Stanford University
Human-Centered
Artificial Intelligence

Policy Brief
HAI Policy & Society
September 2025

Validating Claims About AI: A Policymaker's Guide

Olawale Salaudeen, Anka Reuel, Angelina Wang, and Sanmi Koyejo

When OpenAI claims GPT-4 shows "human-level performance" on graduate exams, or when Anthropic says Claude demonstrates "graduate-level reasoning capabilities," how can policymakers verify these claims are valid? The impact of these assertions goes far beyond company press releases. Potential claims made on benchmark results are increasingly influencing regulatory decisions, investment flows, and model deployment in critical systems.

The problem is one of overstating claims: Companies test their AI models on narrow tasks (e.g., multiple-choice science questions) but then make sweeping claims about broad capabilities based on these narrow task results (e.g., models exhibiting broader "reasoning" or "understanding" based on Q&A benchmarks). Consequently, policymakers and the public are left with limited, potentially misleading assessments of the capabilities of the AI systems that are increasingly permeating their everyday lives and society's safety-critical processes. This pattern appears across AI evaluations more broadly. For example, we may incorrectly conclude that if an AI system accurately solves a benchmark of International Mathematical Olympiad (IMO) problems, it has reached human-expert-level mathematical reasoning. However, this capability also requires common sense, adaptability, metacognition, and much more beyond the scope of the narrow evaluation based on IMO questions. Yet such overgeneralizations are common.

Key Takeaways

AI companies often use benchmarks to test their systems on narrow tasks but then make sweeping claims about broad capabilities like "reasoning" or "understanding." This gap between testing and claims is driving misguided policy decisions and investment choices.

Our systematic, three-step framework helps policymakers separate legitimate AI capabilities from unsupported claims by outlining key questions to ask: What exactly is being claimed? What was actually tested? And do the two match?

Even rigorous benchmarks can mislead: We demonstrate how the respected GPQA science benchmark is often used to support inflated claims about AI reasoning abilities. The issue is not just bad benchmarks; it is how results are interpreted and marketed.

High-stakes decisions about AI regulation, funding, and deployment are already being made based on questionable interpretations of benchmark results. Policymakers should use this framework to demand evidence that actually supports the claims being made.

In our paper "Measurement to Meaning: A Validity-Centered Framework for AI Evaluation," we propose a practical and structured approach that cuts through the hype by asking three simple questions: What exactly is someone claiming about their AI system? What did they actually test using a benchmark? And what is the evidence that their claim is valid based on that test? We focus on five key types of validity that are most relevant for evaluating AI systems today.

Policymakers must assess a growing number of claims about AI systems, including, but not limited to, their capabilities, risks, and societal impacts. We aim to provide policymakers and the public with a formalized, scientifically grounded way to investigate which claims about an AI model are supported — and which aren't. The validation framework presented in this brief is designed to evaluate all such claims. However, given the recent surge in capability claims from AI developers, this brief focuses on how to validate capability-related claims.

Model benchmarks serve as a powerful tool to evaluate AI systems. However, policymakers must work with developers and researchers to more rigorously define, report, and understand evaluations. Our targeted approach demonstrates how to use this systematic, evidence-based framework to cut through the hype and ensure policy decisions are based on solid ground and avoid tremendous miscalculations.

Benchmark performance does not always equal reliable real-world performance or trustworthy decision-making.

Introduction

Benchmarks have long helped align academia, industry, and other stakeholders around defining criteria to measure progress in specific AI systems. Evaluations have primarily aimed at measuring scientific progress — for example, performance on ImageNet, a large-scale image classification benchmark, has been viewed as an indicator of general scientific progress in AI methods. When new optimizers, architectures, or training procedures perform better on benchmarks, they also tend to lead to the development of better models across other tasks.

Today, the focus of evaluation has expanded from benchmarking methods to benchmarking models themselves, where benchmark performance is now taken as a proxy for real-world utility, often without sufficient evidence that this proxy relationship holds. Benchmark performance does not always equal reliable real-world performance or trustworthy decision-making. Model performance on a single benchmark can be overstated by conflating correlation with causation, discounting distribution shifts (where the statistical distribution of data changes between training and deployment), and downplaying the challenges with causal representation (understanding internal behavior based on observed data).

Foundation models, which can operate across diverse tasks out of the box, further complicate the translation of narrow measurements into broad conclusions. Foundation models are not trained — and rarely tested — with a specific task in mind. Instead, in the absence of such concrete use cases, model developers try to test for more general (and often abstract) skills of these general-purpose models, such as "reasoning" or "intelligence," which they assume would be helpful across a variety of tasks to predict broad and diverse downstream utility. However, designing meaningful, valid tests for such abstract capabilities is much harder than designing an evaluation that tests if the model is good at one specific task. Collectively, these trends and tendencies increase the likelihood that companies and researchers may intentionally or unintentionally overstate a model's capabilities.

Our paper builds on prior literature by explicitly arguing that validity (i.e., the degree to which evidence and theory support the interpretations of test scores) depends not just on the measurement and evaluation of a model, but also on the *claim* that is being made about its capabilities. We lay out a three-step validation process for testing capability claims about AI models (see Figure 1). While our framework can also be applied to testing other claims, such as about AI models' risks or other downstream impacts, we focus in this brief specifically on testing claims about AI model capabilities.

First, we must decide the *object* of our claim: Is it a criterion (i.e., something that can directly be measured, such as arithmetic accuracy) or a construct (i.e., something abstract that cannot directly be measured, such as "intelligence")? Second, we must explicitly *state* the claim — that is, what we want to say about the criterion (e.g., "model A can be used as a calculator") or the construct (e.g., "model A is intelligent"). Third, we must identify or perform experiments to gather *evidence* and assess whether it supports the desired claim (e.g., calculator functions may mean arithmetic accuracy, but high intelligence is unlikely) — or, in the case of reported benchmarks, decide if the benchmark truly supports our (or a model developer's) claim. Aligning what is measured, how it is interpreted, and the overarching claim is central to establishing validity.

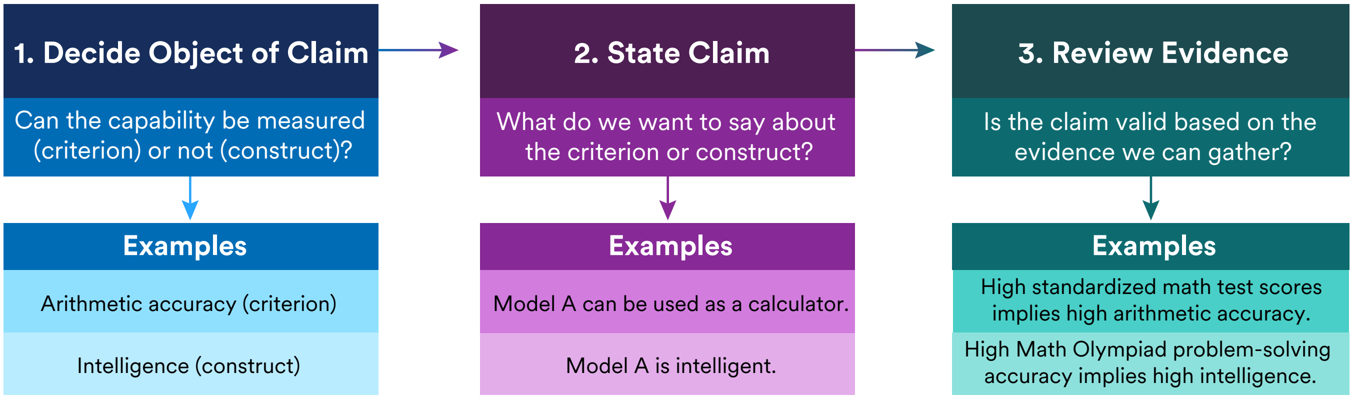
1. Decide Object of Claim
Can the capability be measured (criterion) or not (construct)?
Examples
Arithmetic accuracy (criterion)
Intelligence (construct)
2. State Claim
What do we want to say about the criterion or construct?
Examples
Model A can be used as a calculator.
Model A is intelligent.
3. Review Evidence
Is the claim valid based on the evidence we can gather?
Examples
High standardized math test scores implies high arithmetic accuracy.
High Math Olympiad problem-solving accuracy implies high intelligence.

Figure 1: Three-Step Validation Process for Testing AI Capability Claims

Applying a Claim-Centered Validity Framework for AI Evaluation

To determine to what extent evidence supports desired claims (the third step of our framework), decision-makers should consider what we consider the five most relevant validity types for AI systems and ask themselves the following questions:

- Does the evaluation cover all relevant cases? Known as *content validity*, this is at risk when important aspects of the criterion or construct to be evaluated are missing.
- Does the evaluation correlate with a known, validated standard? Known as *criterion validity*, this is at risk when the evaluation diverges from established, validated benchmarks or when the criterion itself is poorly chosen.
- Does the evaluation truly measure the intended construct? Known as *construct validity*, this is at risk when measurements fail to align with the underlying concept, different parts of the test don't relate to each other in the way the theory predicts, the test picks up on unrelated factors (like language skills or test-taking strategies) instead of the construct, or the construct is not well captured across different levels of ability.
- Does the evaluation generalize across different environments or settings? Known as *external validity*, this is at risk when tests are validated on narrow or unrepresentative populations or with testing conditions that may not reflect real-world scenarios.
- Does the evaluation consider the real-world impact of test interpretation and use? Known as *consequential validity*, this is at risk when results systematically disadvantage certain groups.

To illustrate these problems in practice, we apply our validity framework and this risk lens to real-world LLM benchmarks, including the popular Graduate-Level Google-Proof Q&A (GPQA) benchmark. GPQA relies on 448 graduate-level science questions that even PhD experts answer correctly only 65% of the time. When an AI scores well on GPQA, some AI developers claim

Claims from Graduate-Level Google-Proof Question Answering (GPQA) Benchmark Accuracy Report Card

Claims	Content	Criterion	Construct	External	Consequential
1. AI systems can accurately answer *graduate-level specialized multiple-choice questions* in biology, physics, and chemistry.	OK	OK	OK	OK	⚠
2. AI systems can accurately answer *graduate-level specialized questions* in specialized scientific domains.	⚠	⚠	⚠	⚠	⚠
3. AI systems can exhibit *general reasoning abilities* that can transfer beyond current human specialization.	⚠	✗	✗	✗	⚠

Table 1: A subjective validity scoring of the GPQA benchmark, where blue OKs indicate that the benchmark meets reasonable standards for addressing risks to validity, yellow exclamation marks signal caution, and red cross marks indicate insufficient evidence.

Claiming broader reasoning abilities requires evidence that the benchmark simply doesn't provide.

it has achieved graduate-level "scientific reasoning." But our analysis shows this benchmark actually only supports much narrower claims: The AI can answer multiple-choice questions in three specific science fields. Claiming broader reasoning abilities requires evidence that the benchmark simply doesn't provide.

More specifically, we find that GPQA, for the most part, supports the basic claim that strong performance on the benchmark means AI models can accurately answer graduate-level specialized multiple-choice questions across biology, physics, and chemistry. The benchmark is based on expert-curated questions that mirror a real-world setting, which enhances content validity by ensuring relevance and rigor across subjects and ensures external validity by demonstrating generalization to other external graduate-level assessments beyond GPQA itself. Clear guidance — for example, by the benchmark developers — for how to interpret benchmark results could help improve consequential validity by ensuring that stakeholders don't assume AI models have true general expertise in these three sciences if they score well on the benchmark.

A second possible claim that has been made from GPQA is that high scores mean models can accurately answer graduate-level questions generally across specialized scientific domains. This claim requires more evidence than is provided by the benchmark. For example, regarding construct validity, GPQA's focus on only three sciences and a multiple-choice format limits its ability to capture the overall construct of "specialized scientific knowledge" and may fail to capture deeper analytical reasoning. Including additional domains (e.g., medicine, engineering) and open-ended question formats would better capture general domain-specific scientific competence.

Finally, the claim that GPQA accuracy is evidence of general graduate-level reasoning is largely not supported. To truly support this claim, GPQA would need to, among other things, demonstrate that the benchmark covers diverse reasoning types (content validity), compares performance against other established domain-specific and independent reasoning benchmarks (criterion validity), and generalizes to reasoning tasks outside of science, such as logical puzzles or philosophical reasoning (external validity). This could be accomplished by establishing correlations between the benchmark and real graduate program exams, tracking the model's downstream performance across scientific domains, and other steps.

Without showing that GPQA performance reflects the same underlying capabilities as general reasoning, claims about an AI model outperforming scientists — or humans more broadly — based on GPQA remain unvalidated. The limits of GPQA as a scientific evaluation mechanism underscore the need to distinguish validated reasoning abilities from speculative claims.

Policy Discussion

These validity gaps aren't just academic matters — they can have significant real-world consequences. The EU AI Act, under Article 51, already uses benchmark performance to classify AI risk levels. In the United States, policymakers are similarly turning to AI evaluations as they consider applying existing or new regulations to AI systems. If benchmarks do not actually measure what matters for safety, we could end up with a false sense of security about unsafe systems or unnecessary restrictions on safe ones.

Beyond parsing claims about AI models, U.S. policymakers should also include validity specifications in pre-deployment testing requirements. Companies and researchers using benchmarks to make AI model claims often lack best practices to ensure that their claims are scientifically rigorous. A practical solution exists: Before any AI system gets deployed in critical application areas like healthcare, require developers to clearly state what capability claims their evaluations are designed to support, and why the evaluations are valid evidence of the claims. This is not about slowing down AI development; it is about making sure we are building on solid ground rather than hype.

Policymakers need a systematic way to evaluate AI claims before making regulatory decisions. Our framework provides that systematic approach — a way to demand evidence that matches the scope of the claims being made. This mapping of measurements to valid claims will become all the more important for claims impacting risk, safety, and societal impact, where policy miscalculations could have serious consequences. For example, clear performance guidelines should distinguish validated reasoning abilities from speculative claims, preventing misapplications of AI in scientific decision-making settings like hospitals. Using this framework can help ensure AI policy is evidence-aligned and that policymakers do not fall for misinterpretations or mischaracterizations of AI model evaluations.

Policymakers need a systematic way to evaluate AI claims before making regulatory decisions. Our framework provides that systematic approach.

Advancing measurement science is a crucial step to building an AI evaluations ecosystem that is scientifically grounded and can support evidence-based AI governance mechanisms. These efforts must focus on how to validate claims about AI capabilities. When implemented thoughtfully, our framework does not just prevent bad decisions — it accelerates good ones. When we can trust what AI evaluations actually tell us, we can deploy useful and beneficial AI faster and more safely.

Reference: The original article is accessible at Olawale Salaudeen et al., “**Measurement to Meaning: A Validity-Centered Framework for AI Evaluation**,” arxiv.org, May 13, 2025, revised June 26, 2025, https://arxiv.org/abs/2505.10573.

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**.

Olawale Salaudeen is a postdoctoral associate at MIT in the department of electrical engineering and computer science and the Laboratory for Information and Decision Systems.

Anka Reuel is a PhD candidate in computer science at the Stanford Intelligent Systems Laboratory and the Stanford Trustworthy AI Research (STAIR) lab, and was a graduate fellow at the Stanford Institute for Human-Centered AI (HAI).

Angelina Wang is an assistant professor at Cornell Tech and in the department of information science at Cornell University.

Sanmi Koyejo is an assistant professor in computer science at Stanford University and a faculty affiliate at Stanford HAI, and leads the STAIR Lab.

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**