

Measuring the User, Not the Model: Critical AI Literacy and the Verification Frontier

Derek C. Briggs

University of Colorado Boulder

September 12, 2026

Correspondence: derek.briggs@colorado.edu

Abstract

The AI community routinely speaks of measuring the intelligence or capability of large language models (LLMs). Examined against the core elements of measurement, the Theory, Instrumentation, Scales-and-units, and Modeling (TISM) framework, this usage is difficult to justify. Measurement presupposes variability in the attribute of interest, yet LLM development treats variability, within a model and across models, as a design parameter to be controlled rather than an attribute to be discovered. I argue that an LLM is better understood as a non-stationary measuring instrument than as an object of measurement, and that the variability worth measuring lies in how human beings interact with these systems. I sketch a measurement model for one dimension of AI literacy, critical literacy, the capacity to detect, appraise, and override plausible-but-wrong AI output, and use a simulation to show that this construct is most measurable precisely in domains where verification is expensive and the AI can least help on its own. This is a conceptual work in progress intended as a jump-off point for empirical study.

Keywords: measurement theory; TISM; artificial intelligence; AI literacy; item response theory; large language models.

1 Introduction

Over the past two years, generative AI has moved from novelty to a primary collaborator in my research and teaching. Using it well, however, means reckoning with what Dell'Acqua and colleagues (2023) called a “jagged technological frontier”: LLMs can solve problems that would stump an expert yet stumble on tasks humans find trivial. Asked to write a genuinely funny joke about pelicans, a frontier model offers a serviceable pun — a pelican walks into a bar, and the bartender asks, “Why the big bill?” — competent, but rarely funny in the way a gifted comedian is funny. The same system will, minutes later, debug code written to implement a parameter estimation algorithm. Navigating a

jagged frontier requires domain expertise sufficient to use generative AI critically.

It has become common to describe progress along this frontier as the measurement of model intelligence or capability through multiple-choice benchmarks, head-to-head “arena” leaderboards, and Elo-style ratings (Chollet, 2019; Singh et al., 2025). But is this measurement in a sense widely understood in the human and physical sciences?

In answering this question, I take as a reference point the TISM framework (Briggs, Maul, & McGrane, 2025), under which measurement is conceptualized as the interaction of four components: substantive theory (T) about the construct, instrumentation (I) sensitive to variation in it, scales and units (S) on which results are reported, and modeling (M) of uncertainty. Applying TISM to LLM

“measurement” exposes a missing T — there is no substantive theory about variability in any construct of interest.

2 Is an LLM an Object of Measurement?

Instrumentation, scales, and modeling are not, in principle, beyond reach if the idea to measure something about an LLM. The decisive difficulty is variability. Measurement exists to characterize variation in an attribute, and the human sciences inherited two complementary notions of it (Briggs et al., 2025): intra-individual variation, studied by the psychophysicists (Fechner, Thurstone), and inter-individual variation, the province of the individual-differences tradition (Galton, Spearman).

Neither form of variation is absent in an LLM, but neither has the status that measurement requires. Within a model, output varies with the sampling temperature, with the phrasing and ordering of a prompt, with the surrounding context, and with nondeterminism in inference. Across models, dozens of systems differ in architecture, training data, and configuration, and any one of them can be run under many settings. The point is not that this variation does not exist. It is that it consists of design parameters under the developer's control rather than the attribute of an object that an instrument could discover. Temperature can be set to zero. Sensitivity to prompt wording is treated as a defect to be engineered away. Frontier systems are optimized toward the same benchmarks, which compresses between-model differences rather than preserving them. Measurement in the human sciences presupposes variation that is there to be found. LLM development treats variation as something to be managed and, where possible, removed.

The implication is not that measurement is irrelevant to AI, but that we are pointing it at the wrong object. An LLM is better treated as a basis for better instrumentation — a powerful but non-stationary instrument — than as a person-like object whose latent trait we estimate. The variability that remains abundant, and that matters, is in how humans interact with these instruments. Put simply, we should *measure* human attributes and *evaluate* model capabilities. These are different epistemic activities. Evaluation asks whether an AI system's outputs are fit for a use; measurement asks whether there is a quantitative attribute to be estimated. They should not be conflated.

3 A Model for Critical AI Literacy

If there is a human construct worth measuring here, a strong candidate is generative AI literacy: a set of largely domain-general habits of mind that are nonetheless developed in domain-specific ways. I focus on one dimension, which I will define as *critical AI literacy*: the capacity to detect and repair, or where no objective check exists, to appraise and override, a plausible-but-wrong AI output that is not easily verified.

Consider a single interaction as an item. The stimulus is an LLM response to a prompt. Across the items in a bank, some responses are correct, and some contain one or more errors. When presented with one of these items, a person's task is to decide whether to accept the response as correct, or to override it if there is detectable error. Let $X = 1$ if the person's judgment is correct, meaning that either a correct response was accepted or an erroneous one was overridden, and $X = 0$ otherwise. Let c denote the proportion of responses in the bank that are correct. Because the bank is generated by the model, c is the model's capability on the domain from which the items are drawn. Items fall into two domains indexed by their verifiability v . When $v = 1$ the task is easily verifiable — a coding or statistics problem, a factual question — and ground truth is available at once. When $v = 0$ the task is unverifiable, or only expensively and post hoc verifiable — whether a story is funny for a particular dinner party, whether a presentation succeeds — and any verdict is deferred and social rather than immediate.

Whether a response is erroneous is a property of the item, independent of the person. Let E denote the event that the response is erroneous and C the event that it is correct, so $P(E) = 1 - c$ and $P(C) = c$. A person's behavior on an item is described by two conditional probabilities: $h(\theta) = P(\text{override} | E, \theta)$, the probability that a person with critical literacy θ catches an erroneous response, and $f(\theta) = P(\text{override} | C, \theta)$, the probability that the same person falsely identifies a correct response as erroneous. The decision is correct, $X = 1$, when an erroneous response is overridden or a correct one is accepted. By the law of total probability over the two states of the response,

$$\begin{aligned} P(X = 1 | \theta) &= P(X \\ &= 1 | C, \theta) P(C) + P(X \\ &= 1 | E, \theta) P(E) \\ &= c [1 - f(\theta)] + (1 - c) h(\theta). \quad (1) \end{aligned}$$

This decomposition holds for any h and f . To obtain a model I make two assumptions. First, critical literacy raises the probability of overriding an error without raising the probability of overriding a correct response: $f(\theta) = 0$ for all θ . This is a claim that critical literacy is something akin to a continuum of honed discernment that is generally independent of ingrained skepticism. Second, $h(\theta)$ is a logistic function of the gap between the person's critical literacy and the difficulty of overriding the item:

$$h(\theta) = \frac{1}{1 + e^{-a(\theta - \delta(s))}}. \quad (2)$$

Under these two assumptions,

$$P(X = 1 | \theta) = c + (1 - c) h(\theta). \quad (3)$$

Here c is the accuracy of a person who accepts every response as correct, and it is each item's lower asymptote: as $\theta \rightarrow -\infty$, $h(\theta) \rightarrow 0$ and $P(X = 1 | \theta) \rightarrow c$. A person with no critical literacy takes the LLM responses at face value, so that person's accuracy is defined by the model's capability. The parameter $\delta(s)$ is the value of θ at which a person has an even chance of overriding an error (relative to accepting it); it increases with the seductiveness s of the LLM response, the degree to which stylistic features of the response make a flaw harder to discern. The discrimination a is held constant across items. This is a simplifying assumption made for illustration; the substantive assumption is that override difficulty varies with seductiveness. The model further assumes that items are conditionally independent given θ : that whether a response is correct is independent of the person, and that the person's decision depends only on θ and the item's $\delta(s)$. Formally this is a constrained version of the three-parameter logistic item. The lower asymptote, however, is not a nuisance guessing parameter; it is a substantive parameter of interest.

The assumption $f(\theta) = 0$ is what makes c the lower asymptote, and I make this explicit because it may be unrealistic. That is, there may be people who will tend to override a correct LLM response irrespective of their critical literacy and the seductiveness of the response. Signal detection theory makes the alternative precise. Suppose a person overrides when the evidence that a response is erroneous exceeds a criterion k , and suppose the separation between erroneous and correct responses depends on θ and on s . Then $h(\theta)$ depends on both θ and k , while the probability of overriding a correct response depends on k alone: it reflects the person's

disposition toward skepticism, not their ability to discern an AI error. Under this relaxation the lower asymptote of $h(\theta)$ is no longer 0 but f : as $\theta \rightarrow -\infty$ the person cannot distinguish erroneous from correct responses and overrides both at the same rate, so $h(\theta) \rightarrow f$. Allowing this probability to equal some constant $f > 0$ lowers the floor from c to $c(1 - f) + (1 - c)f$, which is below c whenever $c > .5$, because indiscriminate overriding is costly when the model is usually right. It does not change the conclusion of the simulation that follows, for a reason given in Section 4. A derivation under this more complicated scenario is given in Appendix A.

Figure 1 plots the item characteristic curve for Equation 1. The parameter c is the floor of the curve, the accuracy of a person who accepts every response; as model capability rises, the floor rises and the room for critical literacy to matter shrinks. The parameter $\delta(s)$ is the curve's location; more seductive responses shift the curve to the right. The measurement value of an item is its Fisher information, the expectation of $[P'(\theta)]^2 / [P(\theta)(1 - P(\theta))]$ over a population of users with $\theta \sim N(0, 1)$.

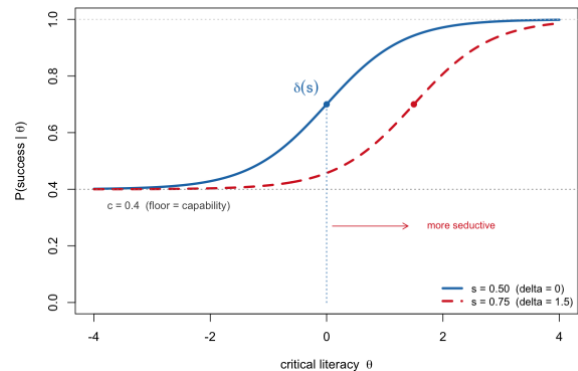


Figure 1: A hypothetical item characteristic curve. The floor c is the accuracy of a person who accepts every response; the location $\delta(s)$ shifts right as the LLM responses grow more seductive.

4 A Simulation: The Verification Frontier

To see how the measurability of critical literacy can be expected to change as models improve, I trace two domains across successive model generations. All computations use the model of Section 3 with $f(\theta) = 0$, $a = 1.5$, $\delta(s) = -3 + 6s$ so that $s \in [0, 1]$ maps onto $\delta \in [-3, 3]$, and $\theta \sim N(0, 1)$. Both domains begin at the same early model ($c = .50, s = .45$). With each generation of models, the LLM responses, and the errors that remain in, become subtler, so seductiveness rises identically in both domains, to

$s = .62$ at the frontier. Capability diverges in the cheaply verifiable domain a clean training signal drives c toward a ceiling near 1 (.96), whereas in the unverifiable domain, where there is no objective signal to optimize against, capability plateaus at .55.

Figure 2 maps the per-item information over the space of capability and seductiveness, with the two domain trajectories overlaid. Information about critical literacy lives on a ridge, the verification frontier, at intermediate seductiveness, and falls toward zero where the model is so capable that it rarely makes an error. The verifiable road climbs rightward into this low-information region; the unverifiable road remains near the ridge.

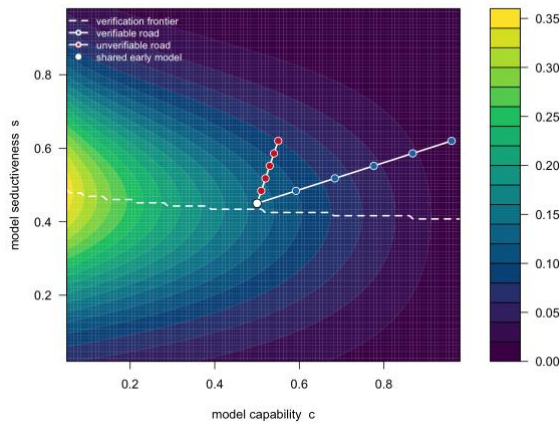


Figure 2: Per-item information about critical literacy over model capability (c) and seductiveness (s). The two roads start at the same early model and fork: the verifiable road slides into the low-information region, the unverifiable road stays on the ridge.

Table 1 gives the numbers. Items needed are computed from the average per-item information $\bar{I}$ as follows. For θ with unit variance, reliability is $\rho = I/(I + 1)$, so $\rho = .80$ requires total information $I = 4$ and $n = 4/\bar{I}$ items. Because $\bar{I}$ is an expectation over the population, n is the number of items of a given type needed to reach that reliability for a randomly drawn person. At the frontier, the verifiable domain yields almost no information per item; one would need roughly 719 items to estimate critical literacy at a reliability of .80. The unverifiable domain still yields usable information and requires about 50. Both domains were equally measurable at the start, at about 30 items.

Model	c	s	Avg. info / item	Items for $\rho = .80$
Shared early LLM	.50	.45	.1323	30
Frontier, verifiable (code)	.96	.62	.0056	719
Frontier, unverifiable (joke)	.55	.62	.0801	50

Table 1: Average Fisher information per item and the number of items needed for a reliability of .80, by model and domain.

The pattern is a consequence of the functional form rather than an empirical finding. As $c \rightarrow 1$, $P(X = 1 | \theta) \rightarrow 1$ for every θ , so the item stops discriminating among people. The simulation illustrates the mechanism; it does not estimate its magnitude, and the specific numbers depend on the parameter values just listed.

The lesson is counter-intuitive: critical AI literacy is most measurable exactly where the AI model, on its own, can help the least. Where verification is cheap and the model is strong, the construct collapses to whether the user bothered to run the check; where verification is expensive and judgment is irreducible, the construct persists.

5 Discussion

This relocation of measurement — from the model to the human — comes with a challenge: we must measure the human through a non-stationary instrument engineered to compress the very variance we wish to detect. The situation resembles measuring penmanship in the age of the word processor, or navigation skill in the age of GPS: the tool raises the floor and can erase the inter-individual differences that once defined the construct.

There is another cost. Because capability enters as a lower asymptote, the item departs from the Rasch model and forfeits specific objectivity: comparisons of critical literacy between two people are no longer invariant to the particular model or task through which they were measured. Measuring a human trait through a capable AI is non-invariant by construction.

Two broader conclusions follow. First, to speak of measuring an LLM's capability as a latent trait is, absent a substantive theory of the construct, largely a category error. Measurement has not become irrelevant; it has been relocated to the human, and made harder to accomplish by the very instrument that occasions it. Second, the AI-

literacy constructs that matter now are socially constructed and in flux; by the standards of physical measurement we are at the thermoscope stage, with much theoretical work still to do (Briggs et al., 2025).

A complementary perspective deserves note. Casabianca (2025), in a position paper titled “Psychometrics is all you need,” argues that AI development and evaluation should be reimagined within a validity framework, using evidence-centered design and the collection of validity evidence to defend the use of an AI application's outputs in a given context. Casabianca works in the argument-based validation tradition (Kane, 2013): her concern is whether evidence supports a claim about the fitness of a system's outputs for a use — a discipline AI evaluation badly needs, and one she grounds in construct definition and rater science (she names the “fluency effect,” our seductiveness, by which judges conflate style with correctness). The present paper asks a prior question from a measurement-realist tradition (Briggs et al., 2025): in what sense is there a quantitative attribute, carrying the variability that measurement presupposes, to be measured at all? For the LLM I argue there is not — which is why we relocate measurement to the human, leaving the evaluation of model capabilities to a validity argument of the kind Casabianca details. As I see it, one measures human attributes and evaluates model capabilities; hers is the rigorous account of the latter, ours of the former. The slippage worth flagging is terminological: what “psychometrics is all you need” supplies is validation, not measurement.

6 Toward an Empirical Program

If critical literacy is most measurable where verification is expensive or deferred, and yet scoring the outcome *X* requires that very verification, has measurement simply been relocated to where it cannot be performed?

The objection conflates two costs that measurement ordinarily keeps apart: the one-time cost of calibrating an instrument and the recurring cost of administering it. We do not verify an examinee's answer as it is given; we compare it against a key fixed when the test was built. The same separation applies here. One can assemble a bank of AI outputs in hard-to-verify domains, each carrying an expert-validated verdict, and pay the verification cost once, during test construction. The bank must contain both correct and erroneous responses in known proportion; without erroneous responses critical literacy

cannot be identified, as Section 4 showed. The simulation above treats a single representative item; a fuller treatment lets the bank's items vary in seductiveness, so that the information returned depends on how well the bank's difficulties are targeted to the population. What matters here is only that the expense is borne once, by the researcher, rather than at each measurement.

Crucially, unverifiable in the wild need not mean unkeyable in the laboratory. Stimuli can be constructed whose correct judgment is known by design: an AI argument with a buried non sequitur, a summary that omits an essential qualifier, a recommendation that overgeneralizes. The seductiveness — the surface fluency that hides the flaw — is genuine, but the key is fixed in advance by expert consensus.

This clarifies what such a measure is for. Its value is predictive: critical literacy estimated in a lab-like setting is worth measuring precisely because it should forecast the quality of a person's judgments in real world practice, where verification is absent and no key will ever arrive — the high-stakes case of a professional relying on AI where errors are subtle and checking is impractical. Demonstrating that laboratory-measured critical literacy predicts real-world success is the central validation task and the natural next study. The open question raised by the simulation — whether, as seductiveness rises while capability plateaus, critical literacy in these domains persists or eventually erodes as even discerning users are fooled — can be posed in exactly this calibrated setting.

None of this rescues the genuinely unverifiable extreme. Where no convergent verdict exists even in principle, there is no key to establish, and critical literacy is not merely expensive to measure but undefined. The tractable target is the broad middle: domains of real, consequential quality in which a defensible verdict exists at a manageable cost.

References

- Briggs, D. C., Maul, A., & McGrane, J. A. (2025). On the nature of measurement. In L. L. Cook & M. J. Pitoniak (Eds.), *Educational measurement* (5th ed., pp. 1513–1578). Oxford University Press.
<https://doi.org/10.1093/oso/9780197654965.003.0021>
- Casabianca, J. M. (2025). Psychometrics is all you need [Preprint]. EdArXiv.
https://doi.org/10.35542/osf.io/7w6pz_v1
- Chollet, F. (2019). On the measure of intelligence. arXiv:1911.01547.
<https://doi.org/10.48550/arXiv.1911.01547>

- Dell'Acqua, F., McFowland, E., Mollick, E. R., Lifshitz-Assaf, H., Kellogg, K., Rajendran, S., Kraye, L., Candelon, F., & Lakhani, K. R. (2023). Navigating the jagged technological frontier (Harvard Business School Working Paper No. 24-013). <https://doi.org/10.2139/ssrn.4573321>
- Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50(1), 1–73. <https://doi.org/10.1111/jedm.12000>
- Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Addison-Wesley.
- Mari, L., Wilson, M., & Maul, A. (2023). *Measurement across the sciences: Developing a shared concept system for measurement* (2nd ed.). Springer. <https://doi.org/10.1007/978-3-031-22448-5>
- Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests*. University of Chicago Press.
- Singh, S., Nan, Y., Wang, A., D'Souza, D., Kapoor, S., Üstün, A., Koyejo, S., Deng, Y., Longpre, S., Smith, N. A., Ermis, B., Fadaee, M., & Hooker, S. (2025). The leaderboard illusion. arXiv:2504.20879. <https://doi.org/10.48550/arXiv.2504.20879>

Appendix A. The Floor Under a Fixed Override Criterion

Section 3 adopts the assumption $f(\theta) = 0$, under which the lower asymptote of an item is c . This appendix derives the lower asymptote when that assumption is relaxed in the way signal detection theory suggests, and shows why the relaxation does not restore information about critical literacy as model capability rises.

Suppose that on each item the person observes evidence e that the response is erroneous and overrides when e exceeds a criterion k . Let erroneous responses generate evidence distributed as $N(d', 1)$ and correct responses as $N(0, 1)$, where the separation $d' = d'(\theta, s)$ increases with critical literacy θ and decreases with seductiveness s . The probabilities of overriding an erroneous response (a hit) and a correct one (a false alarm) are then

$$h(\theta) = \Phi(d'(\theta, s) - k), \quad f = \Phi(-k),$$

where Φ is the standard normal distribution function. The false-alarm probability depends on k alone. It measures the person's disposition to override, not their ability to discriminate. A logistic parameterization with the same structure is $h(\theta) = f + (1 - f)L(\theta)$, with L the

logistic function of Section 3, so that the lower asymptote of h is f rather than 0.

As $\theta \rightarrow -\infty$, $d'(\theta, s) \rightarrow 0$ and erroneous and correct responses become indistinguishable, so the person overrides both at the same rate: $h(\theta) \rightarrow \Phi(-k) = f$. Substituting into the accounting identity $P(X = 1 | \theta) = c(1 - f) + (1 - c)h(\theta)$ gives the lower asymptote

$$\lim_{\theta \rightarrow -\infty} P(X = 1 | \theta) = c(1 - f) + (1 - c)f.$$

Subtracting c ,

$$c(1 - f) + (1 - c)f - c = f(1 - 2c),$$

which is negative whenever $c > .5$ and $f > 0$. The expression is a gain-loss accounting relative to accepting every response, which scores c . Overriding a fraction f of all responses indiscriminately ruins a fraction f of the correct ones, a loss of cf , and by chance repairs a fraction f of the erroneous ones, a gain of $(1 - c)f$. When the model is right more often than not, the loss dominates. Table A1 gives values.

$c \setminus f$	0	.10	.20	.50	1.00
.30	.300	.340	.380	.500	.700
.50	.500	.500	.500	.500	.500
.70	.700	.660	.620	.500	.300
.90	.900	.820	.740	.500	.100
.96	.960	.868	.776	.500	.040

Table A1: Lower asymptote $c(1 - f) + (1 - c)f$ by model capability c (rows) and false-alarm probability f (columns).

The relaxation does not change the conclusion of Section 4. Under a fixed criterion the false-alarm probability does not depend on θ , so as $c \rightarrow 1$ the success probability approaches the constant $1 - f$ for every person, and the item carries no information about critical literacy. Discrimination can be observed only when erroneous responses are present to be discriminated from correct ones.

AI Use Disclosure

The central idea for this manuscript (captured in the title) came to me as an insight after a long conversation with ChatGPT 5.2 in January 2026. The formalization of this idea into the constrained 3PL and the concept of demonstrating the “verification frontier” through a simulation came from Claude (Opus 4.8). The text of this manuscript was generated iteratively through my orchestration using Claude (Opus 4.8). A revision based on reviewer feedback was supported by Claude (Table 5.1).