The Measurement Gap: Why AI Evaluation Needs Psychometrics

A White Paper — Noometric LLC

STEVEN KOSUTH-WOOD

Executive Summary

Artificial Intelligence systems are being deployed at scale into consequential domains — healthcare decisions, hiring, credit, criminal justice — before the field has developed a principled way to evaluate them. The dominant evaluation method, benchmarking, was borrowed from software engineering and was designed for a fundamentally different class of system. It cannot detect the problems that matter most: subtle bias, construct mismeasurement, and evaluation instruments that produce misleading results precisely when accuracy is most critical.

This paper argues that the solution already exist, in a field most AI practitioners have never heard of.

Psychometrics is the scientific discipline developed over the past eighty years to measure human psychological qualities: intelligence, personality, aptitude, bias. These qualities share a defining characteristic with AI behavior: they cannot be directly observed. They must be inferred from measurable outputs. Psychometrics developed rigorous methods for making that inference valid — methods that software engineering never needed because traditional systems operate on transparent, verifiable rules.

Large Language Models do not work using verifiable rules. They acquire structured tendencies from training data — implicit assumptions about language, evidence, and social categories — that function like psychological constructs. Evaluating them with only software engineering tools is analogous to diagnosing a patient using only an engineering inspection checklist. The instrument is not calibrated to what is actually being measured.

Noometric LLC applies psychometric measurement science to AI evaluation. The methodology has three components: defining what constructs the AI is intended to exhibit, validating that evaluation instruments actually measure those constructs, and analyzing fairness at the measurement level rather than the outcome level. This approach is not speculative — it is now formally endorsed by Microsoft Research, the Educational Testing Service, the National Institute of Standards and Technology, and the joint professional standards body of American psychology and education.

This paper explains the problem, the methodology, and the evidence that the approach works.

Contents

1. Introduction

The last decade has seen an unprecedented advancement in Artificial Intelligence (AI). Commercialization of the Large Language Model (LLM) by companies such as OpenAI and Anthropic has made this technology available to virtually everyone and has introduced a rate of change in society that, at times, seems difficult to govern.

A traditional information system works on the rules designed into it and is therefore deterministic in nature. The rules are known because the designers wrote them. The output of the system can be predicted based on those rules. AI-based systems work differently. Their output cannot be predicted from a set of designed rules because there are none. AI systems use the human brain as a model and are trained on data rather than programmed to follow explicit instructions. This makes them extraordinarily capable in some domains but, like humans, their behavior cannot be predicted absolutely.

Two categories of harm have emerged as AI deployment has grown. Hallucinations — outputs that are confidently stated but factually wrong — receive significant public attention and can be mitigated through cross-referencing and verification. The second category, bias, is more dangerous precisely because it is harder to detect. Bias embedded in an AI system's training data can systematically distort outputs in ways that are unknown to the system's operators and invisible to the people affected by those outputs. In order to identify and address bias, there must be a way to measure it.

Software engineering has well-developed techniques for evaluating traditional information systems, but those techniques assume there are known rules to verify against. They were not designed to detect bias, and they cannot reach the layer of an AI system where bias lives.

The field of psychometrics was developed to address exactly this problem — not in AI, but in human psychological assessment. Its methods have been refined for eighty years. This white paper describes the gap between what AI evaluation currently does and what psychometrics makes possible, and how Noometric LLC closes that gap for organizations that need to demonstrate rigorous, defensible AI evaluation.

2. A Brief History of AI Systems

Artificial Intelligence is not new. Natural Language Processing (NLP) and Expert Systems (ES) have been in development since the mid-twentieth century. The first attempts at AI encoded knowledge in the form of explicit rules. An engineer sat down with domain experts — in auto repair, for example — and wrote rules the system would follow. A system built this way could diagnose mechanical failures in automobiles. It was expensive, time-consuming, and limited in scope. The deeper problem was conceptual: a system designed on explicit rules is being imposed on a task that human experts do not actually perform using explicit rules. A skilled mechanic does not have a decision tree in mind. The diagnostic ability comes from experience, pattern recognition, and principles that operate automatically.

Researchers since the 1950s have explored biological neural systems as a model for better AI. Frank Rosenblatt's perceptron (1957) was the first practical computational model of a neuron. The hardware of the era could not support these approaches at useful scale, but since then, processing power has advanced at a rate that has made modern AI possible.

The AI technology in widespread use today is Machine Learning (ML) — a subfield of AI focused on algorithms that enable computers to learn from data rather than from programmed rules. The training data for these systems consists primarily of human-generated content: books, articles, scientific publications, internet text. The models are not programmed; they are shaped by this data. Table 2.1 traces the progression of ML approaches.

Table 2.1 — The Progression of Machine Learning

AI System Type	Era	Description
Expert Systems	1970s–80s	Hand-coded rules and logic; no learning from data
Statistical / Classical ML	1990s–2000s	Models that learn from data but are narrow and task-specific (Naive Bayes, SVMs, etc.)
Pre-transformer Neural NLP	2010s	Word embeddings (Word2Vec, GloVe), RNNs, LSTMs — better at language patterns but task-specific and limited in long-range context
Transformer models	2017–present	Self-attention (Vaswani et al., "Attention is All You Need") allows the model to relate any word to any other word in a sequence, enabling richer understanding at scale — the foundation of modern LLMs

The transformer architecture gave rise to the Large Language Model that has become widely known through products such as ChatGPT and Claude. For the purposes of this paper, the specific architecture matters less than a single fact: modern AI systems must be trained on data that is the product of human behavior, and that data carries human psychological qualities — including human biases.

Table 2.2 shows the most widely used transformer-based model types for text applications, along with the scale and curation level of their training data.

Table 2.2 — Types of Transformer-Based Models

Model Type	Examples	Approx. Training Data	Primary Sources	Curation Level
Encoder-only	BERT, RoBERTa	16 GB – 160 GB	Wikipedia, BooksCorpus, CommonCrawl	High
Encoder-decoder	T5, BART	~750 GB	C4 (filtered CommonCrawl), Wikipedia, task-specific paired data	Medium
Small Language Models (SLMs)	| Phi-2, Phi-3, Mistral 7B	~1 – 1.4 trillion tokens	Web text, synthetic "textbook-quality" data	High
Large Language Models (LLMs)	GPT-3/4, LLaMA 2/3, Claude	~570 GB – trillions of tokens	CommonCrawl, web scrapes, books, Wikipedia, code repositories	Low–medium

The training data volume and curation level figures in Table 2.2 are relevant to what follows: a model trained on trillions of tokens of largely uncurated web text has absorbed an enormous quantity of human opinion, assumption, and prejudice — with limited human review of what it absorbed.

3. The Problem: Where AI Evaluation Falls Short

AI has been adopted broadly and quickly. It is either currently in use by businesses and organizations to improve productivity or under active evaluation for that purpose. Given how deeply AI now touches the everyday lives of individuals, the potential for harm is receiving serious attention from governments, researchers, and the public.

Two forms of harm stand out. Hallucinations — factually incorrect outputs produced with apparent confidence — have been widely reported: court filings citing nonexistent cases, medical summaries with fabricated citations. These are serious but are, in principle, addressable by verifying AI outputs against authoritative sources before relying on them.

Bias is more dangerous in a specific way: it may be unknown to the operators of the system and invisible to those affected by it. Decisions that shape individuals' lives — credit decisions, recidivism risk scores, resume screening — have been demonstrably influenced by AI systems whose bias was not detected until after the harm occurred.

Cathy O'Neil's 2016 book *Weapons of Math Destruction* was among the first widely read accounts of how AI algorithms reinforce existing social inequalities. The same year, a ProPublica investigation found that an AI system used to rate defendants' recidivism risk was significantly more likely to incorrectly flag Black defendants as high-risk compared to white defendants at the same actual risk level.[1]

As documentation of these failures has grown, governments worldwide have begun to regulate AI — most urgently in industries already subject to oversight. In healthcare, the FDA's January 2025 draft guidance explicitly requires bias analysis and mitigation in marketing submissions for AI-enabled medical device software. Critically, this mandate does not come with a specified methodology. Organizations are required to produce a defensible evaluation — but the standard does not tell them how.

This is the regulatory vacuum Noometric fills.

4. Two Classes of Systems

To understand why a new evaluation methodology is needed for AI, the difference between traditional information systems and AI-based systems must be understood. For the purposes of this paper, a distinction between two system classes is useful:

> **Rule-Based Systems: I**nformation systems designed to perform a business function based on pre-existing rules or business logic. The system does what its designers programmed it to do.
>
> **Cognitive Systems:** Systems that process information through an underlying model of biological neural processing. This includes natural cognitive systems — the human mind — and artificial ones, such as an LLM.

The Open Systems Interconnection (OSI) model, a widely used framework in IT systems engineering, provides a useful structure for illustrating the difference. Rather than use all of its layers, two are sufficient here.

> The **physical layer** represents the computer hardware: CPUs, memory, storage.
>
> The **application layer** is where software applications reside: middleware, business logic, user interfaces.

Both rule-based and artificial cognitive systems operate at these two layers. They run on computer hardware and they implement software applications. Artificial cognitive systems add a third layer that rule-based systems do not have.

Table 4.1 — System Layer Comparison

Layer	Elements	Present In
Cognitive	Knowledge, learned associations, language understanding, embedded tendencies	Cognitive systems only
Application	Middleware, web pages, APIs, business logic	Rule-based systems and cognitive systems
Physical	Computer hardware (CPUs, RAM, storage)	Rule-based systems and cognitive systems

The Cognitive Layer is not a tangible software component — it emerges from the model's training and manifests in the model's outputs. The distinction between the cognitive layer and the application layer is analogous to the distinction between the mind and the brain: the brain (hardware) runs processes (software) that produce cognition (the cognitive layer).

This addition is where the evaluation challenge lives. Rule-based systems are deterministic; cognitive systems are not. The cognitive layer is not governed by verifiable rules, so the evaluation methods built for rule-based systems cannot reach it.

5. How Rule-Based Systems Are Evaluated

The standard evaluation methods for rule-based systems are verification and validation — terms defined in systems engineering standards:[2]

> **Verification:** The process of confirming that a system was built according to its design specifications. Did we build what we designed?
>
> **Validation:** The process of confirming that the design meets the needs of its intended users. Did we design the right thing?

These activities work when system behavior is deterministic — when the rules that govern output are known and can be checked against.

The design of a system is expressed through **requirements:** formal statements describing what the system must do. Requirements come in two types:

> **Functional requirements** describe what the system should do and how it should behave.
>
> **Non-functional requirements** describe how the system should perform: speed, reliability, scalability.

The evaluation approach for each type is illustrated by a straightforward example.

Table 5.1 — Evaluating a Simple Multiplication Application

Activity	Description
Requirement	The system shall accept two numbers from the user, multiply them, and output the result.
Design	Accept integers X and Y; compute X × Y; output the result
Verification	Given how multiplication works, will this design produce the correct output? |
Implementation	Write a software program that executes the design.
Test Oracle	| A multiplication table: a predetermined set of inputs with known correct outputs
Validation	Run the oracle against the application. Does it produce the correct output for each input pair?

For a non-functional requirement — say, processing speed — the evaluation takes the form of a **benchmark**: the system's performance is measured against a defined standard.

Table 5.2 — Non-Functional Requirement Evaluation

Activity	Description
Requirement	The system shall process 100 multiplication requests within 10 milliseconds.
Verification	Is this rate of processing achievable on the planned hardware?
Test	A script that sends 100 requests and records total processing time.
Validation	Does the application meet the time requirement?

All of this works because the behavior of a rule-based system is deterministic. Validation is possible because the expected behavior is fully specified in advance. Even for non-functional requirements, where external variables can cause variance, a range of expected results can be predicted. A benchmark for response time always produces a meaningful result because the system's behavior is governed by rules that are known.

When a cognitive layer is introduced — when an LLM is embedded in a system — this assumption breaks down for the qualities that matter most. Benchmarking LLM response time is still valid. Benchmarking LLM bias, fairness, or consistency of reasoning is not, for reasons that the next two sections establish.

6. How Cognitive Systems Are Evaluated: Lessons from Psychology

The human mind is a cognitive system in the same structural sense that an LLM is. The difference is that its physical layer is biological rather than silicon. Psychology faced the same evaluation challenge that AI faces today, decades earlier, and developed a principled solution.

In the early days of psychology, the Behaviorist school of thought took a defensible but ultimately limiting position: because the mind cannot be directly observed, it cannot be studied scientifically. Behaviorists restricted their research to measurable external behavior. The internal workings of the mind were treated as a black box, off limits. This position was scientifically conservative — and it eventually hit a ceiling. Important questions about human behavior could not be answered without looking inside that black box.

The response was the Cognitive school of thought, which held that abstract, unobservable concepts could be studied scientifically, provided that theoretical models of those concepts could be built and tested with empirical evidence. These unobservable concepts were called **constructs.**

> **Construct:** An abstract, unobservable concept used to explain, measure, and predict behavior — such as a personality trait (extraversion), an emotional state (anxiety), or a cognitive capability (intelligence).

The field of **Psychometrics** emerged to give constructs the measurability required for scientific validity. In a landmark 1955 paper, Cronbach and Meehl established the two foundational requirements for measuring a construct:[3]

> 1. Specify a nomological network of theoretical and empirical relationships that give the construct observable implications.
>
> 2. The construct must be embedded within that network in a measurable way. There must be observable indicators of it.

This framework introduced a new category of measurement problem. Once constructs could be measured, the validity of those measurements had to be evaluated. The following concepts emerged as the core measurement quality criteria in psychometrics:

> **Construct Validity:** The degree to which a test accurately measures the theoretical construct it is intended to measure.
>
> **Convergent Validity:** The degree to which a test correlates with other independent tests of the same construct. High correlation is expected if both are measuring the same thing.
>
> **Discriminant Validity**: The degree to which a test does **not** correlate with tests of different constructs. High correlation with an unrelated measure suggests the test is measuring something other than what it claims.

Measurement Invariance: The property that a construct is being measured the same way across different subgroups, so that scores are comparable across groups rather than reflecting group-specific test artifacts.

Interrater Reliability: The degree of agreement among multiple independent raters evaluating the same phenomenon — a measure of whether assessments are objective and reproducible rather than evaluator-dependent.

Each of these measurement properties has been studied by psychometrics for decades, and effective methods for detecting and quantifying failures in each have been developed and validated. They are the same measurement problems that AI evaluation faces today — and which no existing AI evaluation framework has been built to address.

7. Why Benchmarks Fail for AI

The measurement gap facing AI systems was established in the previous section: the introduction of a cognitive layer creates a class of behavior that software engineering evaluation was never built to reach. The response from the AI industry has been to adapt the benchmark approach from software engineering — measuring observable outputs against a reference standard. The problem is that this approach treats the cognitive system as a black box, exactly as the Behaviorists did in early psychology, and is vulnerable to the same limitations they eventually encountered.

The following three cases demonstrate how benchmark-only evaluation fails for AI systems with a cognitive layer, and what psychometric measurement would have revealed in each case.

Case 1: The Benchmark Measures the Wrong Construct

The Bilingual Evaluation Understudy (BLEU) is the standard benchmark for machine translation quality. It measures word-level similarity between a model's translation and a human reference translation. Studies have shown that translation models trained and evaluated on BLEU consistently default to male pronouns for professions such as engineer, doctor, and CEO, and female pronouns for nurse, teacher, and assistant — even when the source language has no grammatical gender (Finnish, Turkish, and Hungarian all lack gendered pronouns).[4]

BLEU scores do not penalize this because the human reference translations in the test set carry the same cultural assumptions. The benchmark is measuring "does it sound like a human translation?" — not "is it free of stereotyping?" These are different constructs. The benchmark was validated against the wrong one.

Psychometric diagnosis: A construct validity assessment would have required the benchmark designers to specify, before building the test, what construct "translation quality" actually means — and whether fairness was part of that construct. Convergent validity testing would have asked whether BLEU scores correlate with independent stereotype-detection measures. Discriminant validity testing would have checked whether BLEU fails to distinguish stereotyped from non-stereotyped output. Each of these steps would have exposed the mismatch before deployment.

Case 2: The Evaluation Instrument Itself Is Biased

In 2018, Joy Buolamwini and Timnit Gebru published the Gender Shades study at the ACM Conference on Fairness, Accountability, and Transparency.[5] Three major commercial facial recognition systems had been tested by their vendors against benchmark datasets and had achieved high accuracy ratings. Buolamwini and Gebru retested them on a more demographically balanced dataset and found error rates up to 34 percentage points higher for darker-skinned women compared to lighter-skinned men. The original benchmarks had used test sets that were predominantly white and male. A biased test set makes a biased model appear accurate — the evaluation concealed the bias rather than revealing it.

Psychometric diagnosis: Measurement Invariance testing via **Differential Item Functioning (**DIF) analysis. DIF asks whether individual test items function identically across demographic subgroups, *after controlling for overall performance level*. A DIF analysis on the original benchmark would have identified that images of darker-skinned women were systematically harder to classify — not because the models lacked general ability, but because the test itself was not measuring the same construct across groups. The benchmark was not measurement-invariant. This is a detectable property; it went undetected because the evaluation methodology had no concept of measurement invariance.

Case 3: Aggregate Accuracy Conceals Subgroup Bias

In 2018, Amazon built an AI system to screen resumes for technical roles and evaluated it against a standard accuracy metric: how well did the system's rankings predict who its hiring managers would have chosen? The system performed well on that benchmark.

The problem was that the benchmark criterion — "who would our hiring managers have chosen?" — was drawn from ten years of historical hiring data in which women were significantly underrepresented in technical roles. The system learned to systematically downrank resumes containing the word "women's" (as in "women's chess club" or "women's college"). It passed its benchmark and was still discriminatory. Amazon eventually scrapped the system.[6]

Psychometric diagnosis: Two psychometric concepts apply here.

Criterion Contamination: The criterion used to validate the model — historical hiring decisions — was itself biased. This is a known psychometric failure mode: when the validation standard reflects the same bias the test is supposed to detect, the evaluation cannot surface the problem.

Measurement Invariance testing via DIF: An analysis at the item (feature) level would have revealed that resume features associated with women functioned differently as predictors across demographic subgroups, even at equivalent overall qualification levels. Aggregate accuracy metrics cannot surface this; subgroup measurement analysis is required.

8. The Noometric Methodology

The preceding sections establish the problem precisely. AI systems have a cognitive layer that traditional software evaluation methods cannot reach. Psychometrics was developed specifically to measure that type of layer in human cognitive systems. Noometric's solution is to apply psychometric measurement science to the evaluation of artificial cognitive systems.

This is a methodological solution, not a product-specific one. The core methodology can be applied wherever AI evaluation is required — regulatory compliance, pre-deployment bias auditing, procurement, fine-tuning validation, or ongoing monitoring. The deliverables vary by context. The underlying measurement approach does not.

The methodology has three components.

Component 1: ***Construct Specification***

Before any measurement begins, the constructs to be evaluated are defined: what quality or capability is being assessed, what its theoretical definition is, and what observable behaviors constitute evidence of it. This step is the foundation of valid measurement. Without it, an evaluation produces a score with no defined referent — a number that cannot be interpreted, challenged, or improved upon.

Construct specification is not optional preprocessing. It is the design act that determines whether the rest of the evaluation is meaningful. The BLEU case illustrates what happens when this step is skipped: a benchmark is built, deployed at scale, and used to validate models for years before anyone specifies what construct it was actually measuring.

Component 2: ***Instrument Design and Validation***

Evaluation instruments — the tests, prompts, tasks, or benchmarks used to elicit model behavior — are themselves subject to psychometric evaluation. A valid instrument:

- Measures the construct it claims to measure (construct validity)
- Correlates with independent measures of the same construct (convergent validity)
- Does not conflate distinct constructs (discriminant validity)
- Produces consistent results under stable conditions (reliability)

These properties are verifiable. An instrument that lacks them produces data that cannot support conclusions about the construct of interest — no matter how large the dataset or how sophisticated the model being tested.

Component 3: ***Measurement-Level Fairness Analysis***

Detecting bias in an AI system requires a different question than current AI evaluation practice typically asks. The standard approach identifies outcome disparities: group A received different scores or results than group B. This is a useful signal, but it is not a measurement of bias.

Outcome disparities can reflect three distinct underlying situations: the groups genuinely differ on the construct being measured; the measurement instrument is miscalibrated for one group; or both. These cannot be distinguished from outcome data alone. The psychometric definition of bias is precise:

Bias occurs when a measure systematically under- or over-represents a construct for a subgroup — not because the construct differs between groups, but because the measurement instrument is miscalibrated for that subgroup.

Differential Item Functioning (DIF) asks whether individual evaluation items function identically across subgroups after controlling for the underlying construct level. If an item produces systematically different results for one subgroup at the same construct level, the item is biased — the problem is in the instrument, not the population.

Differential Algorithmic Functioning (DAF) extends this framework to algorithmic decision systems, including AI outputs. Suk and Han (2025) formally established DAF as an analysis framework, including intersectional analysis across multiple protected characteristics simultaneously.[7] Applied to AI evaluation, DIF and DAF locate bias precisely: they identify not just that a model treats groups differently, but where in the measurement structure that difference originates and whether it reflects construct variance or instrument miscalibration.

Together, these three components constitute a measurement framework that is principled, auditable, and grounded in decades of validated psychometric practice.

What Bias Actually Means — and Why AI Gets It Wrong

In common usage, "AI bias" refers to disparate outcomes across demographic groups. This is a real problem. It is also the surface manifestation of a deeper one.

The psychometric definition above matters because it distinguishes between two different situations that look identical from outside the system. A test that gives different scores to two groups is not necessarily biased — the groups might genuinely differ on the construct being measured. A test is biased when it gives different scores for the same underlying level of the construct. The error is in the instrument, not the population.

AI evaluation almost never makes this distinction. It detects outcome disparities and calls them bias without asking whether the measurement instrument is miscalibrated, what constructs are being measured, or whether those constructs are the appropriate ones to measure.

Why This Matters for Large Language Models

Large Language Models are not neutral text processors. Training datasets introduce selection bias when data chosen for training under- or over-represents subgroups. Training techniques compound this. **Reinforcement Learning from Human Feedback** (RLHF) — a technique used to shape model behavior toward human preferences — introduces the biases of the human trainers who provide that feedback.[8]

These encoded tendencies are not random noise. They are structured: they cluster into coherent patterns that psychologists would recognize as latent constructs — stable, underlying dimensions that organize a wide range of surface behaviors.

An LLM does not simply have "some bias." It has an implicit political epistemology, a latent risk-tolerance model, assumptions about social causality, an embedded theory of what counts as credible evidence. These constructs shape every output the model produces — often invisibly to both operators and users. Identifying and measuring them requires exactly the tools psychometrics was built to provide.

9. Academic and Institutional Validation

The application of psychometrics to AI evaluation is not a speculative proposal. It is an emerging academic field with a growing body of peer-reviewed literature, and it has received formal endorsement from the US government's primary measurement standards body and the world's leading testing institution.

Academic Convergence

The field now has a recognized name — **AI Psychometrics** — and a growing publication record.

Pellert et al. (2024), published in *Perspectives on Psychological Science*, formally established AI Psychometrics as a discipline, demonstrating that psychometric inventories can be used to assess the psychological profiles of LLMs.[9]

A 2025 systematic review documented the field's rapid growth and proposed a three-dimension evaluation framework aligned with Noometric's approach: target construct, measurement method, and validation.[10]

Wang et al. (2023) from Microsoft Research, published in *Communications of the ACM*, validated the core argument in one of the field's most prominent peer-reviewed venues. Their framework maps directly to Noometric's methodology: construct identification → construct measurement → test validation.[11]

Jacobs and Wallach (2021), published at the ACM Conference on Fairness, Accountability, and Transparency — the primary peer-reviewed venue for AI fairness methodology — argued that AI fairness research has a fundamental measurement problem: it defines fairness in terms of outcomes without a theory of what is being measured. The paper explicitly calls for importing psychometric measurement frameworks into AI fairness work.[12] This is the closest existing academic work to Noometric's core thesis.

Domain-Specific Validation

The Psychometrics Centre at Cambridge Judge Business School (2025) applied this approach to generalist medical AI and found that three psychometric factors — reasoning, comprehension, and core language modeling — account for 82% of the variance in LLM performance across 27 cognitive tasks in Stanford's HELM benchmark. The result was published in the *Journal of Medical Internet Research*, targeting the exact regulatory audience — policymakers and health regulators — that Noometric serves.[13]

A 2025 paper in *Nature Machine Intelligence* demonstrated that psychometric frameworks not only evaluate LLM personality traits but can verifiably shape downstream model behavior — showing the approach has both diagnostic and prescriptive value.[14]

Institutional Endorsement

The strongest validation comes not from academic literature but from the institutions that set professional standards for measurement science in the United States.

The **Educational Testing Service** (ETS) — the organization responsible for the SAT, GRE, and most major US standardized assessments, and the world's authoritative body on psychometric methodology for over eighty years — published RR-25-03 (2025): "Responsible AI for Measurement and Learning." It applies the same psychometric validity, reliability, fairness, and DIF framework to AI systems that Noometric proposes. ETS is not a commercial competitor; they apply this framework to their own internal work. They function as the field's credibility anchor.[15]

NIST AI 800-3 (2026), published by the National Institute of Standards and Technology, formally distinguishes benchmark accuracy from generalized accuracy (a psychometric validity concept), explicitly acknowledges that Item Response Theory and psychometric frameworks should be applied to LLM evaluation, and requires evaluators to disclose their statistical assumptions. This is a US government publication endorsing the methodology.[16]

The **AERA/APA/NCME Standards for Educational and Psychological Testing** (7th edition, 2025) — the joint professional standard of the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education — is the authoritative specification for valid psychometric methodology in the United States. The 7th edition, for the first time, explicitly addresses AI and automated scoring. A Noometric engagement aligned to these standards asserts that its methodology meets the same bar as professional licensure examinations.[17]

Summary

Microsoft Research, Cambridge University, ETS, NIST, and the broader academic community have independently arrived at the same conclusion. The regulatory standards bodies have now codified it. The gap between the academic consensus and current AI industry practice is real, documented, and growing. Noometric is the practitioner layer that translates these frameworks into compliance-ready client engagements.

10. Market Demand and Regulatory Mandate

The Market Is Large and Growing

The AI model evaluation platform market was $1.86 billion in 2025, projected to reach $2.36 billion in 2026 — a 27.3% compound annual growth rate driven by early AI deployment failures, expanding regulatory scrutiny, and enterprise AI scaling demands.[18] The AI consulting market overall crossed $65 billion in 2026, growing at over 20% annually, with the evaluation and governance segment growing faster than the overall market.[19]

Global AI governance and compliance spending specifically is projected to reach $2.54 billion in 2026 and $8.23 billion by 2034.[20] The scale of this expansion reflects a structural shift: AI governance is moving from voluntary best practice to mandatory compliance.

A market signal worth noting: in March 2025, CoreWeave acquired Weights & Biases for $1.7 billion explicitly to expand from infrastructure into AI model assessment. Evaluation capability is now considered a strategic asset worth acquiring at billion-dollar scale.[21]

Regulatory Mandates Are Creating Non-Discretionary Demand

Across employment, healthcare, and federal procurement, regulation is requiring AI evaluation that existing frameworks cannot produce.

Employment: Annual independent bias audits are legally required in multiple jurisdictions. New York City Local Law 144 mandates annual audits of automated employment decision tools, with civil penalties for non-compliance and no approved auditor list — creating demand for qualified external methodology. The EEOC has formally extended Title VII to AI hiring tools and made employers, not vendors, fully liable for discriminatory outcomes.

Healthcare: The FDA's January 2025 draft guidance requires bias analysis and mitigation in marketing submissions for AI-enabled medical device software. OCR enforcement actions targeting AI rose 340% in 2025. These requirements specify what must be demonstrated — not how to demonstrate it.

Federal procurement: IEEE 2857-2024 AI performance benchmarking is now mandatory for US federal AI procurement. OMB M-26-04 requires federal agencies purchasing LLMs to request model cards and evaluation artifacts.

State regulation: In 2025, lawmakers in 47 states introduced over 250 bills regulating AI; 33 were signed into law across 21 states. California alone has enacted 38 or more AI-related laws, including SB 53 (effective January 1, 2026), which requires frontier model developers to publish safety protocols, with penalties of up to $1 million per violation.

The critical pattern across all of these mandates is the same: regulators specify the outcome required — a bias audit, an evaluation report, a documented testing protocol — but do not specify the methodology. Organizations must define their own testing approach and defend it. This is the vacancy Noometric fills.

The Methodology Gap Is Currently Unoccupied

A review of current AI bias audit providers finds no commercial competitors applying Differential Item Functioning or psychometric measurement theory to AI evaluation. The gap between what regulators are requiring and what the market currently provides is documented in peer-reviewed literature: publications in **Nature Biomedical Engineering** and **npj Digital Medicine** have explicitly identified that current tools in healthcare AI rely on group fairness statistics with no underlying measurement theory.

The average enterprise AI initiative investment is $1 million to $2.6 million per use case. Noometric's engagement range — $15,000 to $75,000 — represents a small insurance cost against regulatory exposure that can cost ten to one hundred times more, including $1 million per violation under California's SB 53.

11.About Noometric

Noometric LLC is an AI evaluation consulting firm founded on the thesis that eighty years of measurement science should not have to be rediscovered by the AI industry. The firm applies psychometric methodology — construct specification, instrument validation, and measurement-level fairness analysis — to AI evaluation in contexts where rigorous, defensible results are required.

Noometric's approach is grounded in the canonical psychometric literature, aligned with the AERA/APA/NCME professional standards, and consistent with the methodology independently endorsed by ETS, NIST, and the emerging academic field of AI Psychometrics. Engagements are delivered as consulting projects scoped to the client's evaluation need: pre-deployment bias audits, regulatory compliance evaluations, procurement review, or fine-tuning validation.

For further information:

Website: http://noometric.com

Email: steve.kosuthwood@noometric.com

Mail: Attn Noometric LLC - Steven Kosuth-Wood
12312 Alta Panorama
North Tustin, CA 92705

12.Notes

[1] Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016, May 23). Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks. *ProPublica*. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[2] Definitions adapted from IEEE Std 1012-2016, IEEE Standard for System, Software, and Hardware Verification and Validation.

[3] Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281–302. https://doi.org/10.1037/h0040957

[4] Stanovsky, G., Smith, N. A., & Zettlemoyer, L. (2019). Evaluating gender bias in machine translation. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 1679–1684. https://doi.org/10.18653/v1/P19-1164

[5] Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. *Proceedings of Machine Learning Research*, *81*, 1–15. http://proceedings.mlr.press/v81/buolamwini18a.html

[6] Dastin, J. (2018, October 10). Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters*. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G

[7] Suk, Y., & Han, K. T. (2025). Evaluating intersectional fairness using intersectional DAF. *Journal of Educational and Behavioral Statistics*. https://doi.org/10.3102/10769986241269820

[8] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, *35*. https://arxiv.org/abs/2203.02155

[9] Pellert, M., Lechner, C. M., Wagner, C., Rammstedt, B., & Strohmaier, M. (2024). AI psychometrics: Assessing the psychological profiles of large language models through psychometric inventories. *Perspectives on Psychological Science*. https://doi.org/10.1177/17456916231214460

[10] Rao, A., et al. (2025). Large language model psychometrics: A systematic review of evaluation, validation, and enhancement. *arXiv*. https://arxiv.org/abs/2505.08245

[11] Wang, P., Li, L., Chen, L., Zhu, D., Lin, B., Cao, Y., Liu, Q., Liu, T., & Sui, Z. (2023). Evaluating general-purpose AI with psychometrics. *Communications of the ACM*. https://dl.acm.org/doi/full/10.1145/3769688

[12] Jacobs, A. Z., & Wallach, H. (2021). Measurement and fairness. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 375–385. https://dl.acm.org/doi/10.1145/3442188.3445901

[13] The Psychometrics Centre, Cambridge Judge Business School. (2025). Beyond benchmarks: Evaluating generalist medical AI with psychometrics. *Journal of Medical Internet Research*. https://pmc.ncbi.nlm.nih.gov/articles/PMC12129431/

[14] Rao, A., et al. (2025). A psychometric framework for evaluating and shaping personality traits in large language models. *Nature Machine Intelligence*. https://www.nature.com/articles/s42256-025-01115-6

[15] Educational Testing Service. (2025). *Responsible AI for measurement and learning* (ETS Research Report RR-25-03). https://www.ets.org/Media/Research/pdf/RR-25-03.pdf

[16] National Institute of Standards and Technology. (2026). *Expanding the AI evaluation toolbox with statistical models* (NIST AI 800-3). https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.800-3.pdf

[17] American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2025). *Standards for educational and psychological testing* (7th ed.). https://www.testingstandards.net/

[18] GlobeNewswire. (2026, April 14). *AI model evaluation platform market research report 2026*. https://www.globenewswire.com/news-release/2026/04/14/3273083/28124/en/AI-Model-Evaluation-Platform-Market-Research-Report-2026-AWS-Google-Microsoft-and-IBM-Set-Industry-Standards-for-Performance-and-Reliability-Long-term-Forecast-to-2030-and-2035.html

[19] RTS Labs. (2026). *9 best AI consulting firms for enterprises: 2026*. https://rtslabs.com/top-ai-consulting-firms/

[20] SQ Magazine. (2026). *AI compliance cost statistics 2026*. https://sqmagazine.co.uk/ai-compliance-cost-statistics/

[21] Coldewey, D. (2025, March 4). CoreWeave acquires AI developer platform Weights & Biases. *TechCrunch*. https://techcrunch.com/2025/03/04/coreweave-acquires-ai-developer-platform-weights-biases/