From Measurement to Meaning: A Validity-Centered Approach to AI Evaluation

Olawale Salaudeen[1*], Anka Reuel[2*], Ahmed Ahmed[2], Suhana Bedi[2], Zachary Robertson[2], Sudharsan Sundar[2], Angelina Wang[23†], Sanmi Koyejo[2†]

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

Corresponding emails: `olawale@mit.edu; sanmi@cs.stanford.edu.`

Abstract. While the capabilities and utility of AI systems have advanced, rigorous evaluation norms have lagged. Grand claims, such as models achieving general intelligence, are often evaluated with narrow benchmarks, like performance on graduate-level exam questions, which provide a limited and potentially misleading assessment. We provide a structured approach to reasoning about the types of evidence required for an evaluation to sufficiently support a claim. Our framework emphasizes a claim-first and a measurement and evaluation-first approach, where the latter aligns with the contemporary paradigm where various stakeholders provide measurements and evaluations, and users aim to validate claims and decisions by using this measurement to evaluate as a system. We illustrate this 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 *evaluate*[1] an AI system's[2] ability to solve the International Mathematical Olympiad (IMO) by *measuring*[3] its accuracy on such problems (2). Then, we want to validate two distinct *claims*[4] about the model's capabilities with this evaluation:

1. **Claim 1:** The system can solve university-level math problems accurately.
2. **Claim 2:** The system has reached human-level reasoning.

Clearly, asserting Claim 2 requires a much greater inferential leap from the observed evidence (IMO problem-solving accuracy) than Claim 1. If we claim that good performance on IMO problems demonstrates competence in university-level math, the justification is relatively strong: IMO problems often involve advanced undergraduate-level techniques, so proficiency in them provides reasonable evidence of university-level mathematical ability. 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—areas that IMO performance alone does not test. This difference

[1] We define evaluation to be the process used to assess a system.

[*] Equal contribution; [†]Joint senior authorship.

[2] Like Wallach et al. (1), to simplify exposition, we use the term "ML system" to refer to either 1) a single ML model or 2) one or more integrated software components, where at least one component is an ML model.

[3] We define a measurement as a quantitative or qualitative value that represents a specific attribute of a system, based on direct observation, computation, or empirical data.

[4] We define a claim to be an assertion of something about the system (e.g.,"this system can reason").

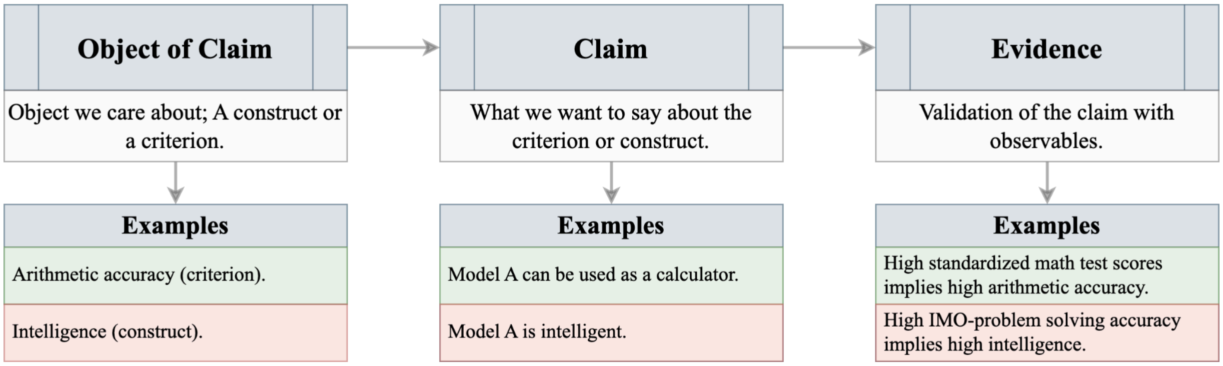

Object of Claim
Object we care about; A construct or a criterion.
Examples
Arithmetic accuracy (criterion).
Intelligence (construct).
Claim
What we want to say about the criterion or construct.
Examples
Model A can be used as a calculator.
Model A is intelligent.
Evidence
Validation of the claim with observables.
Examples
High standardized math test scores implies high arithmetic accuracy.
High IMO-problem solving accuracy implies high intelligence.

Figure 1. The three components are needed to begin the process of validation. First, we must decide what variable our claim is about; is it a criterion, or is it a construct? Then, we must explicitly state the claim. Finally, we must identify 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.

highlights that we must scrutinize an *evaluation* and *measurement* in the context of the claim we wish to support. We also consider validity in this context, defined as the degree to which an outcome, conclusion, or claim is justified by the evidence supporting it.

In measurement theory[5], "measurement validity" refers to the extent to which a test accurately measures what it is intended to measure. Yet, validity is not solely a property of the measurement itself—it also depends on the context of the evaluation it enables and the claims supported by that evaluation. Notably, this is counter to Boorsboom et. al.'s (3) view that that validity is solely a property of the test, whereas we argue that the meaning and strength of any measurement ultimately hinge on the context of its evaluation and the specific claims it is used to support, which is more in line with Messick's view on validity (4, 5).

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. **Claims**, judgments, assertions, decisions, about the system can then be supported by measurements and evaluations. **Measurement instruments** are tools to quantify, assess, or categorize an object and include benchmarks, user studies, and expert assessments (36).

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 problem) to gain some insight about the system's capabilities. To belabor 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.

[5] Measurement theory studies how abstract concepts are quantitatively assessed, ensuring that the chosen measures accurately capture the intended constructs while maintaining validity and reliability. We discuss validity and its treatment in other scientific disciplines in Appendix C.

Table 1. Table of terms.

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., 'university-level math accuracy').	Can be measured directly and often serves as a baseline or gold standard for evaluation.	"University-level exam accuracy" – the system's performance on real university math exams.
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 another example, we can measure the frequency of harmful outputs (e.g., misinformation or offensive responses) from a language model. However, 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 making judgments; we can measure for calibration. For instance, 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.

Despite the term "*measurement validity*," *validity is not merely a property of measurement.* We argue that it depends on the measurement, evaluation, and claim. A core limitation of current AI evaluation discourse is that validity—if considered—often focuses on the measurement–evaluation relationship, i.e., does the measurement sufficiently support the evaluated object (6)? However, this perspective is only complete under the assumption that the object being evaluated is the same as the object we want to make claims about. It also incorrectly suggests that if the measurement does not fully meet the needs of our evaluation, then no claims can be made.

In contrast, our framework allows claims and evaluations to differ and occur in any order. For instance, a policymaker might propose a claim that an AI system is sufficiently safe for deployment, which then prompts researchers to develop new evaluations and measurements to test safety in specific ways. Conversely, policymakers might start with existing evaluations and measurements from academia to inform safety considerations for deployment, effectively reversing the usual order of claim and evaluation. Additionally, even when measurements are limited, we can still derive claims—though such claims are often necessarily more narrow.

Crucially, evaluations and claims need not be provided by the same stakeholders for validity to hold. However, even a well-grounded evaluation by some stakeholders can result in overreaching claims by others if they infer properties beyond the scope of the measured domain.

To assess the validity of a claim derived from an evaluation, we must first carefully consider the *object of the claim* (Figure 1).

1. Is it a **construct**—an abstract variable that cannot be measured directly, like "mathematical reasoning"? Or is it a **criterion**—a directly measurable variable, such as "university-level math problem-solving accuracy"?
2. Furthermore, does the claim refer to the same property that was measured, or does it extend beyond the specific evaluation to infer something about a different property? For example, one might measure IMO problem-solving accuracy as part of an evaluation of mathematical problem-solving ability, then use this evaluation to support a claim about university-level math problem-solving ability, which is a different object.
3. Finally, is the claim supported directly by the measurement (e.g., IMO accuracy implies university-level math problem-solving capability), or does it rely on an intermediate construct (e.g., IMO accuracy implies mathematical reasoning, which in turn implies university-level math problem accuracy)?

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, and the overarching claim is central to establishing validity.

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.

Validity Type	Description	Example (IMO)
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 university-level math problems 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?

Ensuring validity requires five forms of validity, outlined in Table 2. Additionally, Appendix A Table 4 enumerates tools to investigate and establish each form. While other forms of validity exist[6], we identify this set as most relevant for current AI measurement validity gaps in Section 2.

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, these forms of validity are not fully independent and work together to demonstrate validity—we illustrate this in our case studies in Section 3 and Appendix B. Our proposed framework ties claims directly to the requisite standard of evidence provided

[6] Other forms of validity include, for example, face validity. Additional forms of validity forms are given in (7, 8).

by the evaluation. This alignment ensures the appropriate use and interpretation 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, 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 (9) 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" rather than leading to genuine improvements in the model.

Recognizing these challenges, we contribute a structured framework to assess the validity of AI assessments to ensure they are appropriately used and interpreted. Specifically, our **contributions** are:

1. We examine how the relevance of different forms of validity has co-evolved with the progress of AI and the corresponding norms and practices of evaluation.
2. We propose a practical and structured claim-aware framework for identifying the necessary evidence to establish the validity of claims based on AI evaluations.
3. We illustrate our framework through vision and language evaluation case studies, providing concrete, prescriptive examples of validating claims based on evaluations.

2. Validity Gaps in Current AI Evaluations and Related Work

AI evaluation has evolved alongside the complexity of AI tasks. Still, the gap between measurement and the object of claims about AI utility has widened (Appendix D), with deficiencies across all five forms of validity—content, criterion, construct, external, and consequential (Table 2). This widening gap results from changes in how AI systems are evaluated and the claims necessary to imply their utility. Early systems were tested on held-out samples from the same distribution (independent and identically distributed setting), ensuring content validity, i.e., relevant conditions were well-represented. Then, with pretraining and transfer learning, systems were first trained on large datasets like ImageNet (10, 11) and then finetuned for specific tasks, shifting evaluation toward predicting downstream performance, a form of criterion validity (12–14). Studies on spurious correlations (15–22), out-of-distribution generalizability, and biased representations emphasized external (12, 14) and consequential validity (23). However, evaluations remained benchmark-driven, primarily supporting claims of technical progress (24, 25)—this use is surprisingly robust (14, 26). Now, while these evaluation norms align researchers and industry, accelerating advances (27), they fail to predict real-world reliability (9).

The rise of foundation models, which can operate across diverse tasks without finetuning, further complicates this issue. Traditional evaluation methods increasingly fail to capture real-world AI behaviors that require investigating abstract capabilities like trustworthiness and reasoning (28–30).

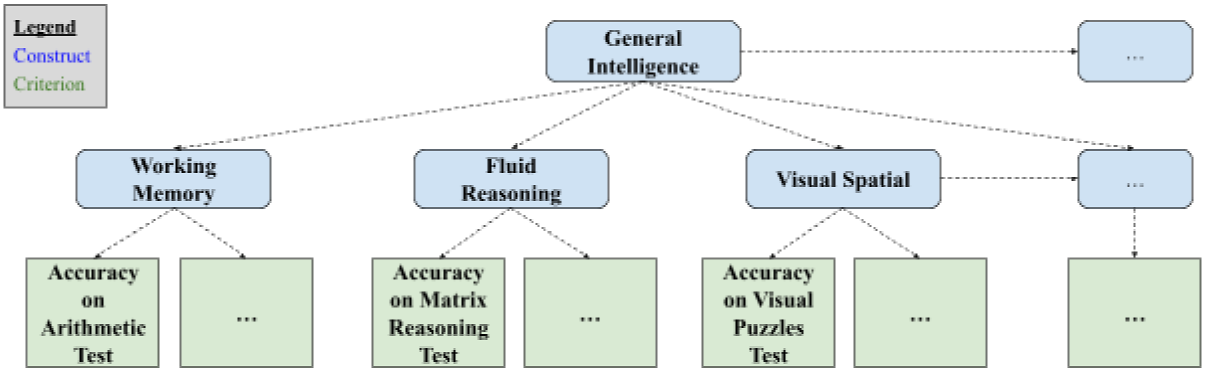


Figure 2. This figure illustrates a nomological network for *human general intelligence* according to the Wechsler Intelligence Scale, adapted from (31). *We note that this breakdown of general intelligence is for illustrative purposes only; it does not necessarily translate to artificial general intelligence.* 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.

Narrow datasets used for “general-purpose” evaluation raises content, construct, and external validity concerns, especially for complex tasks like reasoning (32). Furthermore, these evaluations, already saturating (33), lack criterion validity and fail to predict criteria of real-world applicability (9). At the same time, the socio-technical gap between evaluation results and real-world needs undermines consequential validity (34). Consequently, overgeneralized results erode evaluation credibility (35).

Prior work has demonstrated the need for validity frameworks (36–43), yet much of it has focused on measurement validity. METRICEVAL (39) raises validity concerns stemming from vague benchmark articulation and repurposed datasets in 16 natural language generation metrics. The important work of (1, 44) applied (45)’s measurement theory, critiquing ML evaluations for conflating systematizing a background concept with operationalizing a systemized concept.[7]

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 (36)—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 (46). These networks represent not only 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[8], 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 observable evidence. Consequently, we expand upon (36)’s view of systematization by arguing for the

[7] 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 (45).
[8] The Duhem-Quine thesis emphasizes that scientific claims are interconnected, meaning that rejecting or modifying one hypothesis affects others within the theoretical framework.

importance of broader nomological networks beyond just specifying which definition of a background concept will be used.

(47) 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 the process of designing new measurement instruments for evaluations while we emphasize the importance of the validity of claims from 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 (48).

Ultimately, our framework takes a practical approach, emphasizing that validity is not only a property of measurement and evaluation. As Cronbach and Meehl emphasize (46): "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.

3. A Framework for Claim-Centered Validity Assessment in AI Evaluation

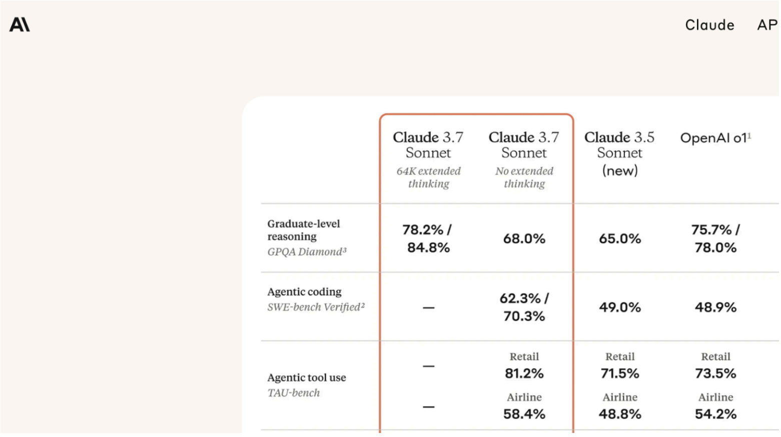
In this section, we categorize when and how different forms of validity are most critical for supporting a claim with measurements and evaluation. While we maintain that all forms of validity are always necessary, some may be trivially satisfied depending on the measurement, evaluation, and claim context. 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 (1, 45).

Establishing validity is an *iterative process* (46, 49), and recognizing the limitations of measurements and evaluations—rather than outright rejecting them when certain validity criteria fall short—requires nuance. This approach is essential for practical utility, enabling us to extract meaningful claims even from evaluations that do not rigorously satisfy all conditions of validity.

A claim can (and perhaps should (50)) be supported by many evaluations and measurements. However, for simplicity and without loss of generality, we focus on a single measurement, and we assume that the measurement is part of an evaluation.

Recall that the object of a claim can be a *criterion*—directly measurable—or a *construct*—abstract and not directly measurable. The primary considerations for investigating validity are determined by the following:

1. Is the object of the claim a criterion (e.g., math exam accuracy) or a construct (e.g., reasoning ability)?
2. Is the measurement the same as the object of the claim (e.g., evaluating IMO problem-solving accuracy when IMO problem-solving is also the object of the claim)?



	Claude 3.7 Sonnet *64K extended thinking*	Claude 3.7 Sonnet *No extended thinking*	Claude 3.5 Sonnet (new)	OpenAI o1[1]
Graduate-level reasoning *GPQA Diamond[3]*	78.2% / 84.8%	68.0%	65.0%	75.7% / 78.0%
Agentic coding *SWE-bench Verified[2]*	—	62.3% / 70.3%	49.0%	48.9%
Agentic tool use *TAU-bench*	— —	Retail 81.2% Airline 58.4%	Retail 71.5% Airline 48.8%	Retail 73.5% Airline 54.2%

Figure 4. 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. Additionally, SWE-bench is used as a proxy for agentic coding, and TAU-bench as a proxy for agentic tool use (51, 52).

3. Does the measurement directly imply the claim, or does it require a mediating construct (e.g., does IMO problem-solving imply university math exam problem-solving, OR does it imply mathematical reasoning, which implies math exam problem-solving)?

The five forms of validity we foreground are relevant in different ways. Generally, we want to validate a claim by directly measuring the object it is about. However, this may not be possible. When we perform a measurement but *the object of the claim is a different criterion* (e.g., evaluate IMO problem solving → make claims about university math problem solving), criterion validity is most important—ensuring 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 university math problem solving), requiring construct validity.

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 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 → make claims about mathematical reasoning).

Importantly, 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. Cronbach and Meehl's nomological network (46) provides a rigorous 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 (46). Figure 2 gives an example. 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 (46).

Next, we illustrate our framework for determining validity. We focus in the main text on GPQA multiple-choice question-answering accuracy as our measurement and evaluation. We then investigate claims of varying generality commonly made from this evaluation (51–53). Additional examples are in Appendix B. Since we directly measure GPQA question-answering accuracy to evaluate the same object, we focus on how the measurement supports the claim henceforth for convenience. 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 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 B.

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, we can start from a claim and use our framework to determine the necessary types of measurements and evaluations to support it, but this can also be done in reverse. (45)'s proposed framework for measurement validity starts with the construct (45), as does (1)'s framework for assessing generative AI. However, increasingly, measurements and evaluations are provided by some set of stakeholders, while other stakeholders are left to make sense of these relative to downstream uses of AI systems. When going through the process of validating claims from measurements and evaluations, the five forms of validity we foreground in Table 2 are not uniformly important or always necessary. In the following sections, we will describe situations where some types of validity are trivially satisfied.

3.1. Scenario 1: The object of the claim is a criterion

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." In this case, the measurement and the criterion object of the claim can be (i) identical, (ii) proxies of the same underlying construct, or (iii) proxies of two different but related underlying constructs. In each case, we must justify why the measurement supports the claim.

In scenario (i), 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 (1, 45).

Table 3. A Graduate-Level Google-Proof Q&A Benchmark (GPQA) (52) 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 Q&A (GPQA) Benchmark Accuracy Report Card					
Claims	Content	Criterion	Construct	External	Consequential
1. AI system can accurately answer ***graduate-level specialized multiple-choice questions*** in biology, physics, and chemistry.	🆗	🆗	🆗	🆗	⚠️
2. AI system can accurately answer ***graduate-level specialized questions*** in specialized scientific domains	⚠️	⚠️	⚠️	⚠️	⚠️
3. AI system exhibit ***general reasoning*** abilities.	⚠️	❌	❌	❌	⚠️

For (ii)-(iii), ideally, we additionally directly establish criterion validity, i.e., establish that the object that is measured is predictive of the desired criterion or an established standard. Then, the existence of these constructs may inform how we choose to establish criterion validity, but we do not need to reason about them directly to establish validity. 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., use 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 "reasoning ability" map to measurable criteria (e.g., "IMO accuracy"). In such cases, evaluations risk being narrow or misleading—a system might excel at algebra tasks yet lack geometric or logical skills, and evaluators could erroneously assume success in one facet implies overall "mathematical reasoning." 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.

Next, we enumerate considerations for validity depending on the measurement, evaluation, and claim.

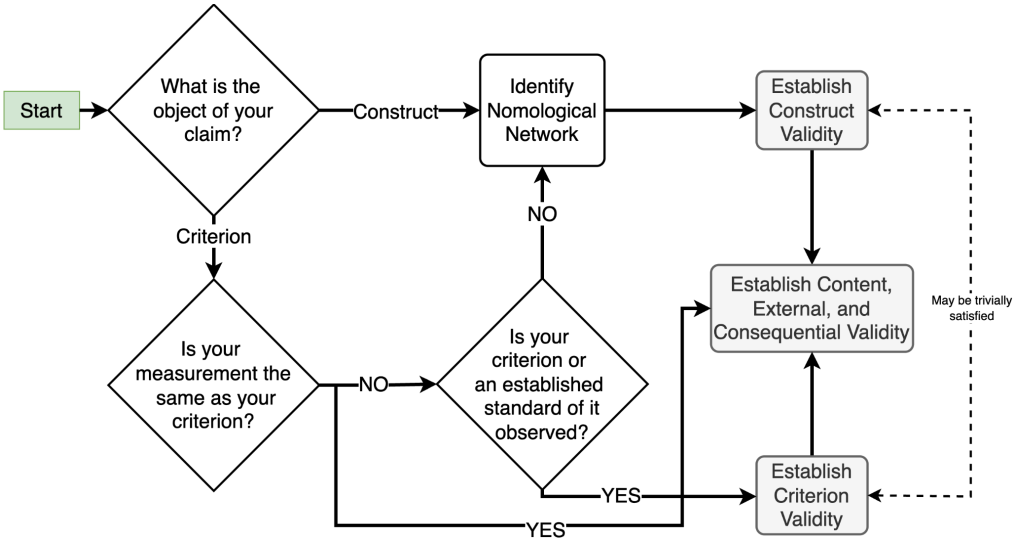

Start
What is the object of your claim?
Construct
Identify Nomological Network
Establish Construct Validity
Criterion
NO
Establish Content, External, and Consequential Validity
May be trivially satisfied
Is your measurement the same as your criterion?
NO
Is your criterion or an established standard of it observed?
YES
Establish Criterion Validity
YES

Figure 3. Decision tree 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.

Identical Measurement and Criterion

Example. Our *object of claim* is a criterion: multiple-choice questions accuracy in physics, chemistry, and biology[9]. We then aim to support *claims* about an AI system's can 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 construction 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.

If content validity holds and one does not expect that the context of measurement is identical to 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.

[9] 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.

To strengthen external validity, validation against diverse question formats and benchmarking across across assessment contexts is necessary.

This setting is commensurate with traditional AI benchmarking practices. Many AI benchmarks have focused on these forms of generalization, including classical generalization (54) and out-of-distribution generalization (55). By ensuring strong content and external validity, such benchmarks provide a solid foundation for validation claims for directly measurable criteria.

Different Measurement and Criterion

Here, the measurement and criterion can be different. Thus, when possible, one should establish criterion validity in addition to the content and external validity described in the example above. When this is not possible, and a nomological network is known, another mechanism to validate a claim is to identify structure in the mapping from constructs to observables that imply a relationship between the measurement and criterion. In this case, construct validity is necessary in addition to content and external.

Example. Our object of claim is scientific expertise, and we want to quantify levels of scientific expertise using 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 and the scientific expertise criterion, i.e.:

- Establish Criterion Validity. Human expert accuracy provides a strong external criterion, 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. To strengthen criterion validity, correlations should be established with real graduate program exams for concurrent validity, and predictive validity studies should track system performance across time and domains.

Alternatively, if we are unable to directly establish predictive validity, we can attempt to leverage the shared construct of scientific reasoning and establish construct validity to justify supporting such a claim with the given measurement if a nomological network is known. In this case, the **measurement and criterion share the same construct.** 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 3.2 on *Object of the claim is a construct,*

When the measurement and criterion are proxies of different constructs, we follow the same process but now we must validate relationships between constructs in addition to their relationships to observables. Doing this also requires knowledge of a nomological network.

Example. Suppose we want to claim that AI systems can reason about the outcomes of different surgical plans. 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. For example, in Figure 2, the accuracy on the arithmetic test and accuracy on the matrix reasoning test must go between working memory and and fluid reasoning.

Now, suppose we want to claim that strong GPQA performance translates into accurate surgical planning; this requires several inferential steps. GPQA (insufficiently) assesses scientific reasoning, while surgical planning 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 surgical decision-making. Without showing that GPQA performance reflects the same underlying capabilities as medical reasoning, claims about AI outperforming surgeons based on GPQA remain unverified.

3.2. Scenario 2. 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., 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, 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 separate reasoning from memorization—models may rely on dataset patterns rather than logical deduction. *Convergent validity* is unverified since GPQA accuracy has not been correlated with other explicit reasoning assessments. *Discriminant validity* is also uncertain, as it remains unclear whether GPQA measures genuine scientific reasoning or simply domain-specific knowledge. Comparing performance to humans with access to google is an attempt to do this. To address these concerns, methods like factor analysis (56) should be conducted to distinguish reasoning from memorization, and performance should be validated against other dedicated reasoning assessments while ensuring it diverges from pure knowledge recall.

Additionally, criterion 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 can support construct validity, as well-designed measurements should reliably predict external outcomes related to the same construct.

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.

4. Establishing Validity

Next, we examine common risks to validity claims 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.

Risks to content validity include coverage deficiency, where important aspects of the construct are missing, and construct irrelevance, where extraneous factors influence scores (5, 57, 58). 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 (59), and unrealistic testing conditions, which may not reflect real-world scenarios (60). Temporal variability and interaction effects can also distort results if performance shifts over time or due to specific environmental factors (61). 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 (62, 63). 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 (64, 65), and complex measurement range, where constructs are not well captured across different levels of ability (5). Convergent validity can suffer from high measurement error (66), which reduces reliability, while discriminant validity can be compromised by construct overlap, where different abilities are not clearly distinguished (67). 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 (5, 68). 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 (69). Policy consequences may emerge if flawed assessments influence high-stakes decisions. These risks can be assessed through anticipatory ethics methods (70), 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. Conclusions

Historically, AI evaluation has been benchmark-driven, focusing on narrow technical progress without critically assessing the validity of broader claims. This was fine in a regime where generative AI was a primarily research endeavor with less far-reaching consequences. However, as general-purpose generative AI systems continue to emerge, these traditional evaluation norms fail to predict real-world utility and risk irresponsible deployment and decision-making for AI systems. To address this, we enumerate five key forms of validity—content, external, criterion, construct, and consequential validity—each playing a critical role in determining whether a measurement and evaluation truly substantiates a given claim.

A fundamental challenge in AI evaluation is the conceptual gap between measured performance and real-world capability. Our claim-centered validity framework systematically bridges this gap, ensuring that AI assessments are rigorous, contextually appropriate, and scientifically robust. By explicitly mapping the relationship between measurements, evaluations, and the claims they are used to support, this framework prevents the overgeneralization of evaluation results and promotes a more accurate understanding of AI capabilities.

This work serves as a call for greater scientific rigor in AI evaluation, emphasizing that AI assessments must be claim-aware, evidence-driven, and methodologically sound. Whether formally articulated or not, the relationship between measurements, evaluations, and the claims they aim to validate implies an underlying nomological network—a structured web of relationships between constructs and criteria. However, the lack of explicit articulation of these networks has hindered progress in systematically aligning measurements and evaluations with AI capabilities. Without a deeper understanding of these relationships, evaluations risk misrepresenting AI capabilities, leading to flawed conclusions and misguided applications.

By adopting a principled approach to AI evaluation validity, we can move beyond surface-level benchmarking and towards a more scientifically grounded, transparent, and reliable assessment of AI systems. This shift is essential for ensuring that AI technologies are developed, evaluated, and deployed responsibly, ultimately fostering more trustworthy AI systems that align with real-world needs and expectations.

This work offers a theoretical foundation for validity-centered AI evaluation, setting the stage for more practical applications and empirical investigations. By clarifying how measurements, evaluations, and claims interact—and by emphasizing nomological networks—we provide a framework that can be adapted to diverse AI domains and tasks. Future research includes a focus on operationalizing this framework in high-stakes contexts and on building robust nomological networks that systematically map AI constructs to measurable variables. Such endeavors will help ensure that evaluations not only gauge narrow performance but also yield trustworthy insights into real-world utility and risk.

Bibliography

1. H. Wallach, *et al.*, Position: Evaluating generative AI systems is a social science measurement challenge. *arXiv [cs.CY]* (2025).
2. E. Glazer, *et al.*, FrontierMath: A benchmark for evaluating advanced mathematical reasoning in AI. *arXiv [cs.AI]* (2024).
3. D. Borsboom, G. J. Mellenbergh, J. van Heerden, The concept of validity. *Psychol. Rev.* **111**, 1061–1071 (2004).
4. S. Messick, Test Validity: A Matter of Consequence. *Soc. Indic. Res.* **45**, 35–44 (1998).
5. S. Messick, Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *Am. Psychol.* **50**, 741–749 (1995).
6. A. P. Gema, *et al.*, Are We Done with MMLU? *arXiv [cs.CL]* (2024).
7. W. M. Lim, A typology of validity: content, face, convergent, discriminant, nomological and predictive validity. *Journal of Trade Science* **12**, 155–179 (2024).
8. D. Hughes, *Psychometric Validity: Establishing the Accuracy and Appropriateness of psychometric measures* (2018).
9. A. Hardy, *et al.*, More than marketing? On the information value of AI benchmarks for practitioners. *arXiv [cs.AI]* (2024).
10. J. Deng, *et al.*, ImageNet: A large-scale hierarchical image database in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, (IEEE, 2009), pp. 248–255.
11. O. Russakovsky, *et al.*, ImageNet Large Scale Visual Recognition Challenge. *arXiv [cs.CV]* (2014).
12. S. Kornblith, J. Shlens, Q. V. Le, Do better ImageNet models transfer better? *arXiv [cs.CV]* (2018).
13. B. Recht, R. Roelofs, L. Schmidt, V. Shankar, Do imagenet classifiers generalize to imagenet? 5389–5400 (2019).
14. O. Salaudeen, M. Hardt, ImageNot: A contrast with ImageNet preserves model rankings. *arXiv [cs.LG]* (2024).

15. O. E. Salaudeen, O. O. Koyejo, Exploiting causal chains for domain generalization. *NeurIPS 2021 Workshop on Distribution Shifts: Connecting Methods and Applications* (2021).

16. O. Salaudeen, S. Koyejo, Causally Inspired Regularization Enables Domain General Representations in *International Conference on Artificial Intelligence and Statistics*, (PMLR, 2024), pp. 3124–3132.

17. D. Lopez-Paz, R. Nishihara, S. Chintala, B. Schölkopf, L. Bottou, Discovering Causal Signals in Images. *arXiv [stat.ML]* (2016).

18. K. Xiao, L. Engstrom, A. Ilyas, A. Madry, Noise or signal: The role of image backgrounds in object recognition. *arXiv [cs.CV]* (2020).

19. M. Arjovsky, L. Bottou, I. Gulrajani, D. Lopez-Paz, Invariant Risk Minimization. *arXiv [stat.ML]* (2019).

20. E. Rosenfeld, P. Ravikumar, A. Risteski, The risks of Invariant Risk Minimization. *Int Conf Learn Represent* **abs/2010.05761** (2020).

21. P. W. Koh, *et al.*, WILDS: A benchmark of in-the-wild distribution shifts. *arXiv [cs.LG]* (2020).

22. I. Gulrajani, D. Lopez-Paz, In search of lost domain generalization. *arXiv [cs.LG]* (2020).

23. A. Wang, O. Russakovsky, Overwriting pretrained bias with finetuning data. *arXiv [cs.CV]* (2023).

24. M. Hardt, B. Recht, Patterns, predictions, and actions: A story about machine learning. *arXiv [cs.LG]* (2021).

25. W. Orr, E. B. Kang, AI as a sport: On the competitive epistemologies of benchmarking in *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, (ACM, 2024).

26. A. Blum, M. Hardt, The Ladder: A reliable leaderboard for machine learning competitions. *arXiv [cs.LG]* (2015).

27. D. Donoho, Data science at the Singularity. *arXiv [stat.OT]* (2023).

28. Z. Wu, *et al.*, Reasoning or reciting? Exploring the capabilities and limitations of language models through counterfactual tasks. *arXiv [cs.CL]* (2023).

29. Y. Wan, *et al.*, LogicAsker: Evaluating and improving the logical reasoning ability of large language models. *arXiv [cs.SE]* (2024).

30. I. Mirzadeh, *et al.*, GSM-symbolic: Understanding the limitations of mathematical reasoning in Large Language Models. *arXiv [cs.LG]* (2024).

31. G. L. Canivez, M. W. Watkins, S. 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**, 458–472 (2017).

32. N. Bostrom, A. Dafoe, C. Flynn, "Public policy and superintelligent AI: A vector field approach" in *Ethics of Artificial Intelligence*, (Oxford University PressNew York, 2020), pp. 293–326.

33. S. Ott, A. Barbosa-Silva, K. Blagec, J. Brauner, M. Samwald, Mapping global dynamics of benchmark creation and saturation in artificial intelligence. *arXiv [cs.AI]* (2022).

34. Q. V. Liao, Z. Xiao, Rethinking model evaluation as narrowing the Socio-technical gap. *arXiv [cs.HC]* (2023).

35. I. D. Raji, E. M. Bender, A. Paullada, E. Denton, A. Hanna, AI and the everything in the whole wide world benchmark. *arXiv [cs.LG]* (2021).

36. A. Z. Jacobs, H. Wallach, Measurement and Fairness in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, (ACM, 2021).

37. M. Saxon, A. Holtzman, P. West, W. Y. Wang, N. Saphra, Benchmarks as microscopes: A call for model metrology. *arXiv [cs.SE]* (2024).

38. A. Subramonian, X. Yuan, H. Daumé III, S. L. Blodgett, It takes two to tango: Navigating conceptualizations of NLP tasks and measurements of performance. *arXiv [cs.CL]* (2023).

39. Z. Xiao, S. Zhang, V. Lai, Q. V. Liao, Evaluating evaluation metrics: A framework for analyzing NLG evaluation metrics using measurement theory. *arXiv [cs.CL]* (2023).

40. S. L. Blodgett, G. Lopez, A. Olteanu, R. Sim, H. 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)*, (Association for Computational Linguistics, 2021), pp. 1004–1015.

41. A. Coston, A. Kawakami, H. Zhu, K. Holstein, H. 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)*, (IEEE, 2023), pp. 690–704.

42. Z. Xiao, *et al.*, Human-centered evaluation and auditing of language models in *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, (ACM, 2024), pp. 1–6.

43. A. Reuel, *et al.*, BetterBench: Assessing AI benchmarks, uncovering issues, and establishing best practices in *The Thirty-Eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, (2024).

44. A. Chouldechova, *et al.*, A shared standard for valid measurement of generative AI systems' capabilities, risks, and impacts. *arXiv [cs.CY]* (2024).

45. R. Adcock, D. Collier, Measurement validity: A shared standard for qualitative and quantitative research. *Am. Polit. Sci. Rev.* **95**, 529–546 (2001).

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

47. Y. L. Liu, *et al.*, ECBD: Evidence-Centered Benchmark Design for NLP. *arXiv [cs.CL]* (2024).

48. NeurIPS 2024 Statistics: Datasets & Benchmarks Track. *Paper Copilot* (2024). Available at: https://papercopilot.com/statistics/neurips-statistics/neurips-2024-statistics-datasets-benchmarks-track/ [Accessed 12 March 2025].

49. T. S. Kuhn, The structure of scientific revolutions. **962** (1997).

50. A. Wang, A. Hertzmann, O. Russakovsky, Benchmark suites instead of leaderboards for evaluating

AI fairness. *Patterns (N. Y.)* **5**, 101080 (2024).

51. Anthropic, Claude 3.7 Sonnet and Claude Code. (2025). Available at: https://www.anthropic.com/news/claude-3-7-sonnet [Accessed 10 March 2025].

52. D. Rein, *et al.*, GPQA: A Graduate-Level Google-Proof Q&A Benchmark. *arXiv [cs.AI]* (2023).

53. B. Buntz, Eureka 2.0: AI is beginning to ace grad-level science, but can you trust it? *Research & Development World* (2025). Available at: https://www.rdworldonline.com/eureka-2-0-ai-is-beginning-to-ace-grad-level-science-but-can-you-trust-it/ [Accessed 10 March 2025].

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

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

56. J.-O. Kim, C. W. Mueller, *Factor analysis: Statistical methods and practical issues* (SAGE Publications, 1979).

57. Standards for Educational & Psychological Testing (2014 Edition). Available at: https://www.aera.net/publications/books/standards-for-educational-psychological-testing-2014-edition [Accessed 10 March 2025].

58. M. Furr, V. Bacharach, Psychometrics: An Introduction (2nd Ed.). (2013).

59. J. Henrich, S. J. Heine, A. Norenzayan, The weirdest people in the world? *Behav. Brain Sci.* **33**, 61–83; discussion 83–135 (2010).

60. J. S. Donald T. Campbell, Experimental and Quasi-Experimental Designs for Research. *Cengage Learning* (1963).

61. D. I. Andonov, *et al.*, Impact of the Covid-19 pandemic on the performance of machine learning algorithms for predicting perioperative mortality. *BMC Med. Inform. Decis. Mak.* **23**, 67 (2023).

62. H. E. Brogden, E. K. Taylor, The theory and classification of criterion bias. *Educ. Psychol. Meas.* **10**, 159–183 (1950).

63. J. T. Austin, P. Villanova, The criterion problem: 1917–1992. *J. Appl. Psychol.* **77**, 836–874 (1992).

64. L. A. Clark, D. Watson, Constructing validity: Basic issues in objective scale development. *Psychol. Assess.* **7**, 309–319 (1995).

65. M. 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 (2023).

66. G. W. Cheung, H. D. Cooper-Thomas, R. S. Lau, L. C. Wang, Reporting reliability, convergent and discriminant validity with structural equation modeling: A review and best-practice recommendations. *Asia Pac. J. Manag.* **41**, 745–783 (2024).

67. J. A. Shaffer, D. DeGeest, A. Li, Tackling the problem of construct proliferation: A guide to assessing the discriminant validity of conceptually related constructs. *Organ. Res. Methods* **19**,

80–110 (2016).

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

69. S. L. Nichols, D. 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, 2007).

70. S. Umbrello, *et al.*, From speculation to reality: Enhancing anticipatory ethics for emerging technologies (ATE) in practice. *Technol. Soc.* **74**, 102325 (2023).

71. D. Hendrycks, *et al.*, Measuring massive multitask language understanding. *arXiv [cs.CY]* (2020).

72. R. Ren, *et al.*, Safetywashing: Do AI safety benchmarks actually measure safety progress? *arXiv [cs.LG]* (2024).

73. Y. Ruan, C. J. Maddison, T. Hashimoto, Observational scaling laws and the predictability of language model performance. *arXiv [cs.LG]* (2024).

74. Y. Wang, *et al.*, MMLU-Pro: A more robust and challenging multi-task language understanding benchmark. *arXiv [cs.CL]* (2024).

75. V. Gupta, D. Pantoja, C. Ross, A. Williams, M. Ung, Changing answer order can decrease MMLU accuracy. *arXiv [cs.CL]* (2024).

76. H. Zhang, *et al.*, A careful examination of large language model performance on grade school arithmetic. *arXiv [cs.CL]* (2024).

77. A. Srivastava, *et al.*, Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models. *arXiv [cs.CL]* (2022).

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

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

80. R. L. Thorndike, *Personnel selection; test and measurement techniques* (J. Wiley, 1949).

81. J. L. Kobrin, B. F. Patterson, E. J. Shaw, K. D. Mattern, S. M. Barbuti, Validity of the SAT® for Predicting First-Year College Grade Point Average. Research Report No. 2008-5. *College Board* (2008).

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

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

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

85. S. Agarwal, D. Roth, “Learning a sparse representation for object detection” in *Computer Vision — ECCV 2002*, Lecture notes in computer science., (Springer Berlin Heidelberg, 2002), pp. 113–127.

86. L. Fei-Fei, R. Fergus, 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*, (IEEE, 2005), pp. 178–178.

87. M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, A. Zisserman, The pascal visual object classes (VOC) challenge. *Int. J. Comput. Vis.* **88**, 303–338 (2010).

88. A. Krizhevsky, G. Hinton, Learning multiple layers of features from tiny images.(2009). **18268744** (2009).

89. T.-Y. Lin, *et al.*, Microsoft COCO: Common objects in context. *arXiv [cs.CV]* (2014).

90. B. Zhou, *et al.*, Semantic understanding of scenes through the ADE20K dataset. *arXiv [cs.CV]* (2016).

91. A. Agrawal, *et al.*, VQA: Visual Question Answering. *arXiv [cs.CL]* (2015).

92. W. Zhou, Y. Zeng, S. Diao, X. Zhang, VLUE: A multi-task benchmark for evaluating vision-language models. *arXiv [cs.CV]* (2022).

93. K. Kafle, C. Kanan, An analysis of visual question answering algorithms. *arXiv [cs.CV]* (2017).

94. P. Rajpurkar, J. Zhang, K. Lopyrev, P. Liang, SQuAD: 100,000+ questions for machine comprehension of text. *arXiv [cs.CL]* (2016).

95. S. R. Bowman, G. Angeli, C. Potts, C. D. Manning, A large annotated corpus for learning natural language inference. *arXiv [cs.CL]* (2015).

96. A. Wang, *et al.*, GLUE: A multi-task benchmark and analysis platform for natural language understanding. *arXiv [cs.CL]* (2018).

97. A. Conneau, D. Kiela, SentEval: An evaluation toolkit for universal sentence representations. *arXiv [cs.CL]* (2018).

98. A. Wang, *et al.*, SuperGLUE: A stickier benchmark for general-purpose language understanding systems. *arXiv [cs.CL]* (2019).

Appendix

Appendix A: Establishing Validity

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

Validity	**Common risks**	**Investigation Tools**	**Evidence Exemplar**
Content Validity			
Content	● 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			
criterion	● 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
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 and floor effects) ● Confounding, e.g., memorization, format	□ Benchmark suites for a construct, e.g., reasoning □ Representation probing, e.g., causal mediation analysis of	□ High correlation with other measures that assess the same construct □ Empirical clustering of model behaviors that align with constructs

		embeddings	
Discriminant	● Construct overlap ● Format induced correlations	□ Orthogonal datasets □ Decomposable metrics	□ Low or non-significant correlation with measures of distinct constructs □ Empirical evidence that evaluation does not overlap with unrelated dimensions
External Validity			
External	● Sample bias ● Unrealistic testing conditions ● Temporal variability ● Interaction effects ● Experimenter effects ● Task-Specific bias	□ Red-Teaming □ Stress testing □ AB 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
Consequential Validity			
Consequential	● 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

Appendix B: Case Studies

GPQA

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 (52). 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.

Claims from Graduate-Level Google-Proof Q&A (GPQA) Benchmark Accuracy Report Card					
Claims	Content	Criterion	Construct	External	Consequential
AI systems can accurately answer ***graduate-level specialized multiple-choice questions*** in biology, physics, and chemistry.	🆗	🆗	🆗	🆗	⚠️
AI systems can accurately answer ***graduate-level specialized questions*** in specialized scientific domains	⚠️	⚠️	⚠️	⚠️	⚠️
AI systems can exhibit ***general reasoning*** abilities that can transfer beyond current human specialization.	⚠️	❌	❌	❌	⚠️

Object of Claim 1: 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 [N] 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 (https://www.youtube.com/watch?v=ZANbujPTvOY). 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.

Object of Claim 2: Q/A in biology, physics, and chemistry accuracy.
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.

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 disciplines beyond the current three to include medicine, engineering, and other scientific domains. Also include non Q/A examples. Conduct coverage analysis across the broader scientific landscape.
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 ⚠️ — as an alternative to 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:* The 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 deep analytical reasoning and synthesis skills.
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 3: Reasoning.
Claim 3. AI models exhibit general reasoning abilities.
Evidence. Accuracy on [N] multiple-choice questions in biology, physics, and chemistry.

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

MMLU

Description of dataset: MMLU (Massive Multitask Language Understanding) is a benchmark designed to test natural language understanding across 57 subjects spanning STEM, humanities, social sciences, and professional fields (71). It consists of multiple-choice questions (four options) drawn from standardized tests like the GRE and medical licensing exams, LSAT exams, and various exams oriented towards domain specific knowledge in the fields listed above.

Table 5. A subjective score for validity — the standard for sufficient is demonstrating 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.

Claims from Graduate-Level Google-Proof Q&A (GPQA) Benchmark Accuracy Report Card					
Claims	Content	Criterion	Construct	External	Consequential
1. Language models can demonstrate broad knowledge across diverse academic and professional subjects.	🆗	🆗	⚠️	⚠️	🆗
2. Language models can perform expert-level reasoning across specialized domains.	⚠️	⚠️	❌	❌	⚠️
3. MMLU performance predicts a model's general language understanding capabilities.	⚠️	⚠️	❌	❌	⚠️

Object of Claim 1: Broad knowledge across diverse subjects
Claim 1. "Language models can demonstrate broad knowledge across diverse academic and professional subjects."
Evidence. MMLU spans 57 subjects across STEM, humanities, social sciences, and professional fields, drawing from practice questions for standardized tests such as the Graduate Record Examination and the United States Medical Licensing Examination.
Validity of Claim from Evidence:

1. Content Validity 🆗
   - *Strength:* MMLU covers an extensive range of domains (57 subjects) spanning STEM, humanities, social sciences, and professional fields.
   - *Weakness:* The multiple-choice format with only four options limits the depth of understanding that can be assessed, and some subjects may have inadequate representation.
   - *Suggestions: Conduct detailed content mapping to ensure proportional representation across domains and expand beyond multiple-choice to include open-ended responses.*
2. Criterion Validity 🆗
   - *Strength:* MMLU has been shown to correlate with downstream performance on other capability oriented tasks, demonstrating

    predictive validity. Related work on benchmarking measured correlation of MMLU scores with the aggregate of scores on MMLU and other capability benchmarks, and found that MMLU to have a very high correlation only behind MedQA and Arc Challenge (72)
    - *Weakness:* There are inconsistencies in how well MMLU correlates with other measures of related capabilities (e.g., models performing well on philosophy but poorly on morality despite their relatedness).
    - *Suggestions: Conduct more systematic studies correlating MMLU performance with other established benchmarks of knowledge across domains.*

3. Construct Validity ⚠️
    - *Strength:* The benchmark draws from standardized tests designed to measure knowledge in respective fields.
    - *Weakness:* MMLU doesn't effectively distinguish between recall and reasoning lacking discriminant validity; high performance could indicate mere memorization from training data scraped from the internet rather than deep understanding.
    - *Suggestions:* Add questions that explicitly test reasoning or precision versus recall, and incorporate analysis of model explanations, not just final answers.
4. External Validity ⚠️
    - *Strength: Using questions from standardized tests provides some real-world grounding.*
    - *Weakness:* Significant issues undermine generalizability: labeling errors (57% of Virology questions contain errors), answer ordering effects, and the constrained multiple-choice format. This suggests an independent reproduction of MMLU might present different results
    - *Suggestions: Implement rigorous quality control (as in MMLU-Pro), test with varied answer orderings, and expand beyond multiple-choice formats.*
5. Consequential Validity ⚠️
    - *Strength:* MMLU has successfully become a standard benchmark driving industry progress in language model development.
    - *Weakness:* There is a risk of overoptimization as models are increasingly designed specifically to perform well on MMLU multiple choice, and might overfit to doing well on easily testable questions rather than broad subject knowledge (Goodhart's Law).
    - *Suggestions: Regularly update the benchmark with new questions and maintain clear documentation about what MMLU does and doesn't measure.*

Object of Claim 2: Expert-level reasoning
Claim 2. "Language models can perform expert-level reasoning across specialized domains."
Evidence. MMLU compares model performance against estimated expert-level accuracy (89.8%) and measures performance across specialized domains from

medicine to formal logic.

Validity of Claim from Evidence:

1. Content Validity ⚠️
   - *Strength:* MMLU includes questions from specialized professional domains that require some domain expertise.
   - *Weakness:* Multiple-choice questions as they are written within MMLU primarily tests factual knowledge rather than complex reasoning processes experts employ.
   - *Suggestions: Include multi-step reasoning problems and questions requiring application of principles to novel scenarios.*
2. Criterion Validity ⚠️
   - *Strength:* Performance is benchmarked against estimated expert-level accuracy (89.8%) so MMLU has a good claim to concurrent validity
   - *Weakness:* The benchmark cannot distinguish between memorized answers and expert reasoning. Error analysis shows 39% of incorrect answers on MMLU-Pro stem from reasoning errors despite correct knowledge, meaning the correlation with correct answers might be spurious.
   - *Suggestions: Incorporate expert validation of both answers and reasoning paths, perhaps through analysis of model explanations*
3. Construct Validity ❌
   - *Strength:* Some questions require application of domain knowledge rather than simple facts.
   - *Weakness:* The benchmark doesn't capture expert reasoning processes, only the final answers lacking structural validity
   - *Suggestions:* Develop metrics to evaluate reasoning quality, not just answer correctness, and include questions that cannot be solved through memorization alone. Elicit experts per domain for their reasoning process, as well as suggestions for relevant question formats and protocols.
4. External Validity ❌
   - *Strength:* Using standardized test questions provides some grounding in real assessment practices. As mentioned earlier, there is some evidence MMLU performance is correlated with performance on other capability benchmarks
   - *Weakness:* Multiple-choice tests do not capture the open-ended, iterative nature of expert reasoning in real-world contexts. Changing answer ordering can also affect scores which an expert should be invariant to.
   - *Suggestions: Develop supplementary benchmarks with more authentic professional tasks and varied formats. Perhaps where the model provides reasoning chains and is evaluated with a reward model calibrated to expert preference*
5. Consequential Validity ⚠️
   - *Strength: The benchmark has helped identify strengths and weaknesses in model capabilities across different domains*
   - *Weakness: High MMLU scores might create an illusion that models can*

replace domain expert judgement, leading to inappropriate applications
    - *Suggestions: Provide clear guidance on the limitations of what MMLU scores indicate about true expert-level reasoning*

Object of Claim 3: Predictive power for general capabilities
Claim 3. "MMLU performance predicts a model's general language understanding capabilities."
Evidence. MMLU has been highly correlated with downstream quality and capability, as noted by industry teams building large language models and supported by research on observational scaling laws.
Validity of Claim from Evidence:

1. Content Validity ⚠️
    - *Strength:* MMLU covers a wide range of domains, providing breadth in assessment that is a non-trivial subset of understanding of “general” topics, if such topics are the enumeration of all academic topics.
    - *Weakness: It doesn't cover all aspects of language understanding, particularly creative, open-ended, or interactive capabilities. It also doesn’t cover areas of knowledge that aren’t readily measured in academic settings*
    - *Suggestions: Supplement with other benchmarks measuring different facets of language understanding and areas that don’t easily map to academic fields of study such as humor*
2. Criterion Validity ⚠️
    - *Strength:* Research on observational scaling laws notes that when running a PCA on evaluation performance of prominent benchmarks against downstream performance, variation in MMLU explains a large fraction of variation (73). As mentioned earlier in claim 1 and claim 2, research shows MMLU scores correlate well with performance on other tasks, supporting its use as a general predictor (72). Combined with the earlier observation that performance is benchmarked against estimated expert-level accuracy (89.8%), this gives MMLU a good claim to concurrent validity.
    - *Weakness:* Correlation patterns are inconsistent across different types of tasks and domains (74)
    - *Suggestions: Develop a more nuanced framework showing which aspects of MMLU best predict which types of downstream capabilities or rely on the observational scaling laws framework*
3. Construct Validity ❌
    - *Strength:* The benchmark captures some aspects of knowledge acquisition and application.
    - *Weakness:* "Natural language understanding" as a construct encompasses much more than multiple-choice question answering, including discourse comprehension, pragmatics, and nuanced interpretation none of which are covered here.
    - *Suggestions:* Clarify the specific sub-constructs of language understanding that MMLU actually measures.

4. External Validity ❌
    - *Strength:* The breadth of subjects provides some basis for generalization, assuming we are focused on breadth and a more shallow definition of generality rather than depth.
    - *Weakness:* MMLU's format and limitations (answer ordering effects, label errors) raise questions about how well scores generalize to real-world language understanding tasks (74, 75).
    - *Suggestions:* Test whether MMLU-high-performing models also excel at real-world tasks requiring language understanding in non-test environments
5. Consequential Validity ⚠️
    - *Strength:* MMLU has influenced productive research directions in language model development, such as BigBench, GPQA, GAIA and other benchmarks that test language models on a broad set of tasks
    - *Weakness:* Over-reliance on MMLU as a general capability metric could lead to narrowly optimized models for the benchmark rather than genuinely more capable ones. This can lead to overstating progress and capabilities of the latest models and systems, i.e. models such as Phi-1 and Mistral which overfits to GSM8k and saw large drops in performance when tested on a new private split (76).
    - *Suggestions:* Develop complementary metrics that capture aspects of language understanding not measured by MMLU, and emphasize a balanced assessment approach.

Appendix C: Validity

Validity refers to the degree to which evidence and theory support the interpretations, conclusions, and decisions drawn from data or measurement. 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 (77) even appears to measure reasoning. However, relying on face validity alone can be misleading. As Charles Mosier (78) 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 (79) 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 (80). 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 (81). In the context of AI evaluation, one may ask "does accuracy on IMO benchmarks predict accuracy in grading university math exams?"

While predictive validity is useful for assessing direct correlations between tests and outcomes, its limitations became apparent when evaluating theoretical abstract constructs rather than simple outcome-based predictions. In their seminal on construct validity, (46) 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 (82):

- 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 (5, 46), 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.

Building on these foundational ideas, Cronbach and Meehl categorized 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 (83) 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, (4, 5) 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.

We primarily follow Cronbach and Meehl's classification of validity types. However, we adopt Messick's view that validity extends beyond measurement properties. Thus, we incorporate consequential validity as an essential consideration in our framework. We, however, diverge from (3)'s view that validity is only a property of the test.

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.

Appendix D: The (Co)Evolution of evaluations and conclusions

Vision

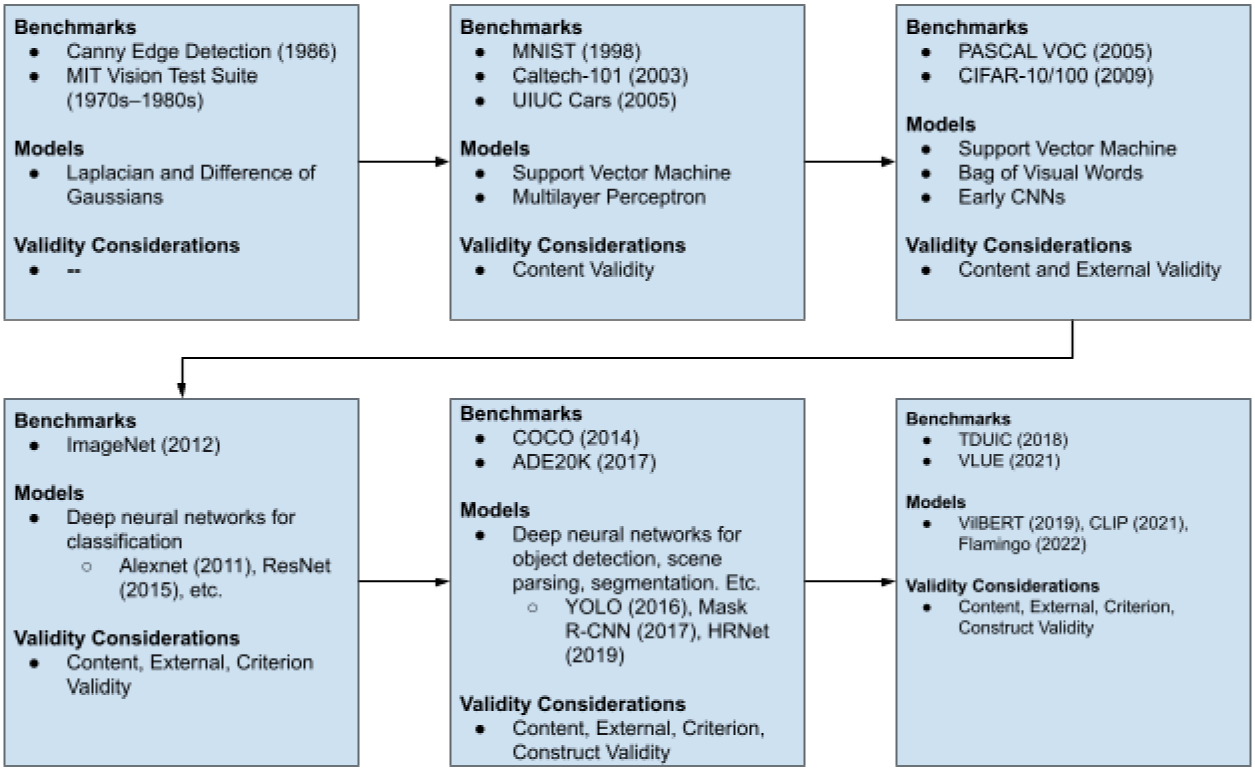

Benchmarks
Canny Edge Detection (1986)
MIT Vision Test Suite (1970s–1980s)
Models
Laplacian and Difference of Gaussians
Validity Considerations
--
Benchmarks
MNIST (1998)
Caltech-101 (2003)
UIUC Cars (2005)
Models
Support Vector Machine
Multilayer Perceptron
Validity Considerations
Content Validity
Benchmarks
PASCAL VOC (2005)
CIFAR-10/100 (2009)
Models
Support Vector Machine
Bag of Visual Words
Early CNNs
Validity Considerations
Content and External Validity
Benchmarks
ImageNet (2012)
Models
Deep neural networks for classification
Alexnet (2011), ResNet (2015), etc.
Validity Considerations
Content, External, Criterion Validity
Benchmarks
COCO (2014)
ADE20K (2017)
Models
Deep neural networks for object detection, scene parsing, segmentation. Etc.
YOLO (2016), Mask R-CNN (2017), HRNet (2019)
Validity Considerations
Content, External, Criterion, Construct Validity
Benchmarks
TDUIC (2018)
VLUE (2021)
Models
ViIBERT (2019), CLIP (2021), Flamingo (2022)
Validity Considerations
Content, External, Criterion, Construct Validity

Figure 4. Coevolution of benchmarks, models, and the type of validity necessary for common conclusions.

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. 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 (84) 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 (85) for vehicle detection and Caltech-101 (2003) (86) 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 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.

By the mid-2000s, large-scale benchmarks such as PASCAL VOC (2007) (87) introduced greater complexity, expanding evaluation beyond simple classification tasks. Later, in the late 2000s, CIFAR-10 and CIFAR-100 (88) further pushed the field toward standardized comparisons in object recognition. The shift from classification to more structured tasks like object detection and segmentation moved benchmarks toward broader contexts, with a stronger emphasis on generalization. Content validity and external validity became increasingly relevant as researchers began evaluating models across multiple datasets and questioning whether benchmark performance was a meaningful proxy for real-world vision tasks.

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 (10) and the ILSVRC (11) 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 general 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 (89), ADE20K (90)) 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 (91), VLUE (92), and TDIUC (93) 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 reasoning and understanding or merely exposed a model's ability to exploit statistical correlations in large datasets. Unlike earlier benchmarks, which primarily focused on accuracy, modern benchmarks aim to evaluate the latent properties of AI systems. 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 researchers 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.

Language

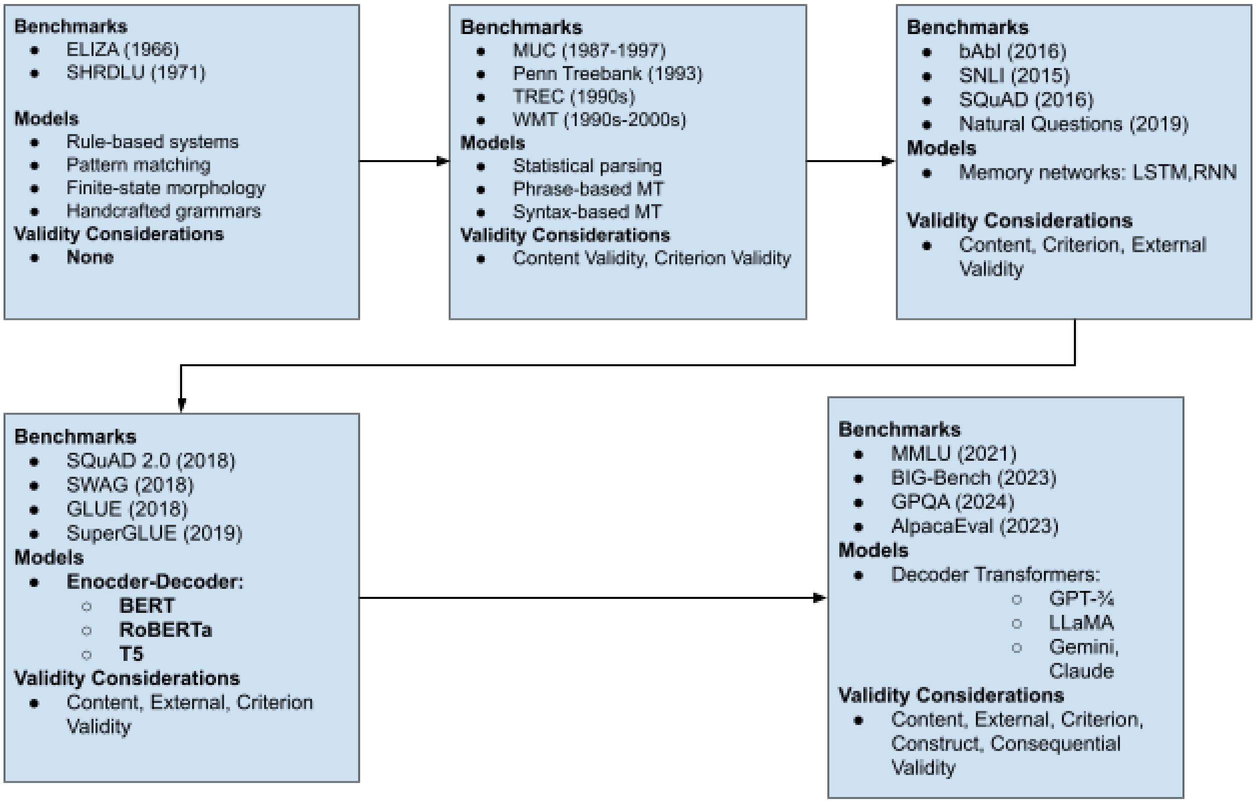


Figure 5. Coevolution of benchmarks, models, and the type of validity necessary for common conclusions for NLP.

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.

In the Blocks World Era (1960s-1980s), NLP evaluation was primarily qualitative and demonstration-based, lacking standardized metrics entirely. Systems like ELIZA (1966) and SHRDLU (1971) 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 and WMT 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 (94) and SNLI (95), 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 (96) and SentEval (97). 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 (98) 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 (71), 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 (73).