benchmarks that last long enough to be meaningful yet evolve fast enough to stay relevant; open data that ensures transparency but risks leakage; resource-intensive evaluations that advance science yet may exclude smaller institutions; and the need to build communities willing to maintain shared infrastructure over time. Addressing these tensions is as much a matter of social organization as of technical design.

5.2 Future directions for benchmarking in ML

Future benchmarks will need to balance stability with adaptability, precision with inclusivity, and automation with human judgment. They will evolve from static datasets into living ecosystems that adapt to progress while preserving comparability. The next decade of benchmarking will likely emphasize sustainability, transparency, and collective stewardship.

Registered experiments can reduce metric shopping and retrofitted narratives. By preregistering hypotheses, metrics, ablations, budgets, and stopping rules, authors make evaluations auditable and negative results publishable, which in turn enables stronger meta-analyses.

Evaluation should be holistic. Next-generation scorecards will report capability and responsibility side by side: calibration, robustness (corruptions, shifts, red teaming), efficiency (latency, memory, energy), fairness (group and individual), interpretability evidence, privacy leakage tests, and security against jailbreak and extraction. Aggregation rules and uncertainty should be explicit.

Humans remain necessary for open-ended judgments. Dynamic arenas and calibrated panels help evaluate helpfulness, harmlessness, and aesthetics, but they introduce cost, drift, and reproducibility challenges. A practical compromise is to maintain a stable, versioned core with documented rater guidelines and inter-annotator agreement, and to use model-based proxies only when regularly re-anchored to human judgments.

As models become agents, system-level evaluation will gain importance. Benchmarks should measure task completion under constraints—tools, time, budget, and risk—while tracking planning errors, tool misuse, long-horizon reliability, and recovery from failure. Sandboxed, auditable environments with event logs and reproducible agent configurations (prompts, tools, rules) enable fair comparison.

Automation will also play a growing role. Evaluator "swarms" and synthetic data can extend coverage and accelerate regression testing, but they must remain accountable and periodically regrounded to curated human judgments. Their role is to assist, not replace, human oversight.

Benchmarks themselves must evolve without breaking comparability. A versioned core (for year-over-year tracking) alongside rotating extensions (new threats, tasks, or domains) can preserve

historical continuity. Transparent changelogs and governance processes should document each revision.

Finally, benchmarking will increasingly depend on collective investment and responsible infrastructure. Sustainable funding, shared repositories, and recognition for maintenance work are needed to avoid overreliance on a few institutions. Compute-aware and carbon-aware protocols, standardized energy reporting, and equitable execution environments will make benchmarking both fairer and greener.

In short, strong benchmarks make capabilities and consequences equally measurable, reward generalization over leaderboard luck, and are designed from the outset to be reviewed, versioned, and trusted by the community.

References

Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Bambrick, et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*, 630(8016):493–500, 2024.

Mubashara Akhtar, Omar Benjelloun, Costanza Conforti, Pieter Gijsbers, Joan Giner-Miguelez, Nitisha Jain, Michael Kuchnik, Quentin Lhoest, Pierre Marcenac, Manil Maskey, et al. Croissant: A metadata format for ML-ready datasets. In *Proceedings of the NeurIPS 2023 Workshop on Datasets for Data-Centric AI*, 2023.

Suhana Bedi, Hejie Cui, Miguel Fuentes, Alyssa Unell, Michael Wornow, Juan M Banda, Nikesh Kotecha, Timothy Keyes, Yifan Mai, Mert Oez, et al. MedHELM: Holistic evaluation of large language models for medical tasks. *arXiv preprint arXiv:2505.23802*, 2025.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623, 2021.

François Chollet. On the measure of intelligence. *arXiv preprint arXiv:1911.01547*, 2019.

Francois Chollet, Mike Knoop, Gregory Kamradt, and Bryan Landers. ARC Prize 2024: Technical report. *arXiv preprint arXiv:2412.04604*, 2024.

Christopher Cieri and Mark Liberman. Issues in corpus creation and distribution: The evolution of the linguistic data consortium. In *Proceedings of the Second International Conference on Language Resources and Evaluation, LREC 2000, 31 May - June 2, 2000, Athens, Greece*. European Language Resources Association, 2000. URL http://www.lrec-conf.org/proceedings/lrec2000/html/summary/209.htm.

Mostafa Dehghani, Yi Tay, Alexey A Gritsenko, Zhe Zhao, Neil Houlsby, Fernando Diaz, Donald Metzler, and Oriol Vinyals. The benchmark lottery. *arXiv preprint arXiv:2107.07002*, 2021.

David Donoho. 50 years of data science. *Journal of Computational and Graphical Statistics*, 26 (4):745–766, 2017a. doi: 10.1080/10618600.2017.1384734. URL https://doi.org/10.1080/10618600.2017.1384734.

David Donoho. 50 years of data science. *Journal of Computational and Graphical Statistics*, 26(4): 745–766, 2017b.

David Donoho. Data Science at the Singularity. *Harvard Data Science Review*, 6(1), jan 29 2024. https://hdsr.mitpress.mit.edu/pub/g9mau4m0.

Richard O Duda, Peter E Hart, et al. *Pattern classification and scene analysis*, volume 3. Wiley New York, 1973.

Kawin Ethayarajh and Dan Jurafsky. Utility is in the eye of the user: A critique of NLP leaderboards. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4846–4853, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.393. URL https://aclanthology.org/2020.emnlp-main.393.

R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annual Eugenics*, 7(Part II):179–188, 1936. The competition protocol was designed by Isabelle Guyon. This challenge was generated using ChaLab for Codalab v1.5.

Elliot Glazer, Ege Erdil, Tamay Besiroglu, Diego Chicharro, Evan Chen, Alex Gunning, et al. FrontierMath: A benchmark for evaluating advanced mathematical reasoning in AI. *arXiv preprint arXiv:2411.04872*, 2024.

Anthony Goldbloom and Ben Hamner. Kaggle. 2010. URL https://www.kaggle.com/.

Guy Gur-Ari, Ethan Dyer, Ambrose Slone, Jascha Sohl-Dickstein, Noah Fiedel, Jaehoon Lee, Daniel Freeman, Aitor Lewkowycz, Anders Andreassen, Gaurav Mishra, Vedant Misra, Vinay Ramasesh, Noah Constant, Rosanne Liu, and et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *CoRR*, abs/2206.04615, 2022. doi: 10.48550/arXiv.2206.04615. URL https://doi.org/10.48550/arXiv.2206.04615.

Isabelle Guyon, Steve Gunn, Asa Ben-Hur, and Gideon Dror. Result analysis of the NIPS 2003 feature selection challenge. In L. Saul, Y. Weiss, and L. Bottou, editors, *Advances in Neural Information Processing Systems*, volume 17. MIT Press, 2004. URL https://proceedings.neurips.cc/paper/2004/file/5e751896e527c862bf67251a474b3819-Paper.pdf.

Isabelle Guyon, Amir Reza Saffari Azar Alamdari, Gideon Dror, and Joachim M. Buhmann. Performance prediction challenge. 2006.

Isabelle Guyon, Gavin C. Cawley, Gideon Dror, and Vincent Lemaire. Results of the active learning challenge. In Isabelle Guyon, Gavin C. Cawley, Gideon Dror, Vincent Lemaire, and Alexander R. Statnikov, editors, *Active Learning and Experimental Design workshop, In conjunction with AISTATS 2010, Sardinia, Italy, May 16, 2010*, volume 16 of *JMLR Proceedings*, pages 19–45. JMLR.org, 2011. URL http://proceedings.mlr.press/v16/guyon11a/guyon11a.pdf.

Moritz Hardt and Benjamin Recht. *Patterns, predictions, and actions: Foundations of machine learning*. Princeton University Press, 2022.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

Carlos E Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik R Narasimhan. SWE-bench: Can language models resolve real-world github issues? In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=VTF8yNQM66.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Applying and improving alphafold at CASP14. *Proteins: Structure, Function, and Bioinformatics*, 89(12): 1711–1721, 2021a.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, 2021b.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with AlphaFold. *nature*, 596(7873):583–589, 2021c.

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. Dynabench: Rethinking benchmarking in NLP. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4110–4124, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.324. URL https://aclanthology.org/2021.naacl-main.324.

R.D. King, C. Feng, and A. Sutherland. StatLog: Comparison of classification algorithm on large real-world problems. *Applied Artificial Intelligence*, 9(3):289–333, 1995. doi: 10.1080/08839519508945477. URL https://doi.org/10.1080/08839519508945477.

Doudou LaLoudouana and Mambobo Bonouliqui Tarare. Data set selection. *Neural Information Processing Systems (NIPS)*, 2002.

Pat Langley. Machine learning as an experimental science. *Machine Learning*, 3:5–8, 1988. doi: 10.1007/BF00115008. URL https://doi.org/10.1007/BF00115008.

Tony Lee, Haoqin Tu, Chi Heem Wong, Zijun Wang, Siwei Yang, Yifan Mai, Yuyin Zhou, Cihang Xie, and Percy Liang. AHELM: A holistic evaluation of audio-language models. *arXiv preprint arXiv:2508.21376*, 2025.

Ziming Li, Qianbo Zang, David Ma, Jiawei Guo, Tuney Zheng, Minghao Liu, Xinyao Niu, Yue Wang, Jian Yang, Jiaheng Liu, et al. AutoKaggle: A multi-agent framework for autonomous data science competitions. *arXiv preprint arXiv:2410.20424*, 2024.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*, 2022.

Mark Liberman. Fred Jelinek. *Computational Linguistics*, 36(4):595–599, 12 2010. ISSN 0891-2017. doi: 10.1162/coli_a_00032. URL https://doi.org/10.1162/coli_a_00032.

Kazuaki Maeda and Stephanie M. Strassel. Annotation tools for large-scale corpus development: Using AGTK at the linguistic data consortium. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation, LREC 2004, May 26-28, 2004, Lisbon, Portugal*. European Language Resources Association, 2004. URL http://www.lrec-conf.org/proceedings/lrec2004/summaries/761.htm.

D. Michie, D.J. Spiegelhalter, and C.C. Taylor. *Machine Learning, Neural and Statistical Classification*. 1994.

John Moult. A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction. *Current opinion in structural biology*, 15(3):285–289, 2005.

Adrien Pavao, Isabelle Guyon, Anne-Catherine Letournel, Dinh-Tuan Tran, Xavier Baro, Hugo Jair Escalante, Sergio Escalera, Tyler Thomas, and Zhen Xu. Codalab competitions: An open source platform to organize scientific challenges. *Journal of Machine Learning Research*, 24(198):1–6, 2023. URL http://jmlr.org/papers/v24/21-1436.html.

Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, Josephine Hu, et al. Humanity's last exam. *arXiv preprint arXiv:2501.14249*, 2025.

Ameya Prabhu, Vishaal Udandarao, Philip Torr, Matthias Bethge, Adel Bibi, and Samuel Albanie. Lifelong benchmarks: Efficient model evaluation in an era of rapid progress. *arXiv preprint arXiv:2402.19472*, 2024.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

Olawale Salaudeen and Moritz Hardt. ImageNot: A contrast with imagenet preserves model rankings. *arXiv preprint arXiv:2404.02112*, 2024.

Roy Schwartz, Jesse Dodge, Noah A Smith, and Oren Etzioni. Green AI. *Communications of the ACM*, 63(12):54–63, 2020.

Andrew W Senior, Richard Evans, John Jumper, James Kirkpatrick, Laurent Sifre, Tim Green, Chongli Qin, Augustin Žídek, Alexander WR Nelson, Alex Bridgland, et al. Improved protein structure prediction using potentials from deep learning. *Nature*, 577(7792):706–710, 2020.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*, 2022.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and policy considerations for deep learning in nlp. *arXiv preprint arXiv:1906.02243*, 2019.

Alan M. Turing. Computing machinery and intelligence. *Mind*, 59(236):433–460, 1950.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, 2018.

Jiahao Ying, Yixin Cao, Yushi Bai, Qianru Sun, Bo Wang, Wei Tang, Zhaojun Ding, Yizhe Yang, Xuanjing Huang, and Shuicheng Yan. Automating dataset updates towards reliable and timely evaluation of large language models. Conference on Neural Information Processing Systems, NeurIPS, 2024.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? *arXiv preprint arXiv:1905.07830*, 2019.

James Zou and Londa Schiebinger. AI can be sexist and racist—it's time to make it fair, 2018.