Evals Are All You Need: A Critical Analysis of AI Benchmarking in Scientific Discovery

Sujn Kumar[1]

[1]Institution Name

September 10, 2025

Abstract

The rapid advancement of artificial intelligence has been accompanied by an equally rapid proliferation of benchmarks designed to measure progress. This paper presents a comprehensive critical analysis of the current state of AI evaluation methodologies across core scientific domains including mathematics, physics, chemistry, biology, and medicine. We argue that while benchmarks have become the de facto standard for measuring AI capabilitieswith models achieving impressive scores on complex evaluations like GPQA (48.9 percentage point improvement) and SWE-bench (67.3 percentage point improvement)a fundamental disconnect persists between these metrics and genuine scientific discovery. Our analysis reveals three systemic flaws: (1) the prevalence of unrealistic benchmarks suffering from data contamination, where test sets share high similarity with training data, (2) the "black box" nature of high-performing models that undermines scientific trust and interpretability, and (3) a reproducibility crisis that makes validation of results prohibitively difficult. We examine successful paradigms like the Critical Assessment of Structure Prediction (CASP), which catalyzed the AlphaFold breakthrough, and propose that the future of AI evaluation lies not in static, skill-based assessments but in "living," iterative frameworks that evaluate an AI system's capacity for the complete scientific discovery loophypothesis generation, experimentation, observation, and refinement. The evidence suggests that current benchmarking practices often measure a model's ability to perform narrow tasks rather than assess the general intelligence required for scientific inquiry, highlighting an urgent need for evaluation methodologies that align with the true nature of scientific progress.

1 Introduction

The evaluation of artificial intelligence systems has become one of the defining challenges of our technological era. As AI models grow increasingly sophisticated and their applications expand across scientific disciplines, the question of how to measure their capabilities has taken on paramount importance. The term "benchmark," originally derived from surveying practice where it denoted a physical mark made on a stationary object of predetermined position and elevation, has evolved in the computational sciences to represent standardized tests against which performance can be measured and compared [1]. In the context of AI, benchmarks have become more than mere measurement toolsthey shape research priorities, drive investment decisions, and increasingly inform regulatory frameworks.

The history of benchmarking in artificial intelligence traces back to the earliest days of the field. The Turing Test, proposed in 1950, can be considered the first AI benchmark, attempting to evaluate whether a machine could exhibit intelligent behavior indistinguishable from a human [2]. As the field matured, more specialized benchmarks emerged. The 1980s saw the development of expert system evaluations, while the 1990s brought standardized datasets like MNIST for digit recognition. The ImageNet Large Scale Visual Recognition Challenge, launched in 2010, marked a watershed moment, demonstrating how well-designed benchmarks could catalyze breakthrough advancesin this case, the deep learning revolution sparked by AlexNet's victory in 2012.

Today's AI landscape presents a paradox. On one hand, models are achieving remarkable scores on increasingly complex benchmarks. The 2025 Stanford HAI AI Index reports that performance on demanding evaluations like MMMU, GPQA, and SWE-bench has increased by 18.8, 48.9, and 67.3 percentage points respectively over just the past year [1]. These metrics suggest rapid progress toward artificial general intelligence. Yet, a growing chorus of researchers argues that these impressive numbers mask a fundamental limitation: the disconnect between benchmark performance and genuine scientific capability [3].

1.1 The Benchmark Paradox

The central tension in contemporary AI evaluation lies in what we term the "benchmark paradox." While benchmarks are designed to be objective measures of progress, they inherently create what they purport to measure. Once a benchmark is established, the entire research community optimizes toward it, potentially at the expense of broader capabilities. This phenomenon, known as Goodhart's Law"when a measure becomes a target, it ceases to be a good measure"has profound implications for AI development in scientific domains.

Consider the evolution of natural language processing benchmarks. The field progressed from simple tasks like part-of-speech tagging to complex reasoning challenges like GPQA (Graduate-level Physics, Chemistry, and Biology Question Answering). Yet despite models achieving near-human or superhuman performance on these tests, their ability to conduct actual scientific research remains limited. As noted in recent critiques, these benchmarks often test "skill" at specific tasks rather than the "intelligence" required to efficiently acquire new skills and tackle novel problems [4].

1.2 Scientific Discovery vs. Scientific Knowledge

The distinction between possessing scientific knowledge and conducting scientific discovery is crucial yet often overlooked in current evaluation paradigms. Traditional benchmarks, from MiniF2F in mathematics to MoleculeNet in chemistry, primarily assess whether models can apply known principles to solve well-defined problems. They test crystallized knowledgethe ability to recall facts, execute algorithms, and apply established methods. However, scientific discovery requires something fundamentally different: the capacity to navigate uncertainty, generate novel hypotheses, design experiments, interpret unexpected results, and synthesize disparate observations into new theoretical frameworks.

The scientific method, refined over centuries since Francis Bacon's Novum Organum, is inherently iterative and creative. It involves not just logical deduction but also intuition, serendipity, and the ability to recognize significance in anomalies. Current AI systems, despite their impressive benchmark scores, struggle with this open-ended, exploratory aspect of science. They excel at problems with clear objectives and well-defined solution spaces but falter when faced with the ambiguity and complexity of real research.

1.3 The Stakes of Evaluation

The importance of developing appropriate evaluation methodologies extends beyond academic interest. Benchmarks have become powerful tools that shape the trajectory of AI development, influence billions in research investment, and increasingly inform policy decisions. The European Union's AI Act and the United States AI Executive Order both reference benchmark performance as criteria for regulation [5]. Companies allocate vast resources to achieve state-of-the-art results on prominent benchmarks, often using these scores as primary metrics for model releases and marketing.

Moreover, in scientific domains, the stakes are particularly high. Flawed evaluations can lead to misplaced confidence in AI systems for critical applications like drug discovery, materials design, or medical diagnosis. The Lo-Hi benchmark study revealed that widely-used molecular property prediction benchmarks like MoleculeNet suffer from severe data leakage, with 56% of test molecules having high similarity to training examples [6]. Such flaws create an illusion of progress, potentially misdirecting research efforts and delaying genuine breakthroughs.

1.4 Organization of This Review

This paper provides a comprehensive examination of the current state of AI benchmarking in scientific domains, synthesizing insights from institutional reports, academic literature, and community-driven initiatives. We begin by analyzing foundational frameworks and general-purpose benchmarks, examining how major initiatives like the Stanford HAI AI Index and NIST's evaluation standards shape the field. We then conduct deep dives into domain-specific benchmarks across mathematics, physics, chemistry, biology, and medicine, highlighting both successes and systemic failures.

Our analysis reveals three critical gaps between current evaluation practices and the requirements of scientific progress: the prevalence of unrealistic benchmarks that enable overfitting rather than generalization, the "black box" problem that undermines scientific trust and interpretability, and a reproducibility crisis that makes independent validation prohibitively difficult. We examine exemplar cases like the Critical Assessment of Structure Prediction (CASP), which demonstrated how well-designed evaluation can catalyze transformative breakthroughs, as evidenced by AlphaFold's solution to the protein folding problem.

Finally, we propose future directions for AI evaluation that move beyond static, task-specific assessments toward dynamic, iterative frameworks that better capture the essence of scientific discovery. We argue for community-wide, neutral evaluation standards, the adoption of "living benchmarks" that evolve with scientific knowledge, and a shift from benchmark-driven to problem-driven development paradigms. The path forward requires not just better benchmarks but a fundamental reconceptualization of what it means to evaluate intelligence in the context of scientific inquiry.

2 The Evolution of Scientific Benchmarking

2.1 Historical Foundations

The concept of standardized evaluation in science predates artificial intelligence by centuries. The Royal Society of London, founded in 1660, established one of the first systematic approaches to scientific validation through peer review and reproducible experimentation. This tradition of rigorous evaluation laid the groundwork for modern benchmarking practices. In mathematics, David Hilbert's famous 23 problems, presented in 1900, served as informal benchmarks that guided mathematical research for decades. Similarly, the Clay Mathematics Institute's Millennium Prize Problems, announced in 2000, continue to serve as grand challenges for the field.

The integration of benchmarking into computer science began with algorithmic complexity analysis in the 1960s and evolved through standard test suites for compiler optimization, database performance, and eventually machine learning. The UCI Machine Learning Repository, established in 1987, provided one of the first centralized collections of datasets for empirical evaluation of learning algorithms. This marked a crucial transition from theoretical analysis to empirical validation as the primary mode of progress assessment in AI.

2.2 The Modern Benchmark Ecosystem

Today's AI benchmark ecosystem comprises multiple layers of evaluation, from micro-benchmarks testing specific capabilities to comprehensive suites assessing general intelligence. The National Institute of Standards and Technology (NIST) has emerged as a key player in establishing rigorous evaluation standards through its AI Test, Evaluation, Validation and Verification (TEVV) framework [7]. Similarly, initiatives like MLCommons have created collaborative platforms where academia, industry, and government converge to develop consensus benchmarks.

The proliferation of benchmarks has been exponential. A recent survey identified over 400 distinct AI benchmarks across various domains, with new evaluations being proposed at an accelerating rate. This explosion reflects both the expanding scope of AI applications and the increasing specialization of research communities. However, it also raises concerns about benchmark fragmentation, where the multiplicity of evaluations makes it difficult to assess genuine progress or compare approaches across different research groups.

3 Methodological Foundations

3.1 Taxonomy of Evaluation Approaches

Current AI evaluation methodologies can be categorized along several dimensions. First, there is the distinction between static and dynamic benchmarks. Static benchmarks, like ImageNet or SQuAD, provide fixed datasets that remain unchanged over time. While this enables consistent comparison across years, it also allows for overfitting and data contamination. Dynamic benchmarks, exemplified by the proposed "living physics benchmark," continuously evolve with new problems and data, preventing memorization but complicating longitudinal comparison [8].

Second, evaluations differ in their scope and granularity. Component-level benchmarks assess specific capabilities like logical reasoning or pattern recognition. System-level benchmarks evaluate end-to-end performance on complex tasks. Meta-benchmarks attempt to measure learning efficiency and generalization across multiple domains. Each approach offers different insights but also has distinct limitations in capturing the full spectrum of intelligence.

Third, there is variation in evaluation protocols. Closed evaluations, where test data is hidden and submissions are limited, prevent overfitting but reduce transparency. Open evaluations promote reproducibility but risk data leakage. Adversarial evaluations, where humans actively try to find failure modes, can reveal robustness issues but may not reflect typical use cases. The choice of protocol significantly impacts both the validity and utility of benchmark results.

3.2 Metrics and Measurement

The selection of appropriate metrics remains one of the most challenging aspects of AI evaluation. Simple accuracy metrics, while easy to interpret, often fail to capture the nuances of scientific problem-solving. For instance, in drug discovery, a model might achieve high accuracy on molecular property prediction while failing to generate genuinely novel compounds-the actual goal of pharmaceutical research.

More sophisticated metrics attempt to address these limitations. The GDT-TS score used in protein structure prediction provides a nuanced measure of structural similarity. The Jaccard index employed in genomics benchmarks assesses set overlap for tasks like gene identification. However, even these specialized metrics can miss crucial aspects of scientific utility, such as interpretability, uncertainty quantification, and the ability to provide mechanistic insights.

References

[1] Stanford HAI. The 2025 AI Index Report. Stanford Institute for Human-Centered Artificial Intelligence, 2025.

[2] A. M. Turing. Computing machinery and intelligence. Mind, 59(236):433-460, 1950.

[3] The Disconnect Between AI Benchmarks and Math Research. Reddit r/MachineLearning Discussion, 2025.

[4] From MIT PhD to AI Startup: Creating Better Benchmarks for AI-Driven Scientific Research. Scientific Discovery Evaluation Framework, 2025.

[5] Can We Trust AI Benchmarks? An Interdisciplinary Review of Current Issues in AI Evaluation. arXiv:2502.06559, 2025.

[6] Lo-Hi: Practical ML Drug Discovery Benchmark. arXiv:2310.06399, 2023.

[7] National Institute of Standards and Technology. Artificial Intelligence Test, Evaluation, Validation and Verification Framework, 2025.

[8] Towards a Large Physics Benchmark. arXiv:2507.21695, 2024.