Measurement to Meaning: A Validity-Centered Framework for AI Evaluation

Olawale Salaudeen[1*†] Anka Reuel[2*] Ahmed Ahmed[2] Suhana Bedi[2]
Zachary Robertson[2] Sudharsan Sundar[2] Ben Domingue[2] Angelina Wang[2,3‡]
Sanmi Koyejo[2‡†]

[1] Massachusetts Institute of Technology [2] Stanford University [3] Cornell Tech

Abstract

While the capabilities and utility of AI systems have advanced, rigorous norms for evaluating these systems have lagged. Grand claims, such as models achieving general reasoning capabilities, are supported with model performance on narrow benchmarks, like performance on graduate-level exam questions, which provide a limited and potentially misleading assessment. We provide a structured approach for reasoning about the types of evaluative claims that can be made given the available evidence. For instance, our framework helps determine whether performance on a mathematical benchmark is an indication of the ability to solve problems on math tests or instead indicates a broader ability to reason. Our framework is well-suited for the contemporary paradigm in machine learning, where various stakeholders provide measurements and evaluations that downstream users use to validate their claims and decisions. At the same time, our framework also informs the construction of evaluations designed to speak to the validity of the relevant claims. By leveraging psychometrics' breakdown of validity, evaluations can prioritize the most critical facets for a given claim, improving empirical utility and decision-making efficacy. We illustrate our framework through detailed case studies of vision and language model evaluations, highlighting how explicitly considering validity strengthens the connection between evaluation evidence and the claims being made.

1 Introduction

Suppose we observe that an AI system can solve International Math Olympiad (IMO) problems accurately (Glazer et al., 2024). Let's consider two claims about what this ability implies:

Claim 1. The system can also solve linear algebra questions from a textbook accurately.

Claim 2. The system has reached human-expert-level mathematical reasoning.

*Equal contribution.
†Corresponding authors: `olawale@mit.edu; sanmi@cs.stanford.edu.`
‡Equal senior authorship.

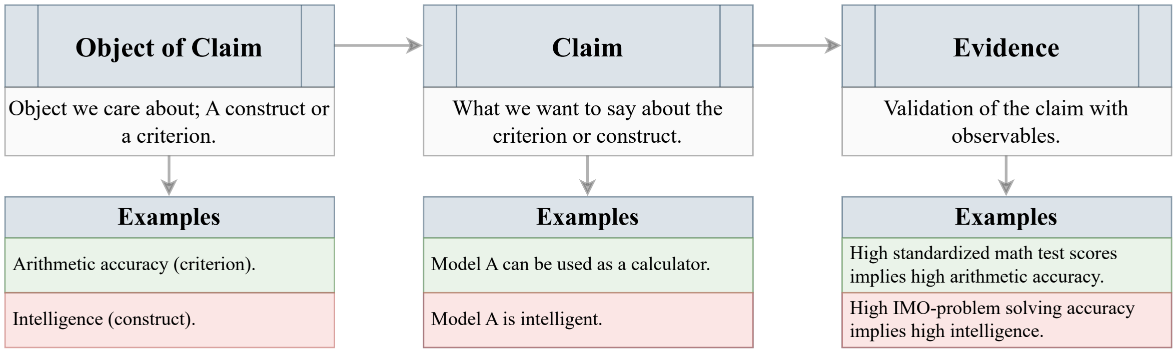

Figure 1: The three components are needed to begin the process of validation. First, we must decide what object our claim is about; is it a criterion, or is it a construct? Then, we must explicitly state the claim. Finally, we must identify or design our evidence and assess whether it supports the desired claim, i.e., do we have a valid claim based on the evidence? Here, a green background indicates that the claim-evidence pair is reasonably well supported. In contrast, a red background means the inferential leap between claim and evidence is larger and less well-supported.

Clearly, asserting Claim 2 requires a much greater inferential leap from the observed evidence (IMO accuracy) than Claim 1. If we claim that good performance on IMO problems demonstrates competence in solving linear algebra questions from a textbook, the justification may be reasonable: IMO problems often involve advanced undergraduate-level techniques, including linear algebra, so proficiency in them provides reasonable evidence for the claim. However, if we claim that the system has reached human-level reasoning, the justification is much weaker. Solving IMO problems primarily requires mathematical problem-solving, but human reasoning encompasses a broader spectrum, including common sense, adaptability, and metacognition, which IMO performance says little about. This difference highlights that we must scrutinize an evaluation and measurement in the context of the claim we wish to support. Thus, we consider validity. Validity refers to the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests (Association et al., 2014).

Historically, measurement theory studies how latent psychological concepts are quantitatively assessed, ensuring that the chosen measures accurately capture the intended constructs while maintaining validity and reliability. *Validity is not inherently a property of the measurement itself—it also depends on the context of the evaluation it enables, the claims supported by that evaluation, and the potential real-world consequences of those claims* (Messick, 1998, 1995; Shepard, 1993; Lissitz and Samuelsen, 2007; Cronbach and Meehl, 1955). While this view of validity is not universal (c.f., Borsboom et al. (2004)), it is widely held and endorsed by several professional organizations focused on measurement in educational and psychological testing Association et al. (2014). We discuss validity and its treatment in other scientific disciplines in Appendix B.

We first clarify some terms (we refer the reader to also carefully examine Table 1). In this work, measurement refers to assigning a quantitative or qualitative value to a specific property of a system (e.g., accuracy or usability), while evaluations are the broader process of interpreting these measurements to provide insights about the system. Measurements and evaluations can then support claims, judgments, assertions, and decisions about the system. Measurement instruments are tools to gather observations or assign values and include benchmarks, user studies, and expert assessments.

Table 1: Table of terms and definitions.

Term	**Definition**	**How does it relate to other terms?**	**Example**
Measurement Instrument	A tool used to gather observations or assign values (e.g., a benchmark, user study, or survey).	Underlies the act of measurement. Evaluation and claims often hinge on data obtained via instruments.	A dataset of IMO math problems (the "instrument") used to gather system accuracy scores.
Measurement	Assigning a quantitative or qualitative value to a property of a system (e.g., accuracy, usability).	Involves applying the instrument and recording results; informs subsequent evaluation.	"The system answers 15 of 20 IMO questions correctly" (accuracy = 75%).
Evaluation	The broader process of interpreting one or more measurements in context.	Translates raw measurement into insights (e.g., domain-specific analysis, comparisons to a baseline).	"Because the system can solve 75% of these IMO problems, it demonstrates proficiency in competition-level algebra."
Claim	An assertion, judgment, or decision made about the system, potentially based on evaluation results.	Draws on evaluation evidence to generalize or conclude something about the system or its capabilities.	"The system exhibits human-level math reasoning skills."
Criterion	A directly measurable or observable concept (e.g., "textbook linear algebra question-answering accuracy").	Can be measured directly and often serves as a baseline or gold standard for evaluation.	"Textbook Linear Algebra question-answering accuracy" – the system's performance on textbook linear algebra questions.
Construct	An abstract concept not directly measurable (e.g., "mathematical reasoning" or "trustworthiness").	Requires an operational definition plus proxies or indicators to measure and evaluate indirectly.	"Mathematical reasoning" – a theoretical ability captured through various problem sets and expert assessments.

For instance, we can measure accuracy on question-answer problems. However, the context of applying this measurement to IMO problem-solving problems (our measurement instrument) makes it an evaluation of a system's accuracy in answering IMO problem-type questions; we are not merely recording an accuracy score but interpreting it in a specific domain context (IMO problems) to gain some insight about the system's capabilities. To emphasize the point, while the measurement is an accuracy score, the interpretation of that measurement as an indicator of math problem-solving capability is an evaluation. Finally, one may then make claims—not necessarily correctly—about general reasoning capabilities to be supported by the evaluation.

For another example, we can measure the frequency of harmful outputs (e.g., misinformation or offensive responses) from a language model. The context of applying this measurement specifically to high-stakes medical advice scenarios (our measurement instrument) transforms it into an evaluation of the system's safety in that domain; we are not merely counting harmful responses but interpreting their potential impact within a clinical setting. From there, one might make broad claims about the system's trustworthiness or readiness for real-world deployment—claims that may be more or less justified depending on whether the measurement truly captures the range of potential harms and aligns with relevant medical standards. This full pipeline is relevant context for establishing validity.

We can measure without evaluating—for example, by collecting raw accuracy scores without concluding their implications. However, to evaluate, we must measure in some form (quantitatively or qualitatively) and then interpret those measurements in a domain-specific context. One might then ask, why measure if not to evaluate a system? We can measure as a means to develop new metrics for future evaluation; we can measure to observe or characterize phenomena before mak-

Table 2: We provide an overview of the different forms of validity considered in this work, along with key questions to ask in their assessment. The standard of evidence for validity depends on the conceptual gap between the measurement and the object of the claim, with broader gaps demanding stronger justification. Certain forms of validity, such as criterion validity, encompass multiple facets that capture different aspects of the evaluation. We adopt a view on validity closest to (Lissitz and Samuelsen, 2007). This includes aspects Cronbach and Meehl (1955) and Messick (1995)'s views on validity. Like (Lissitz and Samuelsen, 2007), we do not unify all facets of validity under construct validity.

Validity Type	Description	Example: IMO Problem Solving → Reasoning
Content Validity	Does your evaluation cover all relevant cases?	Does solving IMO problems sufficiently capture the content relevant to reasoning?
Criterion Validity	Does your evaluation correlate with a known validated standard?	Does IMO problems accuracy predict other external criteria of reasoning, e.g., common sense reasoning benchmarks?
Predictive Validity	Can your evaluation predict downstream outcomes?	–
Concurrent Validity	To what extent does your evaluation agree with another validated assessment under the exact same conditions?	–
Construct Validity	Does your evaluation truly measure the intended construct?	Does IMO problem solving capture all components of reasoning and only components of reasoning?
Structural Validity	Does your evaluation capture the structure of the construct you are measuring?	–
Convergent Validity	Does your evaluation correlate with other measures that assess the same construct?	–
Discriminant Validity	Can your evaluation differentiate between constructs that should be distinct?	–
External Validity	Does your evaluation generalize across different environments or settings?	Does excelling at IMO problems translate to solving textbook linear algebra problems accurately, where the problems are provided in different formats?
Consequential Validity	Does your evaluation consider the real-world impact of test interpretation and use?	Does emphasizing IMO problem-solving in AI development narrow research focus in ways that overlook other essential reasoning skills?

ing judgments; we can measure for calibration. We may measure accuracy to identify patterns or outliers that guide future studies without evaluating whether the system's performance is 'good' or meets real-world requirements.

A core limitation of current AI evaluation discourse is that validity, if considered, often focuses on the measurement–evaluation relationship, i.e., designing measurements that support a predefined evaluated object (Gema et al., 2024; Wallach et al., 2025), an important aspect of validity. Consequently, one may then conclude that if a measurement does not fully meet the needs of the object it was designed to evaluate, then no valid claims can be made, and no insights can be gained; but, just because IMO accuracy does not sufficiently measure reasoning does not mean that it cannot support better-scoped claims.

Additionally, *establishing validity is an iterative process (Cronbach and Meehl, 1955; Kuhn, 1997), there is no single validity checklist or iteration that can be completed to resolve all issues of validity.*

Indeed, tests for human intelligence have existed for over a century (Neisser, 1976; Binet and Simon, 1916), and the quest to improve the validity of tests is ongoing. Recognizing the limitations of measurements and evaluations, rather than outright rejecting them when certain validity criteria fall short, requires nuance; our framework enumerates this nuance. Our approach is essential for practical utility, enabling us to extract meaningful claims even from evaluations that do not rigorously satisfy all conditions of validity.

To assess the validity of a claim derived from an evaluation, Figure 1, we must:

1. Carefully consider the object of the claim. Is it a construct—an abstract object that cannot be measured directly, like 'mathematical reasoning'? Or is it a criterion—a directly measurable object, such as "accuracy on linear algebra questions from a textbook"?

2. Furthermore, does the claim refer to the same property measured, or does it extend beyond the specific evaluation to infer something about a different property? For example, one might measure IMO accuracy as part of an evaluation of mathematical ability, then use this evaluation to support a claim about textbook linear algebra question-answering, which is a different object.

3. Finally, is the claim supported directly by the measurement (e.g., IMO accuracy implies textbook linear algebra question-answering capability), or does it rely on an intermediate construct (e.g., IMO accuracy implies mathematical reasoning, which in turn implies textbook linear algebra question-answering capability)?

These distinctions determine the necessary standards of evidence required before an evaluation can meaningfully support a claim. The alignment between what is measured, how it is interpreted (evaluation), and the overarching claim is central to establishing validity.

Ensuring validity primarily requires five forms of validity from psychometrics (Thorndike, 1949; Cronbach and Meehl, 1955; Messick, 1995; Association et al., 2014; Lissitz and Samuelsen, 2007), outlined in Table 2. Additionally, Section 3 and Appendix B Table 4 enumerates tools to investigate and establish each form. While other forms of validity exist and are relevant[1], we identify this set as most relevant for current AI measurement validity gaps in Section 2. As part of our practically guided framework, we also differentiate between types of claims, where some forms of validity are trivially satisfied and only a subset needs to be focused on (Lissitz and Samuelsen, 2007).

The standard of evidence required to demonstrate validity depends on the conceptual gap between what is actually measured (and how it is evaluated) and the object of the desired claim. The greater the gap, the more arduous the task of establishing validity. Notably, different forms of validity are not independent and work together to demonstrate validity—we illustrate this in our case studies in Section 4 and Appendix D. Our proposed framework ties claims directly to the requisite standard of evidence provided by the evaluation. This alignment ensures the appropriate downstream use of evaluations while also guiding improvements that support more general claims.

This framework is pivotal because AI evaluations inform decisions with real-world consequences. For example, under Article 51 of the EU AI Act (5), benchmarks are explicitly referenced as indicators for classifying AI models according to their systemic risk. Developers of models deemed high-risk must comply with significantly more obligations. If we fail to consider validity in this context,

[1] Other forms of validity include, for example, face validity. Additional forms of validity are given in (Lim, 2024; Hughes, 2018; Borsboom et al., 2004; Association et al., 2014; Hunsley and Meyer, 2003; Reichardt, 2002; Bronfenbrenner, 1977; Neisser, 1976).

such classification may become meaningless because we cannot be certain that what truly matters—namely, the risk posed by these models—is accurately captured by the chosen measurement instruments (e.g., benchmarks). This can lead to a false sense of security. Similarly, measurement instruments are often used within organizations (Hardy et al., 2024) to guide resource allocation and further training aimed at improving a model's capabilities. Yet, if the chosen instrument does not accurately measure the capability developers care about, additional training may simply become an exercise in "teaching to the test" (Jennings and Bearak, 2014) rather than leading to genuine improvements in the model.

Recognizing these challenges, we propose a structured framework to assess the validity of AI assessments, ensuring appropriate use and interpretation. Specifically, **our contributions** are:

1. We examine and identify limitations in how the relevance of different forms of validity has co-evolved with the progress of AI and the corresponding norms and practices of evaluation. (Sec. 2)

2. We enumerate common risks to validity and practical operationalization to investigate and support distinct forms of validity. (Sec. 3)

3. We propose a practical and structured claim-aware framework for identifying the necessary evidence to establish the validity of claims based on AI evaluations. We also enumerate adoptable practices to demonstrate validity. (Sec. 4)

4. We illustrate our framework through vision and language evaluation case studies, providing concrete, prescriptive examples of validating claims based on evaluations. (Sec. 5 and App. D)

2 Validity Gaps in Current AI Evaluations and Related Work

AI evaluation has evolved alongside the systems it measures, but the distance between what we evaluate and what we claim about real-world utility has grown (Appendix C), revealing gaps across all five major forms of validity, content, criterion, construct, external, and consequential (Table 2). These gaps reflect shifts not only in how we evaluate but in what we claim.

Initially, evaluations emphasized held-out test data drawn from the same distribution as training data, which is often called the i.i.d. setting, supporting content validity, where the test environment closely matched the training one. As pretraining became the norm, evaluation shifted to fine-tuning large models like those trained on ImageNet (Deng et al., 2009; Russakovsky et al., 2014; Kornblith et al., 2018; Recht et al., 2019) and measuring downstream performance. This introduced criterion validity, where success on downstream tasks was taken as evidence that pretraining had captured useful representations (Mikolov et al., 2013; Brown et al., 2020; Radford et al., 2021).

As concerns about spurious correlations driven decision-making, distribution shifts (Bai et al., 2025; Salaudeen and Koyejo, 2024; Lopez-Paz et al., 2016; Xiao et al., 2020; Arjovsky et al., 2019; Rosenfeld et al., 2020; Koh et al., 2020; Gulrajani and Lopez-Paz, 2020), and causal representations (Schölkopf et al., 2021; Gichoya et al., 2022) rose, so too did other forms of validity. Recent work has increasingly emphasized external validity, whether evaluation results generalize beyond the training distribution (Salaudeen and Hardt, 2024), consequential validity, whether model deployment leads to desirable outcomes (Wang and Russakovsky, 2023), and construct validity, whether an evaluation actually captures the concept we think it does (Bell et al., 2024; Salaudeen et al., 2025).

Still, evaluation remains largely benchmark-driven, often in service of leaderboard-based progress (Hardt and Recht, 2021; Orr and Kang, 2024). This kind of evaluation is not without merit: when optimizers, architectures, or training procedures improve benchmark performance across multiple tasks, they also tend to improve performance on another new and real-world task, a sort of criterion validity (Blum and Hardt, 2015; Kornblith et al., 2018; Salaudeen and Hardt, 2024). Moreover, these shared benchmarks have helped align academia, industry, and other stakeholders on a criterion to measure progress (Donoho, 2023; Recht, 2024), for instance, ImageNet accuracy (Russakovsky et al., 2014). However, benchmark performance does not always translate to reliable real-world performance or trustworthy decision making (Hardy et al., 2024).

The rise of foundation models, which can operate across diverse tasks out of the box, further complicates this issue. Traditional evaluation methods increasingly fail to capture real-world AI behaviors that require investigating abstract capabilities like intelligence and reasoning to predict broad and diverse downstream utility (Wu et al., 2023; Wan et al., 2024; Mirzadeh et al., 2024).

Narrow datasets used for "general-purpose" evaluation raise content, construct, and external validity concerns, especially for constructs like reasoning (Bostrom et al., 2020; Alaa et al., 2025). Furthermore, these evaluations lack criterion validity and fail to predict criteria of real-world utility (Hardy et al., 2024), and the socio-technical gap between evaluation results and real-world needs undermines consequential validity (Liao and Xiao, 2023; Barocas et al., 2023). Consequently, overgeneralized results erode evaluation credibility (Raji et al., 2021).

Important prior work has demonstrated the need for validity frameworks (Jacobs and Wallach, 2021; Saxon et al., 2024; Subramonian et al., 2023; Xiao et al., 2023; Blodgett et al., 2021; Coston et al., 2023; Xiao et al., 2024; Reuel et al., 2024). However, much of it has focused on validity in the context of the limitations of measurements, focusing on conditions for perfect measurements of nicely defined concepts, which is far from practice. METRICEVAL (Xiao et al., 2023) raises validity concerns stemming from vague benchmark articulation and repurposed datasets in 16 natural language generation metrics. Liu et al. (2024) further challenged benchmarks' ability to measure intended constructs, proposing the Evidence-Centered Benchmark Design (ECBD) framework to ensure rigorous metric selection. However, these works focus on designing new measurement instruments for evaluations. While developing better measurement instruments for better evaluations is also important, and our framework also applies to this task, we find it of practical value to understand what claims can be made from existing evaluations and evidence, given the intractability of creating tailored evaluations for each claim, and the already unwieldy amount of existing benchmarks (Copilot, 2024).

The important work of (Chouldechova et al., 2024; Wallach et al., 2025) applied (Adcock and Collier, 2001)'s measurement theory, critiquing ML evaluations for conflating systematizing a background (conceptual and informal) concept with operationalizing (measuring) a systemized (formalized and structured) concept[2].

Our work complements this literature by explicitly identifying that validity depends not only on the measurement and evaluation but also on the claim intended to be made. Building on Wallach et al. (2025)—who underscore the importance of explicit systematization needed in AI, where concepts often emerge from practice rather than theory—our work clarifies this process in the context of nomological networks (Cronbach and Meehl, 1955). These networks represent not only

[2]According to Adcock and Collier, a background concept is a "broad constellation of meanings and understandings associated with [the] concept," and systematization describes the process of refining and explicitly defining a concept to create a structured and consistent foundation for measurement and analysis (the systematized concept) while operationalization the process of transforming a systematized concept into measurable indicators ().

the relationship between the background concept and systematized concept but also the broader relationships to other background and systematized concepts. In the sense of the Duhem-Quine thesis[3], a nomological network serves as a map of empirical and theoretical relationships, helping manage the holistic nature of scientific testing by clarifying how constructs relate to each other as well as observable evidence. Consequently, we expand upon Wallach et al. (2025)'s view of systematization by arguing for the importance of broader nomological networks beyond just locally specifying which definition of a background concept will be used. Our view subsumes both the view that background concepts (constructs) merely serve as a mechanism for conceptual agreement and the view that they are fundamental properties inherent to AI systems to be discovered and refined through the scientific process. This process must be considered in the emerging science of AI evaluations (Weidinger et al., 2025; Hardt, 2025).

Ultimately, our framework takes a practical approach, emphasizing that validity is not only a property of measurement and evaluation. As Cronbach and Meehl emphasize (Cronbach and Meehl, 1955): "In one sense, it is naive to inquire 'Is this test valid?' One does not validate a test, but only a principle for making inferences. If a test yields many different types of inferences, some can be valid and others invalid." We further enumerate risks, tools, and evidence exemplars to assess whether evaluations meet appropriate validity standards in Section 3 and Table 4.

3 Risks to Validity and Operationalizable Strategies for Mitigation

In this section, we examine common risks to valid claims as a function of limitations in establishing some form of validity and discuss existing tools and methodologies for assessing and strengthening validity in AI assessment. Appendix A Table 4 categorizes in detail key risks, investigation tools, and evidence exemplars across multiple forms of validity in assessment. We summarize here. Importantly, this section makes clear that general-purpose *'benchmarks' are currently an insufficient sole evaluation mechanisms for the real-world utility of AI systems.*

Risks to content validity include coverage deficiency, where important aspects of the construct are missing, and construct irrelevance, where extraneous factors influence scores (Association et al., 2014; Messick, 1995). Imbalanced content can lead to assessments overemphasizing certain skills while neglecting others. These issues can be examined through expert review, adversarial scrutiny, and synthetic data generation, with supporting evidence from explicit content mapping and coverage analysis.

Risks to external validity include sample bias, where the test is validated on a narrow or unrepresentative population (Henrich et al., 2010), and unrealistic testing conditions, which may not reflect real-world scenarios (Donald T. Campbell, 1963). Temporal variability and interaction effects can also distort results if performance shifts over time or due to specific environmental factors (Andonov et al., 2023). These issues can be investigated through stress testing, A/B testing, transfer testing, and population-stratified assessments, with evidence from performance comparisons across different conditions and sensitivity analyses.

Risks to criterion validity include criterion contamination, where extraneous factors influence assessment, and criterion deficiency, where relevant aspects of performance are omitted (Brogden and

[3] The Duhem-Quine thesis emphasizes that scientific claims are interconnected, meaning that rejecting or modifying one hypothesis affects others within the theoretical framework.

Taylor, 1950; Austin and Villanova, 1992). Restricted range limits the ability to detect meaningful relationships if all scores are too similar. These issues can be addressed through real-world longitudinal studies, validated criterion studies, and behavioral testing, with evidence from correlations with gold-standard benchmarks and predictions of real-world utility.

Risks to construct validity can come from structural, convergent, and discriminant validity risks. Structural validity is compromised by poor factor structure, where test items fail to group in expected ways (Clark and Watson, 1995; Elhami Athar, 2023), and complex measurement range, where constructs are not well captured across different levels of ability (Messick, 1995). Convergent validity can suffer from high measurement error (Cheung et al., 2024), which reduces reliability, while discriminant validity can be compromised by construct overlap, where different abilities are not clearly distinguished (Shaffer et al., 2016). These risks can be investigated using hypothesis testing, factor modeling, and benchmark suites, with supporting evidence from item-test correlations and demonstrated non-significant overlap with unrelated constructs.

Risks to consequential validity include bias and fairness issues, where results systematically disadvantage certain groups (Meredith, 1993; Messick, 1995; Randall, 2023). While bias and fairness can themselves be constructs of interest, they are also important to consider in any measurement. Further, unintended incentives can distort behavior if assessment criteria encourage gaming rather than genuine learning (Nichols and Berliner, 2007). Policy consequences may emerge if flawed assessments influence high-stakes decisions. These risks can be assessed through anticipatory ethics methods (Umbrello et al., 2023), societal impact audits, and ethical stress testing, with evidence from stakeholder feedback, improvements in fairness and reliability, and documented real-world impacts.

While this framework highlights key risks and mitigation strategies, additional risks may arise in different contexts, necessitating continuous assessment and refinement.

4 A Framework for Claim-Centered Validity Assessment in AI Evaluation

In this section, we categorize when and how different forms of validity are most critical for supporting a claim with measurements and evaluation, Figure 2. While we maintain that all forms of validity are always necessary, some may be trivially satisfied depending on the measurement, evaluation, and claim context. Rather than applying uniform scrutiny to all forms of validity, we account for context-dependent nuances that make certain forms particularly significant in some cases; this is distinct from previous work (Wallach et al., 2025).

Our framework, as described by Figure 2, is similar to Lissitz and Samuelsen (2007), who challenge the dominant paradigm that all forms of validity are subsumed by construct validity, as advocated by (Messick, 1995). They argue that this approach can obscure critical distinctions between types of test use and the kinds of evidence needed to justify them. They argue that by folding all these aspects into a single, construct-validity-centric framework, the field risks undervaluing empirical utility and pragmatic decision-making goals. For example, if an IMO accuracy reliably predicts college linear algebra question-answering accuracy, it may be considered valid for such uses, even if the construct it measures (e.g., "mathematical reasoning" or "pattern recognition") is poorly defined or theoretically contested. Lissitz and Samuelsen (2007) call for a functional approach to validation, one that aligns the type of evidence required with the specific role the test plays. This perspective opens the door to broader and more flexible evaluation strategies, particularly in

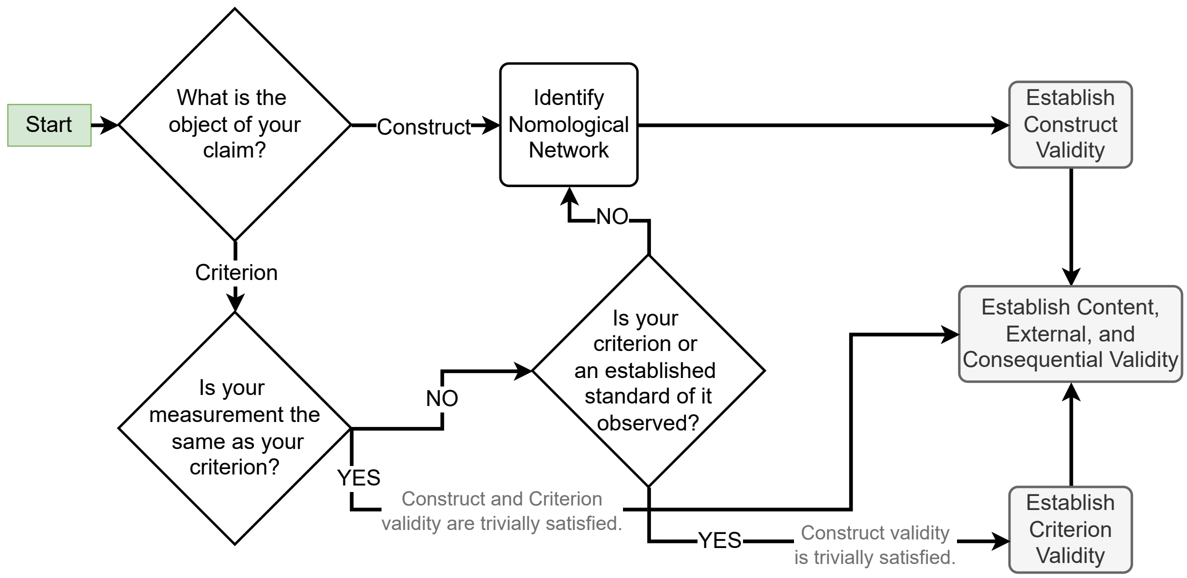

Figure 2: Decision process for establishing validity. For the decision processes that do not directly go through establishing construct or criterion validity, our argument is not that those forms of validity are irrelevant, but rather that they may be trivially satisfied in the context of the measurement, evaluation, and claim.

applied contexts where operational outcomes matter more than construct fidelity.

Specifically, Lissitz and Samuelsen (2007) consider three common relevant activities, which are related to the structure of our framework.

- **Utility Determination.** Validity should be determined by how useful a test is for its intended purpose. A valid test enables appropriate and effective decisions in real-world contexts (Sec. 5.1-5.2).
- **Theoretical Support.** A test is valid insofar as it supports the theory guiding its development or use (Sec. 5.3).
- **Impact Evaluation.** Validity must account for the outcomes of using a test, including whether the decisions it informs lead to beneficial, fair, or intended consequences (Sec. 5.4).

Our frameworks shares this perspective and develops a strategy to implement it for AI systems in the following.

Importantly, a claim can (and should (Wang et al., 2024)) be supported by many evaluations and measurements. However, for simplicity and without loss of generality, we focus on a single measurement and evaluation.

Recall that the object of a claim can be a criterion (directly measurable) or a construct (abstract and not directly measurable). Our primary considerations for investigating and prioritizing forms of validity are determined by the following (Figure 1):

1. Is the object of the claim a criterion (e.g., linear algebra textbook question accuracy) or a construct (e.g., reasoning ability)?

	Claude 3.5 Sonnet	Claude 3 Opus	GPT-4o	Gemini 1.5 Pro	Llama-400b (early snapshot)
Graduate level reasoning *GPQA, Diamond*	**59.4%*** 0-shot CoT	**50.4%** 0-shot CoT	**53.6%** 0-shot CoT	—	—
Undergraduate level knowledge *MMLU*	**88.7%**** 5-shot	**86.8%** 5-shot	—	**85.9%** 5-shot	**86.1%** 5-shot
	88.3% 0-shot CoT	**85.7%** 0-shot CoT	**88.7%** 0-shot CoT	—	—

Figure 3: A real-world example of a developer presenting the Graduate-Level Google-Proof Q&A Benchmark (GPQA) as a proxy for the graduate-level reasoning capabilities of their system (top left). Additionally, SWE-bench Jimenez et al. (2023) is used as a proxy for agentic coding, and τ-bench (Yao et al., 2024) as a proxy for agentic tool use (Anthropic, 2024; Rein et al., 2023).

2. Furthermore, is the measurement the same as the object of the claim (e.g., evaluating IMO accuracy when IMO accuracy is also the object of the claim)?
3. Finally, if the measurement and object of the claim are different, does the measurement directly imply the claim, or does it require a mediating construct (e.g., does IMO accuracy imply linear algebra textbook question accuracy, OR does it imply mathematical reasoning ability, which implies linear algebra textbook question accuracy)?

The five forms of validity we foreground are relevant in different ways. Ideally, we should validate a claim by directly measuring the object it is about in the context in which we want to make the claim. However, this is often not possible. In any case, we must always validate content validity and external validity. Additionally, when we perform a measurement but the object of the claim is a different criterion (e.g., evaluate IMO problem solving → make claims about linear algebra textbook problem solving), criterion validity is most important. Criterion validity ensures that the measurement reliably predicts the object of the claim (predictive validity) or an established external standard of the object of the claim (concurrent validity). When neither the object of claim nor an established external standard is available, we may validate the claim through an intermediate construct (evaluate IMO problem solving → infer mathematical reasoning → make claims about linear algebra textbook problem solving), requiring construct validity of tests of mathematical reasoning with respect to the downstream use of linear algebra textbook problem solving.

When the object of the claim is itself a construct, and we directly measure and evaluate its proxies (e.g., evaluate IMO accuracy → make claims about mathematical reasoning), construct validity is essential to determine whether the measurement genuinely measures the intended construct rather than an unrelated or superficial correlation. This requires *structural validity* to ensure that the measurement reflects the internal structure expected of the construct (e.g., subskills of reasoning are coherently represented), *convergent validity* to demonstrate that the measurement correlates with other measures of the same or closely related constructs (e.g., aligns with expert assessments of reasoning), and *discriminant validity* to show that it does not correlate with measures of unrelated constructs (e.g., verbal fluency or memorization), guarding against misleading associations. This is also necessary when we aim to validate a claim about a construct with measurements and evaluations of proxies of other constructs (e.g., evaluate IMO problem solving → infer logical reasoning

Table 3: A Graduate-Level Google-Proof Q&A Benchmark (GPQA) (Rein et al., 2023) Application. A subjective score for validity—the standard for "reasonable" is demonstrating that obvious risks to invalidity are addressed: 🆗: reasonable; ⚠️: proceed with caution; ❌: insufficient. Even for a score of "reasonable," there will be weaknesses in the evidence. The score is given because the strengths outweigh the weaknesses in terms of determining the validity of the claim from that evidence. This is never a binary classification nor complete, and should rather be a cyclic process—for instance, as our forms of what constitutes graduate-level chemistry may evolve over time and from school to school.

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.	🆗	🆗	🆗	🆗	⚠️
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.	⚠️	❌	❌	❌	⚠️

→ make claims about mathematical reasoning). In this case, a Nomological Network—an understanding of the relationship between logical and mathematical reasoning and their observables—is paramount.

A claim about a construct cannot be validated in isolation—instead, it gains meaning and validity through its relationships with other constructs and observable measures through the scientific process. Cronbach and Meehl's nomological network (Cronbach and Meehl, 1955) provides a rigorous, albeit historically challenging to operationalize, way to reason about constructs within a broader abstract and empirical system, allowing for more scientifically robust validation. A nomological network is a conceptual framework that maps the relationships between constructs and criteria. Figure 4 gives an example from human psychology. An explicit nomological network for AI constructs, while vital, remains a missing piece in efforts toward valid AI evaluations, limiting the establishment of validity when constructs are involved. Although a detailed treatment of nomological networks is beyond the scope of this work, we emphasize their importance in establishing validity and explicitly indicate where they are necessary in our framework. We refer the reader to Cronbach and Meehl's seminal work for more detail (Cronbach and Meehl, 1955).

Next, we illustrate our framework for determining validity.

5 Application of our Framework

We focus in the main text on GPQA (Graduate-Level Google-Proof Question Answering) accuracy as our measurement and evaluation. We then investigate claims of varying generality commonly made from this evaluation (Buntz, 2025; Rein et al., 2023). Additional examples are in Appendix D. This clarifies that a given measurement may not support broad claims, yet it can still be highly useful for supporting more narrowly defined ones. This adds necessary nuance to the discourse on validity in AI assessment.

Table 3 summarizes the following example of applying our framework to assess the validity of

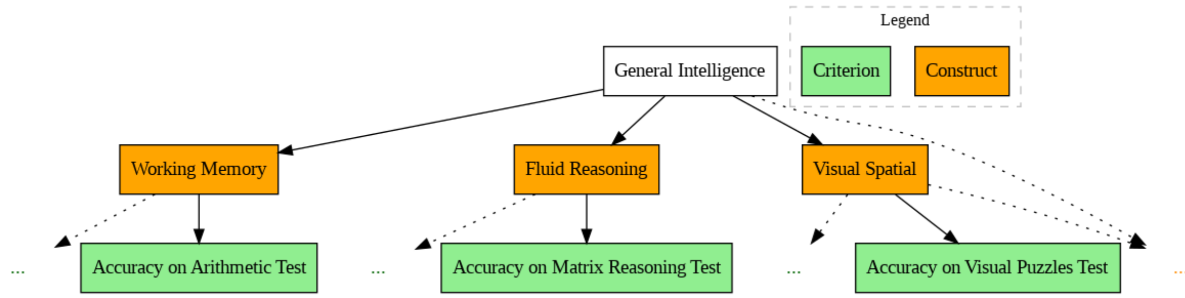

Figure 4: This figure illustrates a nomological network for human general intelligence according to the Wechsler Intelligence Scale, adapted from Canivez et al. (2017). *We note that this breakdown of general intelligence is for illustrative purposes only; it does not necessarily translate to artificial general intelligence.* The network consists of background concepts (blue w/ rounded corners) linked by hypothesized associations, reflecting abstract expectations. Observable indicators (green w/ sharp corners) represent criteria, measurable variables used to assess the constructs. Establishing a robust nomological network is critical to ensuring construct validity by demonstrating abstract coherence, convergent validity, and discriminant validity within empirical research.

claims from GPQA multiple-choice question-answering accuracy. We supplement this section with detailed case studies in the context of evaluating popular vision and/or language AI systems in Appendix D.

Sources of Validity Evidence. While we restrict ourselves to the evidence of validity provided in the GPQA paper for the previous analysis for brevity and simplicity, establishing validity can (and should) be done across multiple asynchronous studies and various stakeholders.

Furthermore, while we can begin with a claim and use our framework to design and identify the necessary types of measurements and evaluations to support it, i.e., the scientific process; we can also do this in reverse. The latter is increasingly common in practice, given the exploratory nature of many of our studies of AI systems. Adcock and Collier (2001)'s framework for measurement validity begins with the construct and reasons about developing valid measurements, as does Wallach et al. (2025)'s framework for evaluating generative AI. Recognizing the iterative nature of validation and that all measurements and evaluations are limited to some degree, our framework allows for the possibility of deriving valid claims even from imperfect measurements. Such claims are necessarily narrower in scope and generality than those supported by more complete or ideal measurements. We see this flexibility as a key practical advantage, particularly in modern contexts where the stakeholders conducting measurements and those making claims from evaluations may differ, and where some information needed to establish full validity may be unavailable.

We first consider the setting where the object of the claim is a criterion, and the desired claim in this setting looks like this: "A higher measurement score predicts a higher/lower criterion score." "A higher GPQA accuracy predicts a higher PhD qualifying exam accuracy." In this case, the measurement and the criterion object of the claim can be (1) identical (GPQA accuracy vs. new GPQA accuracy), (2) proxies of the same underlying construct (GPQA accuracy vs. general scientific question-answering accuracy), or (3) proxies of two different but related underlying constructs (GPQA accuracy (scientific reasoning) vs. medical diagnostic accuracy (medical reasoning)). (4) We then consider the setting where the object of the claim is a construct.

In each case, we must justify why the measurement supports the claim.

5.1 Scenario 1: The object of the claim is a criterion and is measured as evidence

In scenario 1, we are primarily concerned with content and external validity, i.e., does our measurement cover the relevant content of the criterion, and does it generalize to relevant contexts beyond that of the measurement? Construct validity and criterion are trivially satisfied as a consequence of directly measuring and evaluating the object of the claim. This could be because the hard work of systematization and operationalization of the construct we would have otherwise attempted to measure and evaluate has already been done (Adcock and Collier, 2001; Wallach et al., 2025).

Identical Measurement and Criterion. *Example.* Our object of claim is a criterion: multiple-choice questions accuracy in physics, chemistry, and biology[4]. We then aim to support claims about an AI system's accuracy on such questions by measuring and evaluating the system's accuracy on the GPQA dataset; we must:

- Establish content validity: GPQA has expert-curated questions, which enhance content validity by ensuring relevance and rigor across biology, physics, and chemistry, with the performance gap between experts and non-experts indicating effective assessment of specialized knowledge. However, the criteria may inadvertently exclude certain relevant topics, potentially skewing subfield representation. Systematic content mapping and expert diversity analysis can strengthen validity by ensuring comprehensive coverage and mitigating selection biases. In modern AI, red-teaming can also help identify the coverage gaps that hinder content validity (Perez et al., 2022).

If content validity holds and one does not expect that the context of measurement is different than the context of the claim, then the claim can be supported by the measurement. However, if the claim must hold in a different context than the measurement, one must also:

- Establish external validity: The GPQA measurement reflects real-world conditions, with human experts developing questions and a measurement format aligned with academic multiple-choice assessments, so the context of measurement is aligned with the context of the claim in this sense. However, the human assessment may not generalize beyond the measurement context, and without comparison to other multiple-choice science tests, its generalizability remains unverified. To strengthen external validity, validation against diverse question formats, question types, and other variations of context is necessary.

This setting is commensurate with traditional AI benchmarking practices. Many AI benchmarks have focused on these forms of generalization, including classical generalization (Vapnik and Chervonenkis, 1971) and out-of-distribution generalization (Shimodaira, 2000). By ensuring strong content and external validity, such benchmarks provide a solid foundation for validating claims for known and directly measurable criteria.

[4]Thresholding is commonly used in real-world decision-making to transform continuous measurements (e.g., confidence scores) into binary or categorical outcomes (e.g., pass/fail, high risk/low risk). The choice of threshold can profoundly affect both evaluation and claim validity: even a perfectly measured property may lead to an invalid claim if the threshold does not align with the intended context or the actual consequences of misclassification. This is known in psychometrics as standard testing (Cizek and Bunch, 2007).

5.2 Scenarios 2-3: The object of the claim is a criterion, but a different object is measured as evidence

For 2-3, different criteria that are either proxies of the same or different but related mediating constructs. Ideally, we additionally directly establish criterion validity. That is, establish that the object that is measured is predictive of the desired criterion or an established standard. Then, the existence of these mediating constructs may inform how we establish criterion validity, but we do not need to reason about them directly to make valid claims.

Example. Our object of claim is general scientific question-answering, and we want to use GPQA accuracy as evidence. We still need to demonstrate that the measurement covers relevant content and generalizes to all the contexts we want the claim to hold (content and external validity). However, we must additionally establish that the measurement of GPQA accuracy is predictive of the scientific question-answering criterion or a validated standard, i.e.:

- Establish Criterion Validity. Human expert accuracy provides a strong external criterion and validated standard, supporting concurrent validity, while the AI-expert performance gap reinforces the benchmark's credibility. However, there is no evidence of predictive validity, as accuracy has not been tested against future performance on specialized assessments, and concurrent validity remains incomplete without correlations to established external measures of expertise, such as standardized exams in other fields. To strengthen criterion validity, correlations should be established with real graduate program exams for concurrent validity, and predictive validity studies should track the system's downstream performance across scientific domains.

However, if criterion validity is implausible in this way, we may attempt to leverage our understanding of the underlying structure in constructs and their known mapping to observables, when it is available, to establish validity, i.e., a nomological network. Importantly, such a nomological network often does not exist in the current paradigm of AI assessment.

When a nomological network is unknown, establishing validity becomes significantly more difficult, as there is no agreed-upon basis for interpreting how abstract constructs like 'scientific reasoning' map to measurable criteria (e.g., 'GPQA accuracy'). In such cases, evaluations risk being narrow or misleading—a system might excel at scientific reasoning in physics yet still lack scientific reasoning in psychology, and evaluators could erroneously assume success in one facet implies the same in another. Without explicit connections between sub-constructs and corresponding measurements, conflicting results may emerge, and different assessments might rely on unfounded inferences about a system's capabilities. This lack of structure not only obscures whether a measurement provides meaningful evidence for a given claim but also undermines the reliability of validity assessments, leaving practitioners vulnerable to inflated claims and misguided deployment decisions.

When such a network is available, to establish construct validity, we utilize its facets: structural, convergent, and discriminant validity:

- Establish Construct Validity: For brevity, please refer to the subsequent discussion in section 5.3 on when the object of the claim is a construct.

Importantly, when the measurement and criterion are proxies of different constructs, we must also validate relationships between constructs in addition to their relationships to observables. Doing this also requires knowledge of a nomological network. For example, in Figure 4, the accuracy on the arithmetic test and accuracy on the matrix reasoning test must go between working memory and fluid reasoning.

Example. Suppose we want to make a claim about AI systems' reasoning about medical diagnosis based on their reasoning ability. To evaluate this, we must first define what reasoning entails. Suppose a model evaluator interprets reasoning as scientific reasoning and uses the GPQA benchmark to measure it. However, the object of the claim is most related to medical reasoning. At this stage, defining reasoning in this way is neither inherently valid nor invalid.

Now, suppose we want to claim that strong GPQA performance translates into accurate medical diagnosis; this requires several inferential steps. GPQA (insufficiently) assesses scientific reasoning, while medical diagnosis likely relies on medical reasoning, potentially a different subspace of reasoning, according to one's nomological network. Establishing structural validity requires examining whether GPQA captures the key components of general reasoning relevant to medical decision-making. Without showing that GPQA performance reflects the same underlying capabilities as medical reasoning, claims about AI outperforming physicians based on GPQA remain unverified. Establishing convergent validity through latent variable modeling and item-response theory can help demonstrate if the measurement captures variance in the latent subspace shared between the two constructs that determine the outcome criterion.

5.3 Scenario 4. The object of the claim is a construct

In many cases, we want to validate a claim about a construct by evaluating its proxies. This looks like: "A higher measurement score implies a higher latent capability, e.g., GPQA accuracy to scientific reasoning." Then, construct validity is paramount.

Example. Suppose the object of the claim is general reasoning and we want to make a claim about a system's general reasoning ability by measuring GPQA accuracy. Here, we must establish all five of our forms of validity, especially construct validity (recall it is composed of structural, convergent, and discriminant validity):

- Establish Construct Validity: Performance on GPQA aligns with success in structured question-answering tasks, suggesting some reasoning component. However, structural validity is unclear, as the test may not sufficiently capture the rank of reasoning. Convergent validity is unverified since GPQA accuracy has not been correlated with other explicit reasoning assessments, e.g., interactive human evaluation. Discriminant validity is also uncertain, as it remains unclear whether GPQA measures genuine scientific reasoning or simply domain-specific knowledge and memorization. Comparing performance to humans with access to Google is an attempt to do this, though the time constraints on answering the question may add some construct-irrelevant hardness relative to human-AI comparisons. To address these concerns, methods like factor analysis (Kim and Mueller, 1979) should be conducted to distinguish reasoning from memorization, for instance, and performance should be validated against other dedicated reasoning assessments.

Additionally, content and external validity must be established to confirm that the essential aspects of the construct are accurately measured and that findings generalize to unmeasured components. Moreover, criterion validity, when a construct-relevant criterion or established standard is available, can support construct validity since well-designed measurements should reliably predict external outcomes related to the same construct.

5.4 Consequential Validity

Consequential validity examines whether the real-world outcomes of decisions based on an assessment align with its intended purpose. In the case of GPQA, if the benchmark effectively measures scientific reasoning, AI models that perform well on it could support decision-making in scientific research or education. However, there is a risk of overgeneralization—high GPQA accuracy might lead to misinterpreting AI as possessing broad reasoning abilities when it may only excel at structured multiple-choice problems. In this case, there could be harmful consequences like replacing human workers with ill-suited technology.

For strong consequential validity, GPQA measurement must align with the reasoning skills they intend to measure, ensuring AI performance is interpreted within its actual capabilities. Clear performance guidelines should distinguish validated reasoning abilities from speculative claims, preventing misapplications of AI in scientific decision-making.

6 Who Should Care?

We highlight key active stakeholders in the current AI ecosystem; this list is not meant to be exhaustive or elevate the importance of these stakeholders above others.

- *Researchers.* Scientific claims about model capabilities must be grounded in evidence. Our framework helps researchers articulate what a given evaluation supports, distinguish between measurement and claim, and avoid overstating results, making their work more rigorous, interpretable, and useful to others.
- *Policy Makers.* As benchmarks increasingly guide regulatory decisions (e.g., under the EU AI Act), our framework provides a tool to assess whether the measurements being used actually support the claims that matter, like risk, safety, or societal impact. It helps ensure policy is evidence-aligned, not just benchmark-driven.
- *Corporations.* Product claims, deployment readiness, and resource allocation hinge on evaluation. Our framework helps internal teams diagnose whether current evaluations support the intended claims and identify what additional evidence is needed, reducing wasted effort and increasing confidence in deployment decisions.
- *Funders.* Grant agencies, philanthropic foundations, and investors must decide which projects merit scarce resources. By clarifying the link between what is measured and what is claimed, our framework lets funders vet proposals, set verifiable milestones, and track return on investment, ensuring their dollars advance research that delivers on its stated impact.
- *Civil Society.* As AI systems are deployed across sensitive domains, civil society organizations can use our framework to critically interrogate the basis of performance claims. It provides a structured way to ask: what exactly is being claimed, what is being measured, do they align, and how does misalignment affect their direct (personal use) and indirect (integration into critical infrastructure) day-to-day exposure?

Collective Accountability. Validity cannot be outsourced. It depends on alignment between those who develop measurements, perform evaluations, make claims, act on those claims, and are affected by decisions based on them. However, in practice, these roles are often distributed across stakeholders with different incentives and timelines; researchers, industry teams, regulators, and civil society actors rarely operate in lockstep.

Our framework provides shared structure and vocabulary to bridge this gap. It enables researchers to design evaluations that are not only technically sound but claim-aware.

Critically, we emphasize the need for an iterative feedback loop. Evaluations are not one-shot exercises; neither are the claims derived from them. As systems are deployed and additional evidence becomes available, whether through failure cases, distributional shifts, or evolving standards, this evidence must feed back into how we measure, interpret, and assess. A claim that may be valid under one context or at one point in time may become invalid as the world changes or as new use cases emerge.

This feedback loop must also operate across stakeholders. Claims made by one group (e.g., developers asserting safety) can prompt others (e.g., independent researchers or regulators) to test, contest, or refine those claims using new measurements and evaluation strategies. Rather than expecting perfect measurement upfront, the goal is a responsive evaluation ecosystem, one that updates as systems, stakes, and societal expectations evolve.

Working together does not require perfect consensus, but it does require transparency and shared responsibility. Validity is not a checklist; it is a process. By coordinating around that process, stakeholders can better ensure that claims about AI capabilities are not just persuasive but also defensible, grounded, and worthy of trust.

7 Decision Making with Validity in Mind

Decision-making that foregrounds validity is inherently context-dependent; no single, one-size-fits-all method will suffice. The same claim and evaluation can map to a 'go' decision for one context and a 'no-go' decision for another. Given a claim and evidence, we therefore recommend the following steps to help stakeholders systematically translate context-specific validity evidence into defensible decisions.

1. *Define Risk Tolerance.* Establish, up front, the maximum residual risk each stakeholder is willing to accept given the system's context and potential impact.

2. *Define the Scope of the Claim.* Populate our validity framework with markers that indicate where evidence is weak, moderate, or strong; flag unresolved risks; and log newly surfaced unknowns.

3. *Weight by Harm Potential.* Calibrate evidence bar to the gravity of potential harm: high-consequence claims (e.g., reasoning for medical triage) require stringent conditions and independent replication, whereas low-consequence claims (e.g., reasoning for puzzle-solving games) can proceed under proportionally lighter evidence and safeguards.

4. *Collective Review.* Convene developers, domain experts, risk owners, external auditors, and other stakeholders for a decision-focused session that examines the validity of the claim at hand.

5. *Document Decision Making and Share Publicly.* Record both the decision and its rationale: (i) 🆗 risk tolerance fully satisfied; (ii) ⚠️ specific gaps remain but are bounded by mitigation plans and a re-evaluation schedule; (iii) ❌ evidence is inadequate, triggering further data collection or analysis. Archive the reasoning and assign clear owners for any follow-up work.

8 Conclusion

Historically, AI evaluation has centered on benchmarks, emphasizing narrow technical progress while neglecting the validity of broader claims. This was acceptable when generative AI was primarily a research endeavor with limited impact. Today, as general-purpose AI systems are deployed more widely, traditional evaluation practices fail to capture real-world utility and risk enabling premature claims about readiness and robustness. To address this, we introduce a claim-centered framework grounded in five key forms of validity—content, external, criterion, construct, and consequential. We outline how these forms support four common mechanisms for linking evidence to claims, identify risks to each form of validity, and provide practical strategies for mitigation. Unlike prior work, we also offer guidance on reasoning under imperfect measurement conditions.

A central challenge in AI evaluation is the conceptual gap between measured performance and actual capability. Our framework systematically bridges this gap, ensuring that assessments are rigorous, context-sensitive, and scientifically grounded. By explicitly mapping the relationship between measurements, evaluations, and the claims they are used to justify, the framework helps prevent overgeneralization and supports more accurate interpretations of AI performance.

We argue that AI evaluation must be claim-aware, evidence-driven, and methodologically sound. Implicit in any evaluation is a network of relationships between constructs, measurements, and criteria—a nomological network. Yet, the lack of explicit articulation of these networks has hindered the alignment between what we measure and what we aim to understand. Without this clarity, evaluations risk misrepresenting model capabilities, leading to flawed conclusions and inappropriate applications.

By adopting a principled approach to validity, we move beyond surface-level benchmarks toward more transparent and reliable assessments. This shift is critical for developing and deploying AI systems that are trustworthy and aligned with real-world needs.

This work offers a theoretical foundation for validity-centered AI evaluation, setting the stage for empirical investigation. By clarifying how measurements, evaluations, and claims interact—and by emphasizing the role of nomological networks, we present a flexible framework applicable across AI domains and tasks. We highlight existing tools for probing different forms of validity and point to future directions focused on operationalizing this framework in high-stakes settings. *Building robust nomological networks that map AI constructs to measurable variables will be key to ensuring that evaluations yield not only performance metrics but also meaningful insights into real-world utility and risk.*

Acknowledgments

AA is funded by the NSF GRFP and the Knight-Hennessy Fellowship. SB is funded by the Stanford Graduate Fellowship (SGF). SK acknowledges support by NSF 2046795 and 2205329, the

MacArthur Foundation, and Schmidt Sciences. We thank Florian Dorner, Moritz Hardt, Tom Sühr, Sang Truong, and Serena Wang for detailed comments on early drafts.

References

Robert Adcock and David Collier. Measurement validity: A shared standard for qualitative and quantitative research. *Am. Polit. Sci. Rev.*, 95(3):529–546, September 2001.

Shivani Agarwal and Dan Roth. Learning a sparse representation for object detection. In *Computer Vision — ECCV 2002*, Lecture notes in computer science, pages 113–127. Springer Berlin Heidelberg, Berlin, Heidelberg, 2002.

Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C Lawrence Zitnick, Dhruv Batra, and Devi Parikh. VQA: Visual question answering. *arXiv preprint*, May 2015.

Ahmed Alaa, Thomas Hartvigsen, Niloufar Golchini, Shiladitya Dutta, Frances Dean, Inioluwa Deborah Raji, and Travis Zack. Medical large language model benchmarks should prioritize construct validity. *arXiv preprint arXiv:2503.10694*, 2025.

D I Andonov, B Ulm, M Graessner, A Podtschaske, M Blobner, B Jungwirth, and S M Kagerbauer. Impact of the covid-19 pandemic on the performance of machine learning algorithms for predicting perioperative mortality. *BMC Med. Inform. Decis. Mak.*, 23(1):67, April 2023.

Anthropic. Claude 3.5 sonnet. https://www.anthropic.com/news/claude-3-5-sonnet, 2024. [Accessed 05-04-2025].

Anthropic. Claude 3.7 sonnet and claude code. https://www.anthropic.com/news/claude-3-7-sonnet, 2025. Accessed: 2025-3-10.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint*, July 2019.

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, editors. *Standards for Educational and Psychological Testing*. American Educational Research Association, Washington, DC, 2014. ISBN 978-0-935302-35-6.

James T Austin and Peter Villanova. The criterion problem: 1917–1992. *J. Appl. Psychol.*, 77(6): 836–874, December 1992.

Xuechunzi Bai, Angelina Wang, Ilia Sucholutsky, and Thomas L Griffiths. Explicitly unbiased large language models still form biased associations. *Proceedings of the National Academy of Sciences*, 122(8):e2416228122, 2025.

Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and machine learning: Limitations and opportunities*. MIT press, 2023.

Samuel J Bell, Diane Bouchacourt, and Levent Sagun. Reassessing the validity of spurious correlations benchmarks. *arXiv preprint arXiv:2409.04188*, 2024.

Alfred Binet and Théodore Simon. *The Development of Intelligence in Children (The Binet-Simon Scale)*. Williams & Wilkins, Baltimore, MD, 1916. Translated by Elizabeth S. Kite. Originally published in 1905.

Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach. Stereotyping norwegian salmon: An inventory of pitfalls in fairness benchmark datasets. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1004–1015, Stroudsburg, PA, USA, 2021. Association for Computational Linguistics.

Avrim Blum and Moritz Hardt. The ladder: A reliable leaderboard for machine learning competitions. *arXiv preprint*, February 2015.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, et al. Findings of the 2017 conference on machine translation (wmt17). Association for Computational Linguistics, 2017.

Denny Borsboom, Gideon J Mellenbergh, and Jaap van Heerden. The concept of validity. *Psychol. Rev.*, 111(4):1061–1071, October 2004.

Nick Bostrom, Allan Dafoe, and Carrick Flynn. Public policy and superintelligent AI: A vector field approach. In *Ethics of Artificial Intelligence*, pages 293–326. Oxford University PressNew York, September 2020.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. A large annotated corpus for learning natural language inference. *arXiv preprint*, August 2015.

Hubert E Brogden and Erwin K Taylor. The theory and classification of criterion bias. *Educ. Psychol. Meas.*, 10(2):159–183, July 1950.

Urie Bronfenbrenner. Toward an experimental ecology of human development. *American psychologist*, 32(7):513, 1977.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Brian Buntz. Eureka 2.0: AI is beginning to ace grad-level science, but can you trust it? https://www.rdworldonline.com/eureka-2-0-ai-is-beginning-to-ace-grad-level-science-but-can-you-trust-it/, February 2025. Accessed: 2025-3-10.

D T Campbell and D W Fiske. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychol. Bull.*, 56(2):81–105, March 1959.

D T Campbell and J C Stanley. *Experimental and Quasi-Experimental Designs for Research*. Ravenio Books, 2015.

Gary L Canivez, Marley W Watkins, and Stefan C Dombrowski. Structural validity of the wechsler intelligence scale for Children-Fifth edition: Confirmatory factor analyses with the 16 primary and secondary subtests. *Psychol. Assess.*, 29(4):458–472, April 2017.

Gordon W Cheung, Helena D Cooper-Thomas, Rebecca S Lau, and Linda C Wang. Reporting reliability, convergent and discriminant validity with structural equation modeling: A review and best-practice recommendations. *Asia Pac. J. Manag.*, 41(2):745–783, June 2024.

Alexandra Chouldechova, Chad Atalla, Solon Barocas, A Feder Cooper, Emily Corvi, P Alex Dow, Jean Garcia-Gathright, Nicholas Pangakis, Stefanie Reed, Emily Sheng, Dan Vann, Matthew Vogel, Hannah Washington, and Hanna Wallach. A shared standard for valid measurement of generative AI systems' capabilities, risks, and impacts. *arXiv preprint*, December 2024.

Gregory J Cizek and Michael B Bunch. *Standard setting: A guide to establishing and evaluating performance standards on tests.* SAGE Publications Ltd, 2007.

Lee Anna Clark and David Watson. Constructing validity: Basic issues in objective scale development. *Psychol. Assess.*, 7(3):309–319, September 1995.

Alexis Conneau and Douwe Kiela. SentEval: An evaluation toolkit for universal sentence representations. *arXiv preprint*, March 2018.

Paper Copilot. NeurIPS 2024 statistics: Datasets & benchmarks track. https://papercopilot.com/statistics/neurips-statistics/neurips-2024-statistics-datasets-benchmarks-track/, April 2024. Accessed: 2025-3-12.

Amanda Coston, Anna Kawakami, Haiyi Zhu, Ken Holstein, and Hoda Heidari. A validity perspective on evaluating the justified use of data-driven decision-making algorithms. In *2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pages 690–704. IEEE, February 2023.

L J Cronbach and P E Meehl. Construct validity in psychological tests. *Psychol. Bull.*, 52(4): 281–302, July 1955.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. IEEE, June 2009.

Julian Stanley Donald T. Campbell. Experimental and Quasi-Experimental designs for research. *Cengage Learning*, 1963.

David Donoho. Data science at the singularity. *arXiv preprint*, October 2023.

Mojtaba Elhami Athar. The pitfalls of untested assumptions and unwarranted/oversimplistic interpretation of cultural phenomenon: a commentary on sajjadi et al. (2023). *Front. Psychol.*, 14: 1248246, September 2023.

Mark Everingham, Luc Van Gool, Christopher K I Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (VOC) challenge. *Int. J. Comput. Vis.*, 88(2):303–338, June 2010.

Li Fei-Fei, R Fergus, and P Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *2004 Conference on Computer Vision and Pattern Recognition Workshop*, pages 178–178. IEEE, 2005.

Aryo Pradipta Gema, Joshua Ong Jun Leang, Giwon Hong, Alessio Devoto, Alberto Carlo Maria Mancino, Rohit Saxena, Xuanli He, Yu Zhao, Xiaotang Du, Mohammad Reza Ghasemi Madani, Claire Barale, Robert McHardy, Joshua Harris, Jean Kaddour, Emile van Krieken, and Pasquale Minervini. Are we done with MMLU? *arXiv preprint*, June 2024.

Judy Wawira Gichoya, Imon Banerjee, Ananth Reddy Bhimireddy, John L Burns, Leo Anthony Celi, Li-Ching Chen, Ramon Correa, Natalie Dullerud, Marzyeh Ghassemi, Shih-Cheng Huang, et al. Ai recognition of patient race in medical imaging: a modelling study. *The Lancet Digital Health*, 4(6):e406–e414, 2022.

Elliot Glazer, Ege Erdil, Tamay Besiroglu, Diego Chicharro, Evan Chen, Alex Gunning, Caroline Falkman Olsson, Jean-Stanislas Denain, Anson Ho, Emily de Oliveira Santos, Olli Järviniemi, Matthew Barnett, Robert Sandler, Matej Vrzala, Jaime Sevilla, Qiuyu Ren, Elizabeth Pratt, Lionel Levine, Grant Barkley, Natalie Stewart, Bogdan Grechuk, Tetiana Grechuk, Shreepranav Varma Enugandla, and Mark Wildon. FrontierMath: A benchmark for evaluating advanced mathematical reasoning in AI. *arXiv preprint*, November 2024.

Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. *arXiv preprint*, July 2020.

Moritz Hardt. The emerging science of machine learning benchmarks, 2025.

Moritz Hardt and Benjamin Recht. Patterns, predictions, and actions: A story about machine learning. *arXiv preprint*, February 2021.

Amelia Hardy, Anka Reuel, Kiana Jafari Meimandi, Lisa Soder, Allie Griffith, Dylan M Asmar, Sanmi Koyejo, Michael S Bernstein, and Mykel J Kochenderfer. More than marketing? on the information value of AI benchmarks for practitioners. *arXiv preprint*, December 2024.

Donna Harman. Overview of the first trec conference. In *Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 36–47, 1993.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

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

Joseph Henrich, Steven J Heine, and Ara Norenzayan. The weirdest people in the world? *Behav. Brain Sci.*, 33(2-3):61–83; discussion 83–135, June 2010.

David J Hughes. *Psychometric validity: Establishing the accuracy and appropriateness of psychometric measures.* Wiley Online Library, 2018.

John Hunsley and Gregory J Meyer. The incremental validity of psychological testing and assessment: conceptual, methodological, and statistical issues. *Psychological assessment*, 15(4):446, 2003.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, Jayne Seekins, David A Mong, Safwan S Halabi, Jesse K Sandberg, Ricky Jones, David B Larson, Curtis P Langlotz, Bhavik N Patel, Matthew P Lungren, and Andrew Y Ng. CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. *arXiv*, January 2019.

Abigail Z Jacobs and Hanna Wallach. Measurement and fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, New York, NY, USA, March 2021. ACM.

Jennifer L Jennings and Jonathan Marc Bearak. "teaching to the test" in the NCLB era: How test predictability affects our understanding of student performance. *Educ. Res.*, 43(8):381–389, November 2014.

Carlos E Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. Swe-bench: Can language models resolve real-world github issues? *arXiv preprint arXiv:2310.06770*, 2023.

Kushal Kafle and Christopher Kanan. An analysis of visual question answering algorithms. *arXiv preprint*, March 2017.

Jae-On Kim and Charles W Mueller. *Factor analysis: Statistical methods and practical issues.* Quantitative Applications in the Social Sciences. SAGE Publications, Thousand Oaks, CA, February 1979.

Jennifer L Kobrin, Brian F Patterson, Emily J Shaw, Krista D Mattern, and Sandra M Barbuti. Validity of the SAT® for predicting First-Year college grade point average. research report no. 2008-5. *College Board*, 2008.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton A Earnshaw, Imran S Haque, Sara Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. WILDS: A benchmark of in-the-wild distribution shifts. *arXiv preprint*, December 2020.

Simon Kornblith, Jonathon Shlens, and Quoc V Le. Do better ImageNet models transfer better? *arXiv preprint*, May 2018.

Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images.(2009). *University of Toronto*, 18268744, September 2009.

T S Kuhn. *The structure of scientific revolutions*, volume 962. Chicago: University of Chicago press, 1997.

C H Lawshe. A QUANTITATIVE APPROACH TO CONTENT VALIDITY[1]. *Pers. Psychol.*, 28 (4):563–575, December 1975.

Y Lecun, L Bottou, Y Bengio, and P Haffner. Gradient-based learning applied to document recognition. *Proc. IEEE Inst. Electr. Electron. Eng.*, 86(11):2278–2324, 1998.

Q Vera Liao and Ziang Xiao. Rethinking model evaluation as narrowing the socio-technical gap. *arXiv preprint*, May 2023.

Weng Marc Lim. A typology of validity: content, face, convergent, discriminant, nomological and predictive validity. *Journal of Trade Science*, 12(3):155–179, September 2024.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C Lawrence Zitnick, and Piotr Dollár. Microsoft COCO: Common objects in context. *arXiv preprint*, May 2014.

Robert W Lissitz and Karen Samuelsen. A suggested change in terminology and emphasis regarding validity and education. *Educational researcher*, 36(8):437–448, 2007.

Yu Lu Liu, Su Lin Blodgett, Jackie Chi Kit Cheung, Q Vera Liao, Alexandra Olteanu, and Ziang Xiao. ECBD: Evidence-Centered benchmark design for NLP. *arXiv preprint*, June 2024.

David Lopez-Paz, Robert Nishihara, Soumith Chintala, Bernhard Schölkopf, and Léon Bottou. Discovering causal signals in images. *arXiv preprint*, May 2016.

William Meredith. Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, 58(4):525–543, 1993.

Samuel Messick. Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *Am. Psychol.*, 50(9):741–749, September 1995.

Samuel Messick. Test validity: A matter of consequence. *Soc. Indic. Res.*, 45(1/3):35–44, November 1998.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

Iman Mirzadeh, Keivan Alizadeh, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio, and Mehrdad Farajtabar. GSM-symbolic: Understanding the limitations of mathematical reasoning in large language models. *arXiv preprint*, October 2024.

C I Mosier. A critical examination of the concepts of face validity. *Educ. Psychol. Meas.*, 7(2): 191–205, July 1947.

Ulric Neisser. Cognition and reality. principles and implication of cognitive psychology. *San Francisko: WH Freeman and Company*, 1976.

Sharon L Nichols and David C Berliner. Collateral damage: How High-Stakes testing corrupts america's schools. In *Harvard Education Press*. Harvard Education Press. 8 Story Street First Floor, Cambridge, MA 02138. Tel: 888-437-1437; Tel: 617-495-3432; Fax: 978-348-1233; e-mail: hepg@harvard.edu; Web site: http://hepg.org/hep-home/home, March 2007.

Mohammad Sadegh Norouzzadeh, Anh Nguyen, Margaret Kosmala, Alexandra Swanson, Meredith S Palmer, Craig Packer, and Jeff Clune. Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning. *Proc. Natl. Acad. Sci. U. S. A.*, 115(25): E5716–E5725, June 2018.

Curtis G Northcutt, Anish Athalye, and Jonas Mueller. Pervasive label errors in test sets destabilize machine learning benchmarks. *arXiv preprint arXiv:2103.14749*, 2021.

Will Orr and Edward B Kang. AI as a sport: On the competitive epistemologies of benchmarking. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, New York, NY, USA, June 2024. ACM.

Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. Red teaming language models with language models. *arXiv preprint arXiv:2202.03286*, 2022.

Jean Ponce, Tamara L Berg, Mark Everingham, David A Forsyth, Martial Hebert, Svetlana Lazebnik, Marcin Marszalek, Cordelia Schmid, Bryan C Russell, Antonio Torralba, et al. Dataset issues in object recognition. *Toward category-level object recognition*, pages 29–48, 2006.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

Inioluwa Deborah Raji, Emily M Bender, Amandalynne Paullada, Emily Denton, and Alex Hanna. AI and the everything in the whole wide world benchmark. *arXiv preprint*, November 2021.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. *arXiv preprint*, June 2016.

Jennifer Randall. It ain't near 'bout fair: Re-envisioning the bias and sensitivity review process from a justice-oriented antiracist perspective. *Educ. Assess.*, 28(2):68–82, April 2023.

Benjamin Recht. The mechanics of frictionless reproducibility. *Harvard Data Science Review*, 6(1), 2024.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? *PMLR*, pages 5389–5400, May 2019.

Charles S Reichardt. Experimental and quasi-experimental designs for generalized causal inference, 2002.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. GPQA: A Graduate-Level Google-Proof Q&A benchmark. *arXiv preprint*, November 2023.

Anka Reuel, Amelia Hardy, Chandler Smith, Max Lamparth, Malcolm Hardy, and Mykel J Kochenderfer. BetterBench: Assessing AI benchmarks, uncovering issues, and establishing best practices. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, November 2024.

Elan Rosenfeld, Pradeep Ravikumar, and Andrej Risteski. The risks of invariant risk minimization. *Int Conf Learn Represent*, abs/2010.05761, October 2020.

Yangjun Ruan, Chris J Maddison, and Tatsunori Hashimoto. Observational scaling laws and the predictability of language model performance. *arXiv preprint*, May 2024.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C Berg, and Li Fei-Fei. ImageNet large scale visual recognition challenge. *arXiv preprint*, September 2014.

Olawale Salaudeen and Moritz Hardt. ImageNot: A contrast with ImageNet preserves model rankings. *arXiv [cs.LG]*, 2024.

Olawale Salaudeen and Sanmi Koyejo. Causally inspired regularization enables domain general representations. In *International Conference on Artificial Intelligence and Statistics*, pages 3124–3132. PMLR, April 2024.

Olawale Salaudeen, Nicole Chiou, Shiny Weng, and Sanmi Koyejo. Are domain generalization benchmarks with accuracy on the line misspecified? *arXiv preprint arXiv:2504.00186*, 2025.

Michael Saxon, Ari Holtzman, Peter West, William Yang Wang, and Naomi Saphra. Benchmarks as microscopes: A call for model metrology. *arXiv preprint*, July 2024.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

Jonathan A Shaffer, David DeGeest, and Andrew Li. Tackling the problem of construct proliferation: A guide to assessing the discriminant validity of conceptually related constructs. *Organ. Res. Methods*, 19(1):80–110, January 2016.

Lorrie A Shepard. Chapter 9: Evaluating test validity. *Rev. Res. Educ.*, 19(1):405–450, January 1993.

Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *J. Stat. Plan. Inference*, 90(2):227–244, October 2000.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. December 2013.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ambrose Slone, Ameet Rahane, Anantharaman S Iyer, Anders Andreassen, Andrea Madotto, Andrea Santilli, Andreas Stuhlmüller, Andrew Dai, Andrew La, Andrew Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakaş, B Ryan Roberts, Bao Sheng Loe, Barret Zoph, Bartłomiej Bojanowski, Batuhan Özyurt, Behnam Hedayatnia, Behnam Neyshabur, Benjamin Inden, Benno Stein, Berk Ekmekci, Bill Yuchen Lin, Blake Howald, Bryan Orinion, Cameron Diao, Cameron Dour, Catherine Stinson, Cedrick Argueta, César Ferri Ramírez, Chandan Singh, Charles Rathkopf, Chenlin Meng, Chitta Baral, Chiyu Wu, Chris Callison-Burch, Chris Waites, Christian Voigt, Christopher D Manning, Christopher Potts, Cindy Ramirez, Clara E Rivera, Clemencia Siro, Colin Raffel, Courtney Ashcraft, Cristina Garbacea, Damien Sileo, Dan Garrette, Dan Hendrycks, Dan Kilman, Dan Roth, Daniel Freeman, Daniel Khashabi, Daniel Levy, Daniel Moseguí González, Danielle Perszyk, Danny Hernandez, Danqi Chen, Daphne Ippolito, Dar Gilboa, David Dohan, David Drakard, David Jurgens, Debajyoti Datta, Deep Ganguli, Denis Emelin, Denis Kleyko, Deniz Yuret, Derek Chen, Derek Tam, Dieuwke Hupkes, Diganta

Misra, Dilyar Buzan, Dimitri Coelho Mollo, Diyi Yang, Dong-Ho Lee, Dylan Schrader, Ekaterina Shutova, Ekin Dogus Cubuk, Elad Segal, Eleanor Hagerman, Elizabeth Barnes, Elizabeth Donoway, Ellie Pavlick, Emanuele Rodola, Emma Lam, Eric Chu, Eric Tang, Erkut Erdem, Ernie Chang, Ethan A Chi, Ethan Dyer, Ethan Jerzak, Ethan Kim, Eunice Engefu Manyasi, Evgenii Zheltonozhskii, Fanyue Xia, Fatemeh Siar, Fernando Martínez-Plumed, Francesca Happé, Francois Chollet, Frieda Rong, Gaurav Mishra, Genta Indra Winata, Gerard de Melo, Germán Kruszewski, Giambattista Parascandolo, Giorgio Mariani, Gloria Wang, Gonzalo Jaimovitch-López, Gregor Betz, Guy Gur-Ari, Hana Galijasevic, Hannah Kim, Hannah Rashkin, Hannaneh Hajishirzi, Harsh Mehta, Hayden Bogar, Henry Shevlin, Hinrich Schütze, Hiromu Yakura, Hongming Zhang, Hugh Mee Wong, Ian Ng, Isaac Noble, Jaap Jumelet, Jack Geissinger, Jackson Kernion, Jacob Hilton, Jaehoon Lee, Jaime Fernández Fisac, James B Simon, James Koppel, James Zheng, James Zou, Jan Kocoń, Jana Thompson, Janelle Wingfield, Jared Kaplan, Jarema Radom, Jascha Sohl-Dickstein, Jason Phang, Jason Wei, Jason Yosinski, Jekaterina Novikova, Jelle Bosscher, Jennifer Marsh, Jeremy Kim, Jeroen Taal, Jesse Engel, Jesujoba Alabi, Jiacheng Xu, Jiaming Song, Jillian Tang, Joan Waweru, John Burden, John Miller, John U Balis, Jonathan Batchelder, Jonathan Berant, Jörg Frohberg, Jos Rozen, Jose Hernandez-Orallo, Joseph Boudeman, Joseph Guerr, Joseph Jones, Joshua B Tenenbaum, Joshua S Rule, Joyce Chua, Kamil Kanclerz, Karen Livescu, Karl Krauth, Karthik Gopalakrishnan, Katerina Ignatyeva, Katja Markert, Kaustubh D Dhole, Kevin Gimpel, Kevin Omondi, Kory Mathewson, Kristen Chiafullo, Ksenia Shkaruta, Kumar Shridhar, Kyle McDonell, Kyle Richardson, Laria Reynolds, Leo Gao, Li Zhang, Liam Dugan, Lianhui Qin, Lidia Contreras-Ochando, Louis-Philippe Morency, Luca Moschella, Lucas Lam, Lucy Noble, Ludwig Schmidt, Luheng He, Luis Oliveros Colón, Luke Metz, Lütfi Kerem Şenel, Maarten Bosma, Maarten Sap, Maartje ter Hoeve, Maheen Farooqi, Manaal Faruqui, Mantas Mazeika, Marco Baturan, Marco Marelli, Marco Maru, Maria Jose Ramírez Quintana, Marie Tolkiehn, Mario Giulianelli, Martha Lewis, Martin Potthast, Matthew L Leavitt, Matthias Hagen, Mátyás Schubert, Medina Orduna Baitemirova, Melody Arnaud, Melvin McElrath, Michael A Yee, Michael Cohen, Michael Gu, Michael Ivanitskiy, Michael Starritt, Michael Strube, Michal Swkedrowski, Michele Bevilacqua, Michihiro Yasunaga, Mihir Kale, Mike Cain, Mimee Xu, Mirac Suzgun, Mitch Walker, Mo Tiwari, Mohit Bansal, Moin Aminnaseri, Mor Geva, Mozhdeh Gheini, Mukund Varma T, Nanyun Peng, Nathan A Chi, Nayeon Lee, Neta Gur-Ari Krakover, Nicholas Cameron, Nicholas Roberts, Nick Doiron, Nicole Martinez, Nikita Nangia, Niklas Deckers, Niklas Muennighoff, Nitish Shirish Keskar, Niveditha S Iyer, Noah Constant, Noah Fiedel, Nuan Wen, Oliver Zhang, Omar Agha, Omar Elbaghdadi, Omer Levy, Owain Evans, Pablo Antonio Moreno Casares, Parth Doshi, Pascale Fung, Paul Pu Liang, Paul Vicol, Pegah Alipoormolabashi, Peiyuan Liao, Percy Liang, Peter Chang, Peter Eckersley, Phu Mon Htut, Pinyu Hwang, Piotr Miłkowski, Piyush Patil, Pouya Pezeshkpour, Priti Oli, Qiaozhu Mei, Qing Lyu, Qinlang Chen, Rabin Banjade, Rachel Etta Rudolph, Raefer Gabriel, Rahel Habacker, Ramon Risco, Raphaël Millière, Rhythm Garg, Richard Barnes, Rif A Saurous, Riku Arakawa, Robbe Raymaekers, Robert Frank, Rohan Sikand, Roman Novak, Roman Sitelew, Ronan LeBras, Rosanne Liu, Rowan Jacobs, Rui Zhang, Ruslan Salakhutdinov, Ryan Chi, Ryan Lee, Ryan Stovall, Ryan Teehan, Rylan Yang, Sahib Singh, Saif M Mohammad, Sajant Anand, Sam Dillavou, Sam Shleifer, Sam Wiseman, Samuel Gruetter, Samuel R Bowman, Samuel S Schoenholz, Sanghyun Han, Sanjeev Kwatra, Sarah A Rous, Sarik Ghazarian, Sayan Ghosh, Sean Casey, Sebastian Bischoff, Sebastian Gehrmann, Sebastian Schuster, Sepideh Sadeghi, Shadi Hamdan, Sharon Zhou, Shashank Srivastava, Sherry Shi, Shikhar Singh, Shima Asaadi, Shixiang Shane Gu, Shubh Pachchigar, Shubham Toshniwal, Shyam Upadhyay, Shyamolima, Debnath, Siamak Shakeri, Simon Thormeyer, Simone Melzi, Siva Reddy, Sneha Priscilla Makini, Soo-Hwan Lee, Spencer Torene, Sriharsha Hatwar, Stanislas Dehaene,

Stefan Divic, Stefano Ermon, Stella Biderman, Stephanie Lin, Stephen Prasad, Steven T Piantadosi, Stuart M Shieber, Summer Misherghi, Svetlana Kiritchenko, Swaroop Mishra, Tal Linzen, Tal Schuster, Tao Li, Tao Yu, Tariq Ali, Tatsu Hashimoto, Te-Lin Wu, Théo Desbordes, Theodore Rothschild, Thomas Phan, Tianle Wang, Tiberius Nkinyili, Timo Schick, Timofei Kornev, Titus Tunduny, Tobias Gerstenberg, Trenton Chang, Trishala Neeraj, Tushar Khot, Tyler Shultz, Uri Shaham, Vedant Misra, Vera Demberg, Victoria Nyamai, Vikas Raunak, Vinay Ramasesh, Vinay Uday Prabhu, Vishakh Padmakumar, Vivek Srikumar, William Fedus, William Saunders, William Zhang, Wout Vossen, Xiang Ren, Xiaoyu Tong, Xinran Zhao, Xinyi Wu, Xudong Shen, Yadollah Yaghoobzadeh, Yair Lakretz, Yangqiu Song, Yasaman Bahri, Yejin Choi, Yichi Yang, Yiding Hao, Yifu Chen, Yonatan Belinkov, Yu Hou, Yufang Hou, Yuntao Bai, Zachary Seid, Zhuoye Zhao, Zijian Wang, Zijie J Wang, Zirui Wang, and Ziyi Wu. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint*, June 2022.

Arjun Subramonian, Xingdi Yuan, Hal Daumé, III, and Su Lin Blodgett. It takes two to tango: Navigating conceptualizations of NLP tasks and measurements of performance. *arXiv preprint*, May 2023.

Robert L Thorndike. *Personnel selection; test and measurement techniques.* J. Wiley, New York, 1949.

Antonio Torralba and Alexei A Efros. Unbiased look at dataset bias. In *CVPR 2011*, pages 1521–1528. IEEE, 2011.

Steven Umbrello, Michael J Bernstein, Pieter E Vermaas, Anaïs Resseguier, Gustavo Gonzalez, Andrea Porcari, Alexei Grinbaum, and Laurynas Adomaitis. From speculation to reality: Enhancing anticipatory ethics for emerging technologies (ATE) in practice. *Technol. Soc.*, 74(102325): 102325, August 2023.

V N Vapnik and A Ya Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory Probab. Appl.*, 16(2):264–280, January 1971.

Hanna Wallach, Meera Desai, A Feder Cooper, Angelina Wang, Chad Atalla, Solon Barocas, Su Lin Blodgett, Alexandra Chouldechova, Emily Corvi, P Alex Dow, Jean Garcia-Gathright, Alexandra Olteanu, Nicholas Pangakis, Stefanie Reed, Emily Sheng, Dan Vann, Jennifer Wortman Vaughan, Matthew Vogel, Hannah Washington, and Abigail Z Jacobs. Position: Evaluating generative AI systems is a social science measurement challenge. *arXiv preprint*, February 2025.

Yuxuan Wan, Wenxuan Wang, Yiliu Yang, Youliang Yuan, Jen-Tse Huang, Pinjia He, Wenxiang Jiao, and Michael R Lyu. LogicAsker: Evaluating and improving the logical reasoning ability of large language models. *arXiv preprint*, January 2024.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint*, April 2018.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. *arXiv preprint*, May 2019.

Angelina Wang and Olga Russakovsky. Overwriting pretrained bias with finetuning data. *arXiv preprint*, March 2023.

Angelina Wang, Aaron Hertzmann, and Olga Russakovsky. Benchmark suites instead of leaderboards for evaluating AI fairness. *Patterns (N. Y.)*, 5(11):101080, November 2024.

Laura Weidinger, Inioluwa Deborah Raji, Hanna Wallach, Margaret Mitchell, Angelina Wang, Olawale Salaudeen, Rishi Bommasani, Deep Ganguli, Sanmi Koyejo, and William Isaac. Toward an evaluation science for generative ai systems. *arXiv preprint arXiv:2503.05336*, 2025.

Joseph Weizenbaum. Eliza—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1):36–45, 1966.

Terry Winograd. Understanding natural language. *Cognitive psychology*, 3(1):1–191, 1972.

Zhaofeng Wu, Linlu Qiu, Alexis Ross, Ekin Akyürek, Boyuan Chen, Bailin Wang, Najoung Kim, Jacob Andreas, and Yoon Kim. Reasoning or reciting? exploring the capabilities and limitations of language models through counterfactual tasks. *arXiv preprint*, July 2023.

Kai Xiao, Logan Engstrom, Andrew Ilyas, and Aleksander Madry. Noise or signal: The role of image backgrounds in object recognition. *arXiv preprint*, June 2020.

Ziang Xiao, Susu Zhang, Vivian Lai, and Q Vera Liao. Evaluating evaluation metrics: A framework for analyzing NLG evaluation metrics using measurement theory. *arXiv preprint*, May 2023.

Ziang Xiao, Wesley Hanwen Deng, Michelle S Lam, Motahhare Eslami, Juho Kim, Mina Lee, and Q Vera Liao. Human-centered evaluation and auditing of language models. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pages 1–6, New York, NY, USA, May 2024. ACM.

Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. τ-bench: A benchmark for tool-agent-user interaction in real-world domains. *arXiv preprint arXiv:2406.12045*, 2024.

Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12104–12113, 2022.

Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ADE20K dataset. *arXiv preprint*, August 2016.

Wangchunshu Zhou, Yan Zeng, Shizhe Diao, and Xinsong Zhang. VLUE: A multi-task benchmark for evaluating vision-language models. *arXiv preprint*, May 2022.

Appendix Table of Contents

A Evidence of Validity

Table 4: Common risks to validity, investigation tools, and evidence exemplars.

Validity	**Common risks**	**Investigation Tools**	**Evidence Exemplar**
Content Validity	☐ Coverage deficiency ☐ Construct irrelevance ☐ Imbalanced mixture of content	☐ Expert review ☐ Red-teaming / adversarially designed evaluations ☐ Synthetic data generation or edge cases	☐ Documentation of how test items comprehensively cover the construct ☐ Explicit mapping of test content to abstract frameworks or industry standards ☐ Coverage analysis
Criterion Validity	**Predictive and Concurrent Validity** ☐ Criterion contamination ☐ Criterion deficiency ☐ Restricted range ☐ Temporal/other shifts	☐ Real-world longitudinal studies ☐ Real-world behavioral testing ☐ Scaling-law predictive models ☐ Validated criterion studies ☐ Periodic post-deployment testing	☐ Correlation with an existing validated benchmark or gold standard ☐ Evidence that higher scores in evaluation metrics predict real-world utility

Table continues on the next page

Table 4 continued from previous page

Validity	Common risks	Investigation Tools	Evidence Exemplar
Construct Validity	**Structural:** ☐ Rank deficiency ☐ Poor factor structure ☐ Item interdependence ☐ Response format bias ☐ Complex measurement range	☐ Theory building and hypothesis testing ☐ Factor modeling ☐ Studies of process	☐ Observed changes in test performance under controlled conditions ☐ Item-test correlations ☐ Emergent substructures in model behavior
	Convergent: ☐ Irrelevant or weakly related evaluations ☐ High measurement error in scoring ☐ Restricted range (ceiling/floor effects) ☐ Confounding (e.g., memorization, format)	☐ Benchmark suites for a construct (e.g., reasoning) ☐ Representation probing (e.g., causal mediation analysis of embeddings)	☐ High correlation with other measures that assess the same construct ☐ Empirical clustering of model behaviors that align with constructs

Table continues on the next page

Table 4 continued from previous page

Validity	Common risks	Investigation Tools	Evidence Exemplar
	Discriminant: ☐ Construct overlap ☐ Format-induced correlations	☐ Orthogonal datasets ☐ Decomposable metrics	☐ Low or non-significant correlation with measures of distinct constructs ☐ Evidence that evaluation does not overlap with unrelated dimensions
External Validity	☐ Sample bias ☐ Unrealistic testing conditions ☐ Temporal variability ☐ Interaction effects ☐ Experimenter effects ☐ Task-specific bias	☐ Red-teaming ☐ Stress testing ☐ A/B testing ☐ Transfer testing ☐ Population-stratified evaluations	☐ Performance comparisons across different populations, environments, or settings ☐ Sensitivity analysis showing consistent performance under varying conditions ☐ Independent replication of results in different contexts or regions

Table continues on the next page

Table 4 continued from previous page

Validity	Common risks	Investigation Tools	Evidence Exemplar
Consequential Validity	☐ Bias / Fairness ☐ Adaptive overfitting ☐ Misuse of results ☐ Unintended incentives ☐ Policy and systematic consequences ☐ Temporal and other shift	☐ Stakeholder interviews and feedback loops ☐ Societal impact audits ☐ Ethical stress testing ☐ Stakeholder feedback	☐ Documented instances of evaluation-driven improvements in safety, reliability, and fairness ☐ Impact studies

B Validity

Validity refers to the extent to which a test accurately measures what it is intended to measure. Validity has a rich history, originally developed in the context of drawing valid conclusions from tests, much like how we now aim to draw valid conclusions from AI evaluations. One of the earliest forms of validity is face validity, which refers to the extent to which a test appears to measure what it claims to, based on intuitive judgment. For instance, one may ask if symbolic regression from BigBench (Srivastava et al., 2022) even appears to measure reasoning. However, relying on face validity alone can be misleading. As Charles Mosier (Mosier, 1947) famously observed:

"This form [face validity] is also gratifying to the ego of the unwary test constructor. It implies that his knowledge and skill in the area of test construction are so great that he can unerringly design a test with the desired degree of effectiveness in predicting job success or in evaluating defined personality characteristics, and that he can do this so accurately that any further empirical verification is unnecessary. So strong is this ego complex that if statistical verification is sought and found lacking, the data represent something to be explained away by appeal to sampling errors or other convenient rationalization, rather than by scientific evidence which must be admitted into full consideration."

A more structured form of validity emerged with content validity, which ensures that a test comprehensively covers all relevant aspects of the construct it aims to measure. For instance, one may ask if mathematical problem-solving benchmarks cover all relevant aspects of reasoning. Content validity is also typically assessed through expert judgment rather than statistical validation. Charles Lawshe (Lawshe, 1975) later formalized this concept with the Content Validity Ratio (CVR), a method for quantifying expert agreement on test content.

Moving toward empirical rigor, predictive validity assesses a test's ability to forecast an outcome of interest, typically a future outcome. This concept, introduced by Robert Thorndike in the mid-20th century during the rise of standardized testing, became central to fields like educational assessment, employment testing, and aptitude measurement (Thorndike, 1949). For example, the predictive validity of SAT scores for college GPA or cognitive ability tests for job performance has led to their widespread use for other outcomes (Kobrin et al., 2008). In the context of AI evaluation, one may ask "Does accuracy on IMO benchmarks predict accuracy in textbook linear algebra questions?" While predictive validity measures the correlation between a test and a future outcome, concurrent validity measures the correlation between a test and a validated standard applied at the same time under the same conditions. Predictive and concurrent validity make up criterion validity (Association et al., 2014).

While criterion validity is useful for assessing direct correlations between tests and desired criteria, its limitations became apparent when evaluating abstract constructs, like psychological traits, rather than simple outcome-based predictions. In their seminal work on construct validity, (Cronbach and Meehl, 1955) highlighted these limitations. For example, while SAT scores may predict GPA, they may not reliably measure intelligence, as GPA is influenced by grading biases and other factors. Recognizing the risks of relying solely on criterion-based validity, Cronbach and Meehl introduced construct validity, which assesses the extent to which a test truly captures the theoretical construct it purports to measure.

Two key sources of evidence necessary for construct validity introduced by Campbell and Fiske (1959) are (Campbell and Fiske, 1959):

- Convergent validity—the degree to which a test correlates with other measures of the same construct.
- Discriminant validity—the degree to which a test does not correlate with measures of unrelated constructs.

Implicitly, this framework also includes structural validity (Cronbach and Meehl, 1955; Messick, 1995), which examines whether a test's internal structure aligns with the theoretical construct it is designed to measure. This is often assessed using factor analysis or other dimensionality evaluations.

Cronbach and Meehl categorize validity into three primary forms:

1. *Content validity*—ensuring a test comprehensively represents the concept it aims to measure.
2. *Criterion validity*—evaluating how well a test correlates with external measures, which include predictive and concurrent validity. Concurrent validity refers to a test's agreement with a validated measure applied at the same time under the same conditions.
3. *Construct validity*—assessing the theoretical alignment between a test and its intended construct.

Beyond these core types, external validity refers to the extent to which a study's findings can be generalized beyond its specific conditions. External validity examines whether results hold across different populations, settings, and time periods. Campbell and Stanley (Campbell and Stanley, 2015) were among the first to systematically define external validity, identifying factors like selection bias and situational specificity as risks to generalizability.

In response to Cronbach and Meehl's framework, which emphasized the theoretical and statistical relationships between measures, (Messick, 1995, 1998) introduced consequential validity on the basis that validity is not just about measurement accuracy but also about the real-world impact of test interpretation and use. However, unlike (Messick, 1995), we do not unify all facets of validity under construct validity. We adopt the view of (Lissitz and Samuelsen, 2007) where the use of a measurement determines what is necessary to support validity. Importantly, this may not require construct validity.

Borsboom et al. (2004) offers a different view: validity is a property of the test itself, and a test is valid if and only if it measures the construct it purports to measure. In this view, questions of use or consequence are orthogonal to validity; what matters is whether the test causally reflects variation in the construct. This perspective draws a clear boundary between measurement and interpretation, placing the burden of validity squarely on the psychometric relationship between construct and test score. While theoretically clean, this stance omits considerations critical to our context, namely, how test outputs are used to make decisions. We, therefore, depart from Borsboom's definition, instead adopting a broader view in which validity also encompasses downstream consequences and use cases, particularly when evaluating AI systems deployed in high-stakes settings.

While these validity concepts were originally developed for psychological and educational testing, they provide a powerful lens for evaluating AI models. In the next section, we examine how these classical validity forms translate into the context of modern AI evaluation.

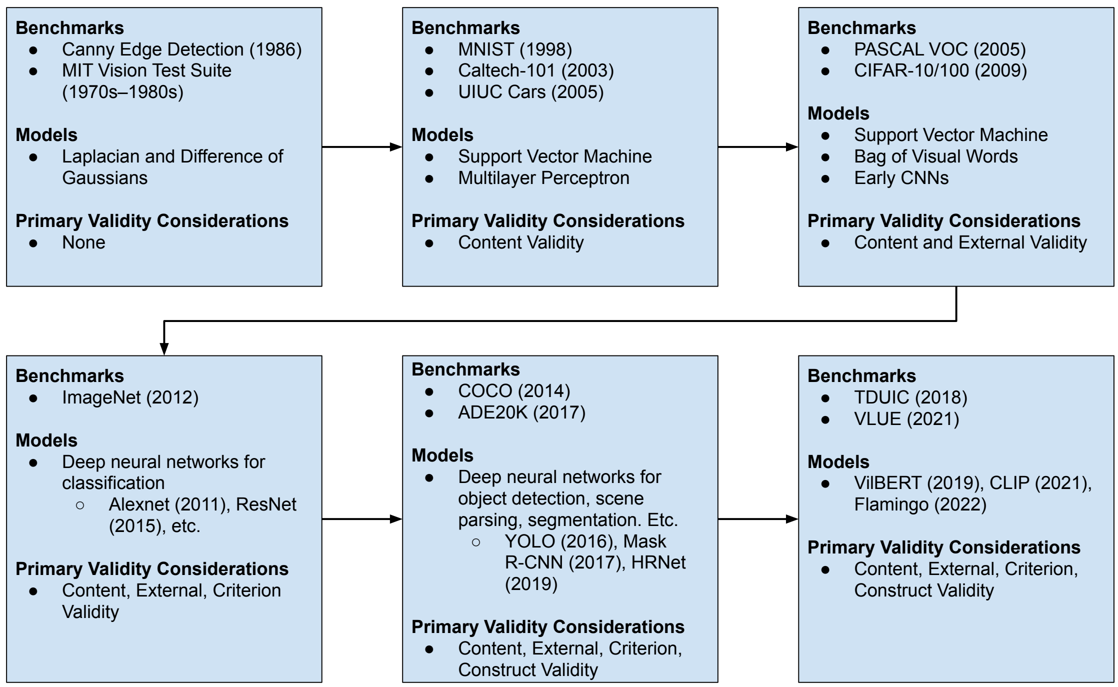

Figure 5: Coevolution of benchmarks, models, and the type of validity necessary for common conclusions for vision.

C The (Co)Evolution of evaluations and claims

C.1 Vision

The evolution of AI benchmarks has been closely tied to the kinds of conclusions researchers aimed to draw and the evidence available at the time—Figure 5. In the 1960s to 1980s, benchmarks were hyper-localized, focusing on narrowly defined technical tasks like edge detection and simple shape recognition. The goal was primarily technical exploration—improving algorithmic efficiency—so the scope of conclusions was very narrow and directly supported by the evaluations carried out.

In the 1990s, AI benchmarks became more structured and began incorporating more applied tasks. A notable example is MNIST (Lecun et al., 1998) for handwritten digit classification, which provided a standardized way to evaluate machine learning models. This trend continued into the early 2000s, with datasets such as UIUC Cars (Agarwal and Roth, 2002) for vehicle detection and Caltech-101 (2003) (Fei-Fei et al., 2005) for object recognition. While these benchmarks remained narrow in scope, they represented a step toward evaluating AI on more applied tasks, bridging the gap between theoretical research and practical applications. However, evaluations were still primarily designed for well-defined technical interests, with conclusions remaining local—focused on determining which

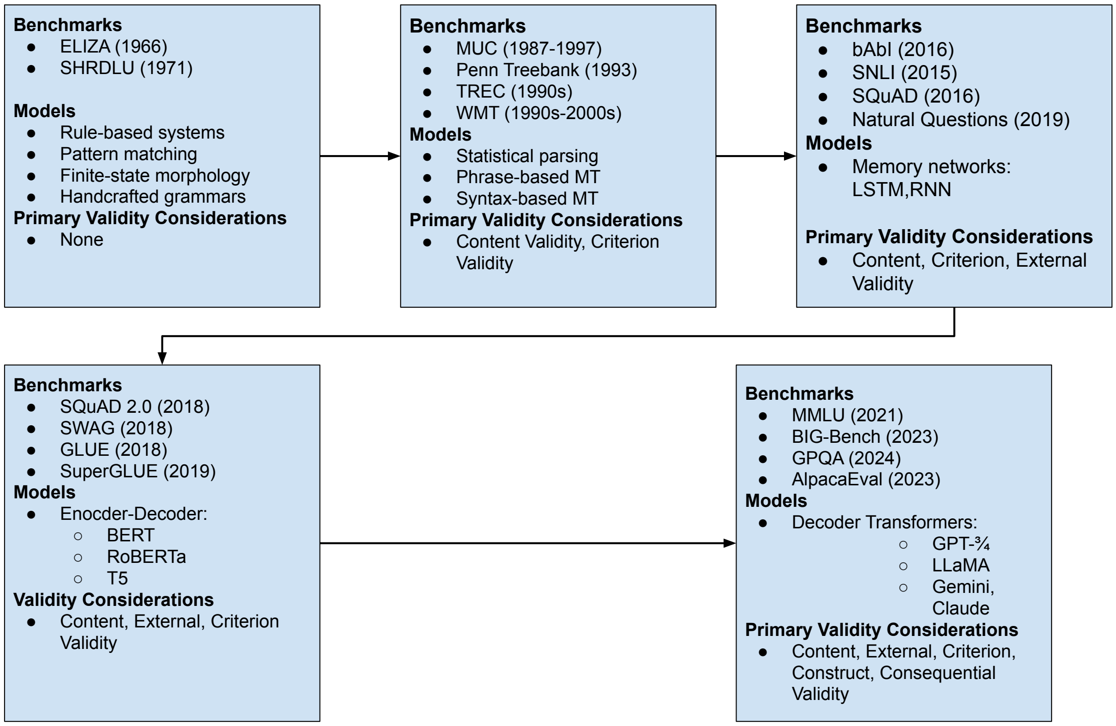

Figure 6: Coevolution of benchmarks, models, and the type of validity necessary for common conclusions for language.

techniques were most effective for the specific task being evaluated. During this period, researchers also became increasingly aware of content validity, recognizing that different datasets captured different aspects of classification tasks, which in turn influenced dataset design and evaluation methodologies (Ponce et al., 2006; Torralba and Efros, 2011).

By the mid-2000s, large-scale benchmarks such as PASCAL VOC (2007) (Everingham et al., 2010) introduced greater complexity, expanding evaluation beyond simple classification tasks. Later, in the late 2000s, CIFAR-10 and CIFAR-100 (Krizhevsky and Hinton, 2009) further pushed the field toward standardized comparisons in object recognition. During this period, criterion validity also gained prominence, as benchmark results were increasingly used to compare models in ways that suggested performance rankings carried external significance. However, construct validity remained largely unexplored—models were evaluated based on their outputs rather than on the reasoning processes behind their decisions. As a result, while evaluations became more sophisticated, they remained focused on performance metrics rather than deeper insights into model behavior. By this stage, the focus of AI evaluation began shifting from isolated dataset-specific improvements to broader claims about model robustness and transferability across different domains.

The 2010s marked a turning point with the ImageNet revolution. The introduction of ImageNet (Deng et al., 2009) and the ILSVRC (Russakovsky et al., 2014) competition (2010) provided large-scale, diverse, and complex benchmarks that dramatically reshaped AI research. During the early 2010s, the focus remained on improving accuracy in image classification and object detection. However, by the mid-2010s, AI evaluation expanded beyond leaderboards to real-world applications, particularly in medical imaging and autonomous driving. Researchers increasingly recognized the importance of content validity and external validity, leading to the widespread practice of testing models across multiple datasets to assess robustness.

As benchmark results gained influence, criterion validity became central—accuracy on ImageNet was frequently treated as a proxy for predicting downstream AI capabilities in vision. However, construct validity remained largely unaddressed in the early years. By the mid-2010s, early concerns emerged as researchers identified shortcut learning, adversarial vulnerabilities, and spurious correlations, leading to growing interest in understanding how models made decisions beyond raw accuracy. The rise of segmentation (COCO (Lin et al., 2014), ADE20K (Zhou et al., 2016)) and video analysis benchmarks (Kinetics, AVA) reflected an effort to capture more complex real-world tasks, but fundamental concerns about model robustness and bias persisted.

In the 2020s, the rise of multimodal and foundation models introduced even greater evaluation challenges. Benchmarks such as VQA (Agrawal et al., 2015), VLUE (Zhou et al., 2022), and TDIUC (Kafle and Kanan, 2017) attempted to assess multimodal reasoning, but defining what these benchmarks truly measured became increasingly difficult. Construct validity became a major concern as researchers debated whether these benchmarks genuinely assessed constructs like reasoning and understanding or merely exposed a model's ability to exploit statistical correlations in large datasets (Sec. 2). Unlike earlier benchmarks, which primarily focused on accuracy, modern benchmarks aim to evaluate the latent properties of AI systems Anthropic (2024). However, fundamental questions about the validity of these evaluations remain unresolved, particularly in assessing generalization, robustness, and true reasoning ability.

Across these decades, benchmarks evolved alongside the conclusions stakeholders sought to make. Early benchmarks required little discussion of validity because they were purely technical exercises. As AI models became more ambitious and claims about their capabilities expanded, benchmarks had to keep up—introducing concerns about content, external, and criterion validity. More recently, as AI systems move toward multimodal reasoning and foundation models, discussions of construct validity have become central. As models grow in complexity, the challenge is no longer just about designing better benchmarks—it's about defining what those benchmarks are actually supposed to measure in the first place.

C.2 Language

Language model benchmarks have seen an evolution from focusing on primarily basic questions of criterion validity against human performance to more nuanced considerations of other validity in more recent years—Figure 6. In the Blocks World Era (1960s-1980s), NLP evaluation was primarily qualitative and demonstration-based, lacking standardized metrics entirely. Systems like ELIZA (1966) (Weizenbaum, 1966) and SHRDLU (1971) (Winograd, 1972) were evaluated through anecdotal observations of how users interacted with them in highly constrained environments. ELIZA simulated a psychotherapist using simple pattern matching, while SHRDLU operated in a "blocks world" where users could issue commands to manipulate virtual objects. Validity considerations during this era were minimal and largely implicit. Content validity was severely limited by extremely narrow domains, criterion validity was nonexistent without standardized measurements, and construct validity wasn't addressed as researchers weren't attempting to measure specific capabilities like “reasoning” or “understanding.” External validity was particularly weak as systems couldn’t generalize beyond their constrained environments. Success was measured simply by the system’s ability to maintain seemingly intelligent conversations or follow instructions rather than through quantitative performance metrics or validity criteria. The North Star Era (1990s-2000s) marked a paradigm shift toward empirical evaluation with standardized benchmarks inspired by information retrieval traditions, where benchmarks with quantitative metrics and clearly defined train, validation ,and test split gave the field a proverbial “North Star” to aim towards. Initiatives like the Message Understanding Conferences (MUC) and the Penn Treebank established common datasets, clearly defined tasks, and metrics such as precision, recall, and F-score for comparing systems. This era introduced the first rigorous validity considerations, though still narrow in scope. Benchmarks like TREC (Harman, 1993) and WMT (Bojar et al., 2017) established improved criterion validity through standardized metrics that allowed consistent measurement across systems and time. Content validity improved but remained limited to specific linguistic tasks. Nascent construct validity concerns emerged as researchers began considering what abilities their tasks were actually measuring. However, external validity remained largely unaddressed as benchmarks weren’t designed to generalize beyond their specific contexts. Consequential validity still wasn’t a major consideration, as NLP applications weren’t yet widely deployed with significant societal impact.

In the early 2010s, many language benchmarks, such as SQuAD (Rajpurkar et al., 2016) and SNLI (Bowman et al., 2015), focused on individual tasks such as reading comprehension or natural language claims such as entailment or contradiction. The primary focus was on establishing baseline comparisons against human performance to create criterion validity for the benchmarks. However, such benchmarks had limitations to other aspect,s such as content validity due to limited focus on specific linguistic tasks and face validity due to narrow objectives and methods used to solve the task (both SQuAD and SNLI can be cast as relatively simple classification problems for which we can measure a gold standard of correctness). Other validity types were not heavily considered at this time.

In the mid to late 2010s, the field began to focus more on multi-task evaluation, which was represented by benchmarks such as GLUE (Wang et al., 2018) and SentEval (Conneau and Kiela, 2018). During this time, emerging validity concerns became prominent. More sophisticated human baselines were required to maintain criterion validity,and broader task coverage led to great content validity. However, concerns about the underlying mechanisms that could explain performance began to emerge, which reflects early concerns about construct validity.

In the late 2010s there were key changes in language model evaluation. Benchmarks like SuperGLUE (Wang et al., 2019) aimed to resolve validity concerns with rigorous multi-annotator baselines, broader task selection, more attention to the demographics of annotators, and the first considerations of social impact and gaming. However, the lack of structural validity evidence and external validation remained as challenges. There were also few analyses of convergent/discriminant validity in studies.

The 2020s marked a shift toward comprehensive knowledge evaluation with benchmarks like MMLU (Hendrycks et al., 2020), reflecting a growing recognition that language models were advancing beyond narrow linguistic tasks to broader knowledge and reasoning capabilities. MMLU introduced several innovations in validity considerations: it established expert-level performance as the criterion validity benchmark rather than average human performance, expanded content validity through coverage of 57 subjects across multiple domains, and highlighted crucial external validity concerns through studies showing sensitivity to answer ordering and other conditions that should not have an effect on the downstream performance for an "intelligent" agent (as measured with respect to an expert). The evolution of MMLU reflects broader trends in the field's approach to validity. Earlier benchmarks like SQuAD primarily focused on criterion validity through human performance comparisons, while MMLU attempted to address multiple validity types simultaneously. However, new challenges emerged: convergent validity became more complex as models showed inconsistent performance across related tasks (e.g., philosophy versus morality questions), and discriminant validity concerns arose around distinguishing between memorization and reasoning capabilities. This progression has led to the current state of language model evaluation, characterized by greater sophistication in validity considerations but also a clearer recognition of inherent limitations. Recent work has highlighted the need for better convergent validity across benchmarks and more robust methods for assessing reasoning abilities. The field has moved from treating benchmarks as simple performance metrics to viewing them as complex instruments requiring multiple types of validation evidence (Ruan et al., 2024).

D Case Studies

D.1 GPQA

Table 5: A Graduate-Level Google-Proof Question Answering Benchmark (GPQA) (Rein et al., 2023) Application. A subjective score for validity—the standard for "reasonable" is demonstrating that obvious risks to invalidity are addressed: 🆗: reasonable; ⚠️: proceed with caution; ❌: insufficient. Even for a score of "reasonable," there will be weaknesses in the evidence. The score is given because the strengths outweigh the weaknesses in terms of determining the validity of the claim from that evidence. This is never a binary classification nor complete, and should rather be a cyclic process—for instance, as our forms of what constitutes graduate-level chemistry may evolve over time and from school to school.

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.	🆗	🆗	🆗	🆗	⚠️
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.	⚠️	❌	❌	❌	⚠️

Description of dataset. The GPQA (Graduate-Level Google-Proof Question Answering) benchmark is a challenging dataset comprising 448 multiple-choice questions crafted by domain experts in biology, physics, and chemistry (Rein et al., 2023). These questions are designed to be exceptionally difficult, with experts holding or pursuing PhDs in the respective fields achieving an accuracy of 65% (74% when excluding clear mistakes identified retrospectively). Notably, highly skilled non-expert validators, even with unrestricted web access and spending over 30 minutes per question, attained only 34% accuracy, underscoring the "Google-proof" nature of the dataset. State-of-the-art AI systems also find this benchmark challenging; for instance, a GPT-4 based model achieved 39% accuracy. The GPQA dataset serves as a valuable resource for developing scalable oversight methods, aiming to enable human experts to effectively supervise and extract truthful information from AI systems that may surpass human capabilities.

Object of Claim: Multiple-choice questions in biology, physics, and chemistry accuracy.
Claim 1: AI models can accurately answer graduate-level specialized multiple-choice questions in biology, physics, and chemistry — criterion is accuracy on such questions.
Evidence: Accuracy on multiple-choice questions in biology, physics, and chemistry.

Validity of Claim from Evidence:

1. **Content Validity** 🆗

   - *Strength:* Expert-curated questions ensure high-quality, relevant content across key topics in biology, physics, and chemistry. The performance gap between experts and non-experts confirms the questions assess specialized knowledge.

- *Weakness:* The dataset's construction criteria may exclude some relevant questions, potentially leading to over- or underrepresentation of certain subfields.
- *Suggestions:* Conduct systematic content mapping across subfields to ensure balanced representation. Include expert diversity analysis to mitigate potential biases in question selection.

2. **Criterion Validity** 🆗
   - *Strength:* Human expert accuracy provides a meaningful external criterion, reinforcing concurrent validity.
   - *Weakness:* Criterion validity could be stronger with comparisons to other specialized science Q/A benchmarks. Predictive validity is untested—no evidence that GPQA accuracy predicts future performance on exams or coursework, for example.
   - *Suggestions:* Compare performance with established science Q&A benchmarks. Conduct longitudinal studies tracking how benchmark performance predicts success on real graduate exams.

3. **Construct Validity** 🆗
   - Since the claim is strictly about accuracy on a defined criterion, construct validity is not necessary to evaluate this specific claim.

4. **External Validity** 🆗
   - *Strength:* The test mirrors a real-world setting—human experts develop the questions, and the evaluation format aligns with academic multiple-choice assessments. GPQA includes diverse topics within its disciplines.
   - *Weakness:* Similar to the criterion validity gap, GPQA accuracy is not compared to other multiple-choice science tests, leaving external generalization unverified.
   - *Suggestions:* Validate against different question formats and compare performance across multiple science benchmarks.

5. **Consequential Validity** ⚠️
   - *Strength:* The AI-expert performance gap prevents premature claims of AI superiority, mitigating risks of overestimating AI scientific knowledge. However, models have quickly improved in this benchmark[5]. GPQA-trained models could support science education as study tools.
   - *Weakness:* If AI models reach high accuracy, stakeholders may overgeneralize their competence, assuming they have true expertise in physics, biology, and chemistry, despite lacking deeper scientific reasoning skills.
   - *Suggestions:* Develop clear guidance for stakeholders on interpreting results. Create documentation explicitly distinguishing multiple-choice performance from broader scientific expertise.

[5] https://www.youtube.com/watch?v=ZANbujPTvOY.

Object of Claim: Domain-specific scientific competency.
Claim 2: AI models can accurately answer graduate-level questions in specialized scientific domains—criterion is accuracy on such questions.
Evidence: Accuracy on [N] multiple-choice questions in biology, physics, and chemistry.

Validity of Claim from Evidence:

1. **Content Validity** ⚠️
   - *Strength:* Expert-curated, high-quality questions covering key topics in biology, physics, and chemistry. Non-expert performance gap supports specialization.
   - *Weakness:* Limited to three disciplines, excluding other specialized scientific domains (e.g., medicine, engineering). Only Q/A questions, excluding fill-in-the-blank or open-ended questions.
   - *Suggestions:* Expand questions to include other scientific subdomains. Conduct systematic content mapping across subfields to ensure balanced representation. Include expert diversity analysis to mitigate potential biases in question selection.

2. **Criterion Validity** ⚠️
   - *Strength:* Human expert accuracy serves as a strong external criterion (concurrent validity). AI-expert performance gap reinforces benchmark credibility.
   - *Weakness:* No predictive validity—GPQA accuracy is not tested against future performance on other specialized assessments.
   - *Suggestions:* Establish correlations with performance on real graduate program assessments. Develop predictive validity studies tracking model performance across time and domains.

3. **Construct Validity** ⚠️ *(importantly, this may be trivially satisfied if we have strong enough criterion validity.)*
   - *Strength:* Expert-curated questions in biology, physics, and chemistry are designed to capture fundamental aspects of specialized scientific knowledge. This suggests that the construct measured—domain-specific scientific competence—has meaningful representation, and high accuracy should correlate with understanding key scientific principles.
   - *Weakness:* GPQA's focus on biology, physics, and chemistry limits its ability to capture the overall construct of "specialized scientific knowledge," as other fields like medicine and engineering require different reasoning and knowledge structures. Moreover, the paper does not provide evidence linking GPQA performance to external measures of scientific competence (such as standardized test scores), leaving its alignment with related constructs unclear. Finally, the multiple-choice format may favor recognition or memorization over deeper analytical reasoning, potentially failing to capture key facets like synthesis and in-depth understanding.
   - *Suggestions:* To improve construct validity, expand GPQA to include additional domains (e.g., medicine, engineering) and correlate its scores with independent standardized assessments to establish convergent and discriminant validity. Additionally, incorporating alternative formats like open-ended questions and problem-solving tasks will better capture domain-specific scientific competence.

4. **External Validity** ⚠️

   - *Strength:* Real-world, expert-created multiple-choice questions ensure relevance. Coverage across multiple subfields increases generalization within biology, physics, and chemistry.
   - *Weakness:* No evidence of generalization to other science assessments (e.g., (non-)multiple choice PhD qualifying exams).
   - *Suggestions:* Test generalization to other assessment formats including written exams, oral defenses, and research proposal evaluations.

5. **Consequential Validity** ⚠️

   - *Strength:* AI-expert performance gap prevents overstating AI's scientific capabilities; models could support science education.
   - *Weakness:* Risk of overgeneralization—high scores may be misinterpreted as broad scientific expertise beyond tested domains.
   - *Suggestions:* Create clear limitations documentation highlighting specific domains where evidence supports or doesn't support performance claims.

Object of Claim: Reasoning.
Claim 3: AI models exhibit general reasoning abilities.
Evidence: Accuracy on [N] multiple-choice questions in biology, physics, and chemistry.

Validity of Claim from Evidence:

1. **Content Validity** ⚠️

   - *Strength:* Covers multiple scientific disciplines, requiring some level of reasoning beyond factual recall.
   - *Weakness:* Multiple-choice format limits assessment of forms of reasoning like logical deduction, or abstract problem-solving.
   - *Suggestions:* Develop specific reasoning-focused questions that isolate logical deduction from domain knowledge. Include diverse reasoning types (inductive, deductive, abductive).

2. **Criterion Validity** ❌

   - *Strength:* Human expert accuracy serves as a real-world external criterion, and the AI-expert performance gap indicates a meaningful benchmark for reasoning capabilities.
   - *Weakness:* GPQA tests factual and applied knowledge rather than abstract reasoning skills. No predictive validity—performance on GPQA is not tested against other established reasoning benchmarks (e.g., LSAT-style logical reasoning or problem-solving tests).
   - *Suggestions:* Compare performance against established reasoning benchmarks like LSAT, GRE analytical, and domain-independent logical reasoning tests.

3. **Construct Validity** ❌

- *Strength:* AI performance on GPQA correlates with success in structured question-answering tasks, suggesting some reasoning component. Additionally, the dataset can distinguish between human experts and non-experts.
- *Weakness:* Does not separate reasoning from memorization—AI models may exploit dataset patterns rather than apply logical deduction. While non-experts with access to Google perform worse than experts, non-experts are given a limited time per question, which may not sufficiently show that models have not been trained on such questions. No convergent validity—GPQA accuracy is not correlated with performance on explicit reasoning assessments. No discriminant validity—It is unclear whether GPQA measures reasoning ability or just domain-specific knowledge.
- *Suggestions:* Conduct factor analysis to distinguish reasoning from memorization. Demonstrate convergent validity with dedicated reasoning assessments and discriminant validity from pure knowledge recall.

4. **External Validity** ❌

   - *Strength:* GPQA questions require problem-solving across multiple disciplines, increasing the likelihood that some reasoning ability is being tested.
   - *Weakness:* Reasoning should generalize across domains, but GPQA only includes three scientific fields. No evidence that AI models with high GPQA accuracy perform well on general reasoning tasks outside science (e.g., logical puzzles, mathematical proofs, legal or philosophical reasoning).
   - *Suggestions:* Test performance on reasoning tasks across non-scientific domains including logic puzzles, mathematical proofs, and philosophical arguments.

5. **Consequential Validity** ⚠️

   - *Strength:* If GPQA successfully measures reasoning, AI models excelling on it could serve as decision-support tools in scientific research or education.
   - *Weakness:* Overgeneralization risk—high GPQA accuracy may lead to misinterpreting AI as possessing broad, human-like reasoning abilities when it may only excel at structured multiple-choice problems.
   - *Suggestions:* Develop clear performance interpretation guidelines specifying which reasoning capabilities are supported by evidence versus which remain speculative.

D.2 ImageNet

Table 6: An ImageNet (Deng et al., 2009; Russakovsky et al., 2014) Application. A subjective score for validity—the standard for "reasonable" is demonstrating that obvious risks to invalidity are addressed: 🆗: reasonable; ⚠️: proceed with caution; ❌: insufficient. Even for a score of "reasonable," there will be weaknesses in the evidence. The score is given because the strengths outweigh the weaknesses in determining the validity of the claim from that evidence. This evaluation is an iterative process, acknowledging that both the benchmark and its interpretations may evolve over time.

Claims from ImageNet Validity Assessment Report Card					
Claims	**Content**	**Criterion**	**Construct**	**External**	**Consequential**
1. ImageNet tests how well models learn complex associations between images and labels.	🆗	🆗	🆗	🆗	⚠️
2. ImageNet gauges the ability to learn semantically general visual features for object classification.	⚠️	🆗	⚠️	⚠️	⚠️
3. ImageNet measures overall visual understanding of a model.	❌	❌	❌	❌	❌

Description of dataset. ImageNet (Deng et al., 2009; Russakovsky et al., 2014) (specifically ILSVRC 2012) is a benchmark for predicting an image's label from a fixed set of 1000 diverse categories. The dataset—curated primarily from Flickr with human annotation—is evaluated using accuracy/error rate and precision/recall metrics.

Object of Claim: Predictive accuracy.
Claim 1: Model architectures can learn to accurately predict predefined image labels.
Evidence: Performance on accuracy/error rate and precision/recall metrics.

Validity of Claim 1 from Evidence:

1. **Content Validity** 🆗

   - *Strength:* The dataset covers 1000 diverse categories with extensive natural variability—including differences in poses, lighting, backgrounds, and fine-grained distinctions (e.g., different dog breeds)—making it well-suited to assess image–label associations.
   - *Weakness:* It is confined to static, natural RGB images and does not include other modalities (e.g., grayscale medical images or hyperspectral data) or dynamic contextual information (e.g., actions or inter-object relationships). Label noise may also affect accuracy metrics (Northcutt et al., 2021).
   - *Suggestions:* Clearly specify that ImageNet targets static natural images, and consider integrating supplementary datasets to represent additional image types or contextual settings.

2. **Criterion Validity** 🆗

   - *Strength:* There is robust evidence that performance on ImageNet is both predictive of downstream task success (models excelling on ImageNet often perform well on benchmarks such as CIFAR or Caltech, and in real-world applications like wildlife classification (Norouzzadeh et al., 2018)) and concurrent with human-annotated labels under similar conditions (He et al., 2020; Zhai et al., 2022; Kornblith et al., 2018).

3. **External Validity** 🆗

   - *Strength:* The dataset is representative of real-world natural images, and its utility has been demonstrated under varying conditions (differences in image quality, size, and even in applications to non-traditional domains such as medical imaging (Irvin et al., 2019) and adversarially constructed settings (Salaudeen and Hardt, 2024). Note, this is not about trained model performance (e.g., Recht et al. (2019)); it is about the external validity of model ability to learn and predict accurately, i.e., necessitates training and evaluating in a new setting rather than transporting trained models to a new setting.

4. **Construct Validity** 🆗

   - Since the claim is strictly about accuracy on a defined criterion, construct validity is not necessary to evaluate this specific claim.

5. **Consequential Validity** ⚠️

   - *Strength:* The clear quantification of labeling accuracy offers a concrete performance metric, facilitating transparent and reproducible comparisons.
   - *Weakness:* There is a risk that high ImageNet accuracy may be misinterpreted as reflecting comprehensive visual understanding, potentially leading to overconfident real-world deployments.
   - *Suggestions:* Advise stakeholders that ImageNet performance should be interpreted strictly as a measure of static image classification and that complementary evaluations are necessary to assess broader aspects of visual intelligence.

Object of Claim: Learning of semantically general visual features.
Claim 2: ImageNet evaluates the ability of models to learn transferable visual features that are useful for object classification.
Evidence: Performance gains in fine-tuning tasks when using models pretrained on ImageNet, compared to those trained from scratch.

Validity of Claim 2 from Evidence:

1. **Content Validity** ⚠️

   - *Strength:* The wide coverage of natural image phenomena—including fine-grained details and numerous object classes—supports the learning of varied and versatile visual features.
   - *Weakness:* It may not comprehensively represent features present in non-natural or synthetic environments, nor fully capture abstract contextual cues.
   - *Suggestions:* Consider integrating supplementary datasets that include synthetic, non-natural, or contextually complex images to achieve a more comprehensive assessment.

2. **Criterion Validity** 🆗

- *Strength:* Empirical studies (e.g., Kornblith et al. (2018)) show that ImageNet pretraining is strongly predictive of improved fine-tuning and transfer learning outcomes and that performance is concurrent with established classification tasks, addressing both the predictive and concurrent dimensions.
- *Weakness:* Although the predictive correlation is robust, direct and extensive concurrent comparisons with alternative feature assessment methods are less common.
- *Suggestions:* Enhance validation by conducting side-by-side evaluations comparing learned features across different pretraining methods and downstream tasks.

3. **Construct Validity**

- *Strength:* The improvement in fine-tuning performance suggests that the learned features are semantically rich and transferable. This provides evidence of structural validity (as features capture fundamental visual components), convergent validity (via correlation with downstream task performance), and discriminant validity (in differentiating meaningful features from noise).
- *Weakness:* It is challenging to definitively establish that these benefits are due to genuine generalization of visual features rather than overfitting to ImageNet-specific patterns, leaving the discriminant aspect less clear.
- *Suggestions:* Continually perform in-depth analyses—such as saliency mapping or kernel visualization—to further elucidate the nature of the learned features and clarify the extent of structural, convergent, and discriminant validity (Simonyan et al., 2013).

4. **External Validity**

- *Strength:* The benefits of ImageNet pretraining have been observed across multiple downstream benchmarks, suggesting that the learned features generalize beyond the confines of natural images (Kornblith et al., 2018; He et al., 2020).
- *Weakness:* The degree of generalizability across the span of domains (e.g., synthetic or non-natural images) remains to be fully validated.
- *Suggestions:* Broaden external validation by pretraining on a more diverse set of data and assessing performance on cross-domain tasks.

5. **Consequential Validity**

- *Strength:* The transformative impact of ImageNet pretraining in advancing computer vision is well-documented, highlighting its practical benefits.
- *Weakness:* An overreliance on fine-tuning improvements may obscure limitations in the intrinsic quality of the learned features, risking overgeneralization regarding model capability.
- *Suggestions:* Clearly communicate that fine-tuning gains indicate enhanced performance in specific settings rather than a comprehensive measure of visual feature quality; encourage complementary evaluations focused specifically on feature robustness.

Object of Claim: Visual understanding.
Claim 3: ImageNet provides an indication of a model's overall visual understanding beyond simple

label prediction or isolated feature representation.
Evidence: Performance on the standard classification task under controlled evaluation conditions, independent of training context.

Validity of Claim 3 from Evidence:

1. **Content Validity** ✗
    - *Strength:* The task of image classification is well-defined and widely used as a proxy for certain aspects of visual understanding.
    - *Weakness:* Relying solely on classification does not capture the full range of visual understanding, which includes spatial reasoning, object detection, contextual awareness, and causal interpretation. Understanding is multitask, including detection, segmentation, etc., which are not sufficiently investigated.
    - *Suggestions:* Complement the classification task with additional evaluations—such as object detection, visual question answering, or spatial reasoning challenges—to more fully capture the construct.
2. **Criterion Validity** ✗
    - *Strength:* Classification accuracy is a clear and quantifiable metric that enables direct comparison across models, addressing both predictive and concurrent aspects to some degree.
    - *Weakness:* There is limited evidence that high performance on this narrow task reliably predicts the broader and deeper aspects of overall visual understanding.
    - *Suggestions:* Compare ImageNet classification results with those from benchmarks explicitly designed to evaluate advanced visual reasoning and interpretative skills.
3. **Construct Validity** ✗
    - *Strength:* Operationalizing visual understanding as performance on image labeling provides a measurable framework that reflects a basic structural organization of visual recognition. However, it offers only limited convergent evidence with tasks requiring integrated reasoning and does not fully differentiate (discriminant validity) between mere pattern recognition and comprehensive understanding.
    - *Weakness:* This narrow operational approach may oversimplify the construct, favoring models that exploit dataset biases rather than achieving holistic visual comprehension.
    - *Suggestions:* Introduce complementary evaluation tasks (e.g., visual question answering or spatial reasoning challenges) to capture additional dimensions of visual understanding and enhance assessments of structural, convergent, and discriminant validity.
4. **External Validity** ✗
    - *Strength:* ImageNet's evaluation framework is reproducible, and similar performance trends have been observed across related image-based tasks.
    - *Weakness:* Its ability to generalize to tasks requiring integrated reasoning, spatial awareness, and contextual interpretation remains unconfirmed.

- *Suggestions:* Validate the broader aspects of visual understanding by employing a wider array of benchmarks that emphasize multidimensional reasoning and contextual evaluation.

5. **Consequential Validity** ✗

   - *Strength:* The benchmark has stimulated important discussions on the limitations of measuring visual intelligence solely via classification, underscoring the need for more comprehensive evaluation methods.
   - *Weakness:* High classification accuracy might be erroneously interpreted as evidence of complete visual understanding, potentially misleading real-world applications.
   - *Suggestions:* Provide clear guidelines on the interpretative scope of ImageNet results and promote complementary measures to capture the full spectrum of visual intelligence.