Graduating the Benchmark Scale: Lessons from Thermometry

Sean Trott
Department of Psychology
Rutgers University-Newark
sean.trott@rutgers.edu

Oisín Parkinson-Coombs
Department of Psychosocial Science
University of Bergen
oparkinson@ethz.ch

Abstract

Benchmarks for assessing large language model (LLM) capabilities have been criticized for a lack of *construct validity*. Here, we focus on an often overlooked dimension of a benchmark's validity: namely, the functional mapping between a benchmark's numerical score and the underlying quantity the benchmark purports to measure. What licenses the assumption that equivalent intervals on a scale correspond to equivalent differences in the underlying capability? We argue that this question is not merely theoretical: the form of this mapping (e.g., linear vs. logarithmic vs. exponential) could and should influence decisions about deployment and regulatory policy. Drawing on work from the history and philosophy of science, we discuss an analogous problem in the early history of thermometry termed the *problem of nomic measurement*, as well as the epistemic practices that enabled scientists to overcome these challenges. We then ask whether a similar process of *epistemic iteration* can overcome this problem in benchmarking. Despite clear differences between temperature and "capabilities" as constructs, we argue that some modest success could be achievable in the domain of benchmarking—but that this depends crucially on the clear articulation of a researcher's goals and theoretical commitments.

1 Introduction

Benchmarks designed to assess the "capabilities" of large language models (LLMs) ostensibly play an important role in the LLM research and policy ecosystem: in principle, an LLM's performance on a benchmark should influence our inferences about that LLM's underlying abilities, as well as our decisions about whether the system is safe to deploy (METR, 2026). Yet in practice, benchmarks are a major driver of disagreement and debate; these debates often center around their *construct validity*, i.e., whether they actually measure what they are designed to measure (Raji et al., 2021; Saxon et al., 2024; Bean et al., 2025; Wallach et al., 2025; Weidinger et al., 2025; Salaudeen et al., 2025).

Specific validity critiques include skepticism about whether behavior on a benchmark generalizes to the complexity of the real world (Raji et al., 2021; Saxon et al., 2024) and whether the same task "means the same thing" for humans and LLMs (Trott et al., 2023; Hu et al., 2025; Ivanova, 2025; Trott et al., 2026). We broadly agree with these critiques. Here, we focus our attention on a frequently overlooked dimension of benchmark validity: namely, whether a benchmark's numerical scores can be meaningfully interpreted as reflecting degrees of some underlying capability. Put another way: what grounds the mapping between measured performance and the quantity we claim to be measuring?

2 The Problem of Nomic Measurement

Validating a novel instrument faces a problem of circularity, which philosopher and historian of science Hasok Chang has referred to as the *problem of nomic measurement*. Chang (2004, p. 59) writes:

1. We want to measure quantity X.
2. Quantity X is not directly observable, so we infer it from another Y, which is directly observable.
3. For this inference we need a law that expresses X as a function of Y, as follows: $X = f(Y)$.
4. The form of this function f cannot be discovered or tested empirically, because that would involve knowing the values of both Y and X, and X is the unknown variable that we are trying to measure.

Chang (2004) describes this circularity in the context of thermometry: having established that

Proceedings of the Workshop on Evaluating Evaluations (EvalEval), pages 111–115

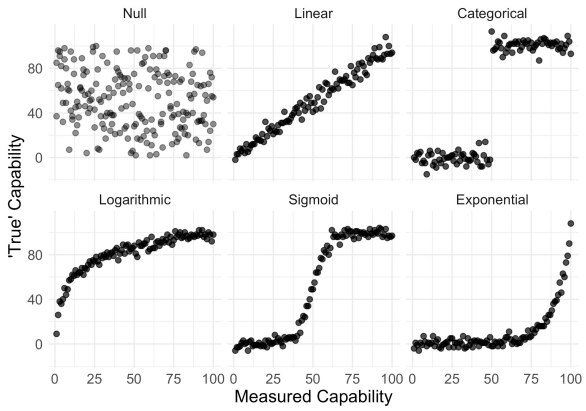

Figure 1: Six hypothetical functional mappings between a measured capability and the (assumed) "true quantity". These scenarios are not intended to be exhaustive.

(say) mercury expands when heated and contracts when cooled, and having identified the "endpoints" of this scale (freezing and boiling), how did scientists *graduate* the intermediate points of the scale without some independent, external criterion? Most options make unjustifiable assumptions. For instance, dividing the scale into equal intervals assumes that mercury expands linearly with temperature. Similarly, the "method of mixtures" approach—in which intermediate points are triangulated by combining various quantities of frozen and boiling water and drawing inferences about the resulting temperature from the ratio of these respective quantities—assumes (incorrectly) that the heat capacity of a liquid is constant with respect to its temperature.

The solution, as discussed in Section 4, ultimately depended on a process of epistemic iteration: rather than seeking a single axiomatic criterion, scientists made progress by iteratively refining fixed points, instruments, and theoretical assumptions in concert. Progress on the problem of nomic measurement in particular came from a shift in emphasis (led by Henri Regnault) towards minimal empirical criteria, such as determining the *comparability* of different instruments (Chang, 2004). First, however, we consider the analogous problem in the construction and design of LLM benchmarks.

3 From Thermometers to LLM Benchmarks

Suppose a task is designed to assess intermediate programming ability in Python.[1] The task consists of 100 questions with objectively correct answers; as such, LLM performance can be described on a scale ranging from 0 (no questions answered correctly) to 100 (all questions answered correctly). We can set aside, for now, the question of whether these items are representative at all of the construct more generally (Yarkoni, 2022; Saxon et al., 2024; Raji et al., 2021). Let us assume, instead, the two criteria cited by Borsboom et al. (2004) for establishing construct validity: (i) the "true" construct exists in some form; and (ii) the true construct bears some causal relation to measured programming ability. Even under these generous assumptions, a fundamental problem remains: what are we licensed to infer about "true" programming ability from measured programming ability?

As Figure 1 depicts, there are a number of possible *functional mappings* between a measurement and the underlying quantity. This relationship could be linear (i.e., unit increases in X correspond to linear increases or decreases in Y), but it could also be (at minimum): categorical (i.e., some threshold in X indexes the presence or absence of Y, but further changes in X do not index further variation in Y); logarithmic (i.e., increases in X lead to "diminishing returns" in Y); sigmoidal (roughly analogous to the categorical relationship); or even exponential (i.e., unit increases in X correspond to compounding gains in Y).

These hypothetical functional forms can be made more intuitive by considering a pair of specific situations:

a. Model M_1 scores 30 and model M_2 scores 50 on the programming benchmark.

b. Model M_1 scores 50 and model M_2 scores 70 on the programming benchmark.

In both situations, the models vary in measured programming ability by 20 points. Should we infer that the difference in "true" programming ability between M_1 and M_2 is equivalent in (a) and (b)? As Figure 1 makes clear, this depends on the form of the relationship between measured and "true" programming ability. If the relationship is linear, the intervals are equivalent in (a) and (b). But if the relationship is logarithmic, the interval in (a) indicates a larger "true" gap than the interval in (b)—and the reverse is true of an exponential relationship. Moreover, if the mapping is categorical,

[1]The arguments below are not specific to programming ability, and could instead be made about a range of capacities, including Theory of Mind (Hu et al., 2025; Ullman, 2023), mathematical reasoning, and more.

all that matters is whether the interval spans the detection threshold.

Benchmarks might inform practical decisions about which models are safe to deploy in which situations, or even which regulatory policies to craft. For example, suppose X is not measured programming ability but some "behavior of concern", such as persuasion ability and propensity (Jones and Bergen, 2026): stakeholders might (justifiably) make very different decisions depending on whether the functional form is linear, logarithmic, or exponential.

This dilemma is arguably even more challenging than the problem faced by scientists working in thermometry. LLM "capabilities" are more abstract—and their objective "reality" more questionable—than temperature; moreover, as we argue below (Section 5), navigating questions of validity depends on a clear articulation of one's research goals (Larroulet Philippi, 2021). However, if there is a solution, insights can nonetheless be drawn from the historical successes of thermometry, even if those successes primarily serve to highlight disanalogies between the situations.

4 Epistemic Iteration: A Way Out?

Clearly, thermometry has achieved remarkable success despite numerous challenges: a reliable thermometer can be purchased at a neighborhood pharmacy, with little thought given to whether this instrument provides reliable, valid measurements—or how its validity was determined. How did scientists overcome the various circularities inherent to validating a new instrument?

Chang (2004) argues that this success can be attributed to a process termed *epistemic iteration* (see also Chang, 2015, 2017). Rather than assuming measurements must be fully justified or "grounded" in axiomatic claims, Chang (2004) suggests that validation proceeds by a series of successive approximations, in which each stage builds on (but is not strictly entailed by) the last. This is broadly consistent with a *coherentist* approach to epistemology, in which individual claims are justified not in isolation but in reference to a broader "web" of mutually supportive observations or beliefs.

With respect to temperature specifically, Chang (2004, p. 47) suggests that thermometry likely followed several distinct stages: first, temperature "measurements" directly reflected bodily sensation (i.e., sensitivity to hot and cold); second, thermoscopes were developed to assess ordinal (but not numerical) changes in temperature, and were validated with respect to bodily sensation; and third, numerical thermometers were devised with the use of fixed points (e.g., the freezing and boiling points of water) and rigorous comparison of different instruments (e.g., Henri Regnault's work assessing the comparability of different air thermometers).

Crucially, these stages did not proceed via strict hierarchical justification. For instance, bodily sensation may have guided the initial validation of thermoscopes, but was not viewed as the "ground truth": indeed, thermoscopes could correct mistaken sensory impressions. Similarly, the use of fixed points was both iterative and contingent: the boiling point of water was sufficiently stable under most circumstances to serve as an initial anchor, but was later shown to depend on a range of external conditions (e.g., barometric pressure), which is why it was eventually displaced by steam temperature as a superior fixed point.

Progress was not linear. Yet several criteria appear to have facilitated the success of epistemic iteration here: first, each stage had a provisional but functional starting point, which was subject to improvement; second, scientists made extensive use of independent instruments for cross-checking and validating measurements; and third, the underlying phenomenon (temperature) was sufficiently clear and well-defined to recognize when progress was being made, however incremental.

5 What Are Benchmarks For?

There are, of course, a number of disanalogies between temperature and LLM capabilities—and between the practices of thermometry and LLM benchmarking (or "model metrology" (Saxon et al., 2024)). Even if one grants the objective existence of LLM "capabilities", they are (clearly) less intimately linked to direct sensory experience than temperature, making the initial stage of epistemic iteration more challenging. Moreover, it is not clear which "fixed points" (provisional or otherwise) might be used to anchor a scale. Human performance (i.e., accuracy or time-to-completion) is one possibility (METR, 2026), but human behavior is variable (likely moreso than the boiling point of water), and may not be directly commensurable with LLM behavior (Trott et al., 2023; Ivanova, 2025); indeed, the fact that construct validity remains a major point of debate within psychome-

tric research on humans should be an indication that the use of human baselines—while obviously advisable—may be more theoretically and empirically problematic than the use of freezing and boiling points in thermometry.

How much do these disanalogies matter? In our view, the extent to which the problem of nomic measurement is actually a problem—and how—depends crucially on one's **research goal**, i.e., the intended use (theoretical or practical) of a benchmark. As Larroulet Philippi (2021) argues, the validity of a measure cannot be easily disentangled from its explanatory context and its ultimate purpose: for example, concerns about how the thermometer is "graduated" presumably depend on the level of precision needed, e.g., whether one needs to differentiate exact degrees of temperature or simply rank substances on an ordinal scale. To take a more relevant example: an evaluation of model capabilities (e.g., reasoning) could in principle be designed to discriminate finely among individuals at the upper end of the distribution, or alternatively to detect meaningful differences in the middle range. Even holding the functional mapping constant, a benchmark that discriminates well among frontier models may be uninformative about differences among weaker systems, and vice versa. Evaluating a benchmark's validity thus requires knowing what it is being asked to do.

One possible goal is **ranking models** (e.g., in terms of their reasoning capacity, or in terms of their "degree of alignment"). Here, the problem of "graduating" a benchmark does not need to be solved. A benchmark merely needs to approximate the true rank ordering of models, which is accomplished by all functional mappings in Figure 1 except for the "Null" scenario. Of course, this goal assumes that such an ordered ranking is in principle possible, and some scenarios (e.g., the "Categorical" function) may lead to misleading inferences about relative differences between models on the basis of an ordered ranking.

Another possible goal is behavioral **detection**, e.g., of a capability or dangerous behavior. As with an ordinal scale, a detector does not need to provide a fully graduated scale. Instead, it needs to reliably discriminate between systems that do and do not exhibit the behavior in question, which requires the identification of a meaningful threshold. This corresponds to the categorical (and sigmoidal) mappings in Figure 1.

The third (and most ambitious) possibility is assessing **degrees** of a capability or construct, analogous to an interval/ratio scale. This is the goal for which the problem of nomic measurement (Chang, 2004) is most acute: scores must reflect not only rank order, but meaningful differences in the underlying quantity. Achieving such a goal could be intractable either because of questionable theoretical assumptions (e.g., it may be inappropriate to characterize reasoning ability as a quantity) or because of insufficient external validation criteria (i.e., to triangulate the quantity in question). Some work, however, has attempted to address these challenges empirically, e.g., identifying the "laws" relating model properties to benchmark performance (Kaplan et al., 2020), or grounding performance with other criteria (Schaeffer et al., 2025).

These disparate goals, in turn, may serve a variety of **inferential functions**. Researchers may be interested in drawing theoretical conclusions, e.g., about the conditions under which certain cognitive behaviors emerge (Trott et al., 2026; Kouwenhoven et al., 2026); alternatively, they might be seeking guidance in crafting policy or forecasting future changes in LLM capabilities (METR, 2026). The nature of these broader inferential aims, in turn, constrains which of the above goals is most appropriate. Even within a given inferential function, one's level of ambition may vary, e.g., predicting whether models will eventually pass some critical capability threshold (detection) vs. predicting the rate of improvement of some capability (degree). These distinctions affect the stringency of the measurement problem at play.

Our argument here is not that researchers should focus on one goal or another. Rather, our view is that researchers should define and clearly articulate their goal, as well as the theoretical commitments undergirding the selection of a particular benchmark in serving that goal (Alexandrova and Haybron, 2016). This articulation would not necessarily solve the problem of nomic measurement (see Section 2), but it would clarify which version of the problem needs to be resolved (see Section 3). In some cases, the inferential demands of a research question may actually be more modest than the implicit assumptions of current benchmarking practices suggest; in other cases, researchers might realize they should adopt a different—perhaps more epistemically cautious—vocabulary for describing measured differences on a scale (e.g., ordinal vs. numerical).

References

Anna Alexandrova and Daniel M. Haybron. 2016. Is Construct Validation Valid? *Philosophy of Science*, 83(5):1098–1109.

Andrew M Bean, Ryan Othniel Kearns, Angelika Romanou, Franziska Sofia Hafner, Harry Mayne, Jan Batzner, Negar Foroutan, Chris Schmitz, Karolina Korgul, Hunar Batra, and 1 others. 2025. Measuring what matters: Construct validity in large language model benchmarks. *arXiv preprint arXiv:2511.04703*.

Denny Borsboom, Gideon J. Mellenbergh, and Jaap Van Heerden. 2004. The Concept of Validity. *Psychological Review*, 111(4):1061–1071.

Hasok Chang. 2004. *Inventing Temperature: Measurement and Scientific Progress*. Oxford University Press. Google-Books-ID: yVOuV8qJkxMC.

Hasok Chang. 2015. The rising of chemical natural kinds through epistemic iteration. In *Natural kinds and classification in scientific practice*, pages 33–46. Routledge.

Hasok Chang. 2017. Epistemic iteration and natural kinds: Realism and pluralism in taxonomy. *Philosophical issues in psychiatry IV: Psychiatric nosology*, pages 229–245.

Jennifer Hu, Felix Sosa, and Tomer Ullman. 2025. Re-evaluating Theory of Mind evaluation in large language models. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 380(1932):20230499.

Anna A. Ivanova. 2025. How to evaluate the cognitive abilities of LLMs. *Nature Human Behaviour*, 9(2):230–233.

Cameron Jones and Benjamin Bergen. 2026. Lies, damned lies, and language statistics: a comprehensive review of risks from manipulation, persuasion, and deception with large language models. *Artificial Intelligence Review*.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Tom Kouwenhoven, Michiel van der Meer, and Max van Duijn. 2026. Traces of social competence in large language models. *arXiv preprint arXiv:2603.04161*.

Cristian Larroulet Philippi. 2021. Valid for What? On the Very Idea of Unconditional Validity. *Philosophy of the Social Sciences*, 51(2):151–175. Publisher: SAGE Publications Inc.

METR. 2026. Task-completion time horizons of frontier ai models. `https://metr.org/time-horizons/`.

Inioluwa Deborah Raji, Emily M Bender, Amandalynne Paullada, Emily Denton, and Alex Hanna. 2021. Ai and the everything in the whole wide world benchmark. *arXiv preprint arXiv:2111.15366*.

Olawale Salaudeen, Anka Reuel, Ahmed Ahmed, Suhana Bedi, Zachary Robertson, Sudharsan Sundar, Ben Domingue, Angelina Wang, and Sanmi Koyejo. 2025. Measurement to meaning: A validity-centered framework for ai evaluation. *arXiv preprint arXiv:2505.10573*.

Michael Saxon, Ari Holtzman, Peter West, William Yang Wang, and Naomi Saphra. 2024. Benchmarks as Microscopes: A Call for Model Metrology. *arXiv preprint*. ArXiv:2407.16711 [cs].

Rylan Schaeffer, Punit Singh Koura, Binh Tang, Ranjan Subramanian, Aaditya K Singh, Todor Mihaylov, Prajjwal Bhargava, Lovish Madaan, Niladri S Chatterji, Vedanuj Goswami, and 1 others. 2025. Correlating and predicting human evaluations of language models from natural language processing benchmarks. *arXiv preprint arXiv:2502.18339*.

Sean Trott, Cameron Jones, Tyler Chang, James Michaelov, and Benjamin Bergen. 2023. Do Large Language Models Know What Humans Know? *Cognitive Science*, 47(7):e13309.

Sean Trott, Samuel Taylor, Cameron Jones, James A. Michaelov, and Pamela D. Rivière. 2026. Language Statistics and False Belief Reasoning: Evidence from 41 Open-Weight LMs. *arXiv preprint*. ArXiv:2602.16085 [cs].

Tomer Ullman. 2023. Large Language Models Fail on Trivial Alterations to Theory-of-Mind Tasks. *arXiv preprint*. ArXiv:2302.08399 [cs].

Hanna Wallach, Meera Desai, A. Feder Cooper, Angelina Wang, Chad Atalla, Solon Barocas, Su Lin Blodgett, Alexandra Chouldechova, Emily Corvi, P. Alex Dow, Jean Garcia-Gathright, Alexandra Olteanu, Nicholas J Pangakis, Stefanie Reed, Emily Sheng, Dan Vann, Jennifer Wortman Vaughan, Matthew Vogel, Hannah Washington, and Abigail Z. Jacobs. 2025. Position: Evaluating generative AI systems is a social science measurement challenge. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 82232–82251. PMLR.

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

Tal Yarkoni. 2022. The generalizability crisis. *Behavioral and Brain Sciences*, 45:e1.