

Machine psychometrics for artificial agents

Konstantina Sokratous^{a,†}, Joachim Vandekerckhove^b, and Clinton P. Davis-Stober^a

Artificial intelligence systems are increasingly evaluated as behavioral agents, yet many evaluation practices still treat model outputs as isolated responses to be scored for accuracy, safety, bias, or preference alignment. We argue that this response-level paradigm is insufficient for systems whose behavior is generated from probability distributions and changes across contexts, versions, and interventions. Human psychometrics and cognitive modeling offer a useful foundation because they were developed to infer latent traits, capacities, preferences, and decision processes from indirect behavioral data. However, artificial agents create a different measurement regime. Their next-token and response-level probability distributions can often be inspected directly, allowing uncertainty, calibration, bias, and response instability to be modeled as latent properties rather than inferred only from overt responses. Artificial agents can also be remeasured and intervened on under controlled conditions, enabling parameter-level assessment of stability, drift, improvement, degradation, and causal sensitivity across prompts, model updates, fine-tuning, retrieval changes, and ablations. We propose *machine psychometrics* as a distribution-aware measurement framework for artificial agents. In this framework, computational cognitive models serve as interpretable interfaces between probability distributions and latent parameters, allowing model behavior to be compared, validated, and monitored without assuming that human psychological constructs transfer unchanged to machines. This approach shifts AI evaluation from scoring what a model says to measuring the probabilistic structure that generates its behavior.

Machine Psychometrics, Generative AI, Cognitive Modeling

Artificial intelligence systems are increasingly embedded in settings that require more than prediction, with artificial agents acting as tutors, advisors, research assistants, evaluators, software engineers, and decision-support tools across various domains (Albashrawi, 2025; Haque & Siddique, 2026; Lai et al., 2024; Raza et al., 2025; Chowa et al., 2026; Gu et al., 2026). We use “artificial agents” or simply “agents” to refer to large language models and related generative AI systems whose behavior can be probed through prompts that yield free-form or structured responses. As these systems become more capable and more widely deployed, methods are needed to evaluate whether they exhibit stable, interpretable, and validly measurable behavioral properties. Such evaluations are ideally reliable across contexts, calibrated under uncertainty, sensitive to relevant distinctions, invariant under superficial prompt changes, and responsive to interventions in ways that support theoretical interpretation.

Current evaluation practices have made substantial progress toward this goal, but they remain largely organized around output-level performance. Benchmarks such as the Massive Multitask Language Understanding data set and related task batteries measure accuracy across broad domains (Hendrycks et al., 2021), holistic evaluation frameworks such as the Holistic Evaluation of Language Models expand the scope of assessment to include accuracy, calibration, robustness, fairness, toxicity, efficiency, and other dimensions (Liang et al., 2023; Bommasani et al., 2023) and behavioral testing approaches such as CheckList expose systematic model failures that aggregate accuracy scores can obscure (Ribeiro et al., 2020). These efforts have been indispensable for making model evaluation more transparent and multidimensional. However, they still typically treat the observable response as the primary desired datum. An agent is scored on what it says, se-

lects, rejects, or generates. The latent structure that gives rise to that response and the distribution of alternatives considered, the uncertainty encoded in that distribution, and the decision process by which one response is selected over another, usually all remain outside the measurement model.

Psychological science has long faced an analogous problem as human traits, capacities, preferences, and cognitive processes are not directly observable. They must be inferred from sparse behavioral traces often including choices, ratings, response times, verbal reports, task performance, and patterns of error. A formal measurement science—psychometrics—was developed for making such inferences under explicit assumptions about reliability, validity, dimensionality, and construct representation. Along similar lines, cognitive modelers developed a complementary strategy where instead of treating the behavior as a surface outcome, it is decomposed into latent processes and interpretable parameters.

The advantage of these methods is that they were built under the strict measurement constraints of the context within which they operated. With human participants, scientists cannot directly observe probability distributions over possible responses before a choice is made. Confidence and ability must be inferred from response latency, choice consistency, self-report, or post hoc judgment. Repeated testing is limited by fatigue, learning, memory, motivation, practice effects, and attrition. Experimental intervention is constrained by both ethics and biology, as one cannot freely ablate a cognitive subsystem, duplicate a participant, or roll back a person to a prior state.

Human data are therefore complex not simply because behavior is variable, but because the measurement situation is indirect, noisy, sparse, and tightly constrained. The great sophistication of modern psychometric measurement methods is in large part in response to these limitations (e.g., Cronbach et al., 1963).

Artificial agents preserve the need for psychometric evaluation (Wang et al., 2026), but they change the measurement situation. Like human respondents, artificial agents produce behavior that is context-sensitive, construct-laden, and difficult to interpret from

^aUniversity of Missouri; ^bUniversity of California, Irvine

All authors contributed to the final draft.

[†]Corresponding author. E-mail: ksokratous@missouri.edu

The authors declare no conflicts of interest.

isolated responses alone. However, unlike human respondents, artificial agents can expose aspects of the generative process that are usually hidden in human testing. For instance, autoregressive language models define probability distributions over possible next tokens, from which sampled outputs are generated (Vaswani et al., 2017; Brown et al., 2020). When these probabilities are accessible through open model weights, log-probability interfaces, or controlled decoding procedures, they provide a qualitatively different object of measurement that allows evaluators to treat responses as part of a larger distribution directly.

Consider a clinical decision-support model asked to evaluate whether a symptom profile indicates the presence of a disease. Two models may both return the same final judgment (e.g., “disease likely”), which at the response level will appear identical, but one model may internally assign overwhelming probability to the disease-present judgment, whereas the other may distribute probability more evenly across disease-present, disease-absent, and indeterminate responses. A conventional benchmark that scores only the final recommendation treats these cases as equivalent while a distribution-aware psychometric approach does not.

This observation motivates the central claim of this paper: that there is a need for a *machine psychometrics* that goes beyond copying human tests for artificial respondents, but rather adapts psychometric and cognitive-modeling principles to the affordances of artificial agents. By *affordances*, we mean properties of a system that make particular forms of measurement possible (Gibson, 1979). Artificial agents generally possess two properties that distinguish them from human respondents: (1) They afford *access*, at least in principle and often in practice, to probability distributions over possible responses. The basic unit of psychometric analysis changes from treating a response as a fixed answer to be scored to evaluators modeling the distribution from which that response was sampled. (2) They afford *control*: artificial agents can be remeasured and intervened on under controlled conditions. They can be probed repeatedly across seeds, prompts, decoding settings, model versions, or deployment contexts, and they can be deliberately modified through prompting, retrieval changes, fine-tuning, data perturbation, or ablation. This makes it possible to track whether estimated latent properties are stable, drifting, improving, or changing in theoretically expected ways.

Because these affordances are currently largely unutilized in psychometric evaluation, we argue for a theory of machine psychometrics that (a) is organized around the distinctive data structures artificial agents provide as a new measurement object, and that (b) capitalizes on controlled re-measurement and intervention to permit new validation designs. Computational cognitive models can then provide a bridge between the two as they translate response distributions into interpretable parameters that can be compared across models, contexts, and experimental manipulations.

Affordance 1: Access

The central measurement opportunity created by artificial agents is that their overt responses are generated from probability distributions that can in principle be inspected directly.¹ Autoregressive language models assign probabilities to possible continuations of a sequence by estimating a conditional distribution over the next token given the preceding context (Bengio et al., 2003; Vaswani et al., 2017; Brown et al., 2020). A generated answer is therefore the result of a decoding process applied to a distribution over possible next tokens and by extension possible continuations (Holtzman

et al., 2020). Standard evaluation usually observes only the final continuation involving the answer selected, the refusal produced or the recommendation stated. However, distribution-aware methods can treat the probability distribution underlying that continuation as measurement data. This matters because an observed response collapses useful information that is available before an answer is selected or sampled. A system may respond “disease likely” but the probability structure behind that answer might indicate strong commitment, near indifference between disease-present and disease-absent options or substantial probability assigned to an indeterminate response. These can suggest and reflect different forms of uncertainty, decisiveness, and response organization. In human testing, such properties must be inferred indirectly from repeated choices, response times, confidence reports, and/or model assumptions. In artificial agents, however, probability outputs allow these properties to be modeled directly, provided that the relevant response space has been defined.

The first methodological challenge is therefore that response probabilities are natively defined over tokens, whereas psychometric constructs are defined over substantively meaningful response categories. A vocabulary distribution is not, by itself, a measurement model. The same decision or judgment may be expressed through many token sequences such as “positive”, “disease present”, “likely malignant”, or simply “yes.” Conversely, similar surface forms may imply different judgments in different contexts. For this reason, machine psychometrics requires an explicit mapping from token space to construct space where the relevant distribution is not necessarily the raw next-token distribution, but the induced distribution over theoretically specified alternatives.

This mapping step is part of construct definition. For example, in a diagnostic classification task, the response space may be “disease present,” “disease absent,” and “indeterminate.” In a triage task, it may distinguish “immediate intervention” from “additional testing,” and so on. In each case, probability mass must be aggregated over tokens or completions that express the same substantive alternative. The resulting distribution is interpretable only because the response categories have been specified in relation to the construct being measured.

This point mirrors a central principle of human psychometrics where validity goes beyond the statistical properties of responses, addressing the interpretation of scores as well (Cronbach & Meehl, 1956; Messick, 1994; Borsboom et al., 2004). The same principle applies to probability outputs: A distribution can support psychometric inference only if the categories over which it is defined have a defensible relation to the target construct. If the construct is diagnostic uncertainty, the response space must preserve degrees or forms of diagnostic commitment, if it is risk sensitivity, the alternatives must differ meaningfully in risk, so on and so forth. Access to probabilities is therefore a necessary step, but measurement only begins when those probabilities are linked to constructs.

Computational cognitive models as probability-to-construct interfaces. Probability outputs only become psychometrically useful when they are linked to models that translate distributional patterns into interpretable parameters. The interpretation of a parameter as a meaningful construct is in part licensed by its role in a proposed data-generating process (Embretson, 2010; Friend, 2019). For example, in a calibration model, an intercept can represent systematic overprediction or underprediction, while a slope can represent whether probability judgments are too extreme or too compressed. In a signal-detection model, separate parameters can represent sensitivity to evidence and response bias. In an

¹ Although possibly only by the creators of the system.

item-response model, parameters can represent model ability and item difficulty. These quantities are interpretable because they connect observable probability patterns to generatively specified behavioral properties.

In human research, computational cognitive models formalize assumptions about how latent processes generate observable behavior, allowing researchers to estimate quantities such as evidence sensitivity, response bias, decision threshold, learning rate, utility, or subjective value (Ratcliff, 1978; Daw, 2011; Wilson & Collins, 2019). In machine psychometrics, analogous models can be fit to sampled responses and, when available, to the probability distributions that artificial systems produce.

The value of cognitive models is interpretive as they provide a vocabulary for describing artificial-agent behavior without relying only on surface benchmark scores or importing human trait labels uncritically. A clinical decision-maker can be described as overconfident, conservative, sensitive to diagnostic evidence, biased toward positive classifications, or insensitive to ambiguous cases. These descriptions are not claims that the model has human-like mental states – rather, they provide parameterized summaries of behavior, often formal descriptions of how the system distributes probability over a defined response space under a specified measurement model and design. Their usefulness depends on whether those parameters are stable, predictive, and responsive to theoretically meaningful changes. For more reading into the field of computational cognitive modeling readers can consult Sun (2008).

It is thus worth clarifying the gap in current psychometric approaches to AI evaluation. One line of work administers human psychological inventories to artificial agents, asking whether models exhibit personality traits, values, or other human-like profiles (Pellert et al., 2024; Miotto et al., 2022; Li et al., 2024; Serapio-García et al., 2025), while another line of research uses psychometric tools, especially item response theory, to analyze benchmarks, estimate item difficulty, and improve model comparisons. Both are useful, but neither fully exploits the distinctive probability outputs of generative systems. This is where we believe we can leverage distribution-aware approaches.

A basic simulated medical classification and calibration example. To make the logic concrete, consider a simple simulated medical classification task. The purpose of the example is not to evaluate a particular deployed model, but to show how distribution-aware measurement can reveal underlying differences between agents that may be obscured by response-level scoring. We simulate three diagnostic systems that evaluate the same set of cases and assign a probability that a disease is present. Each case is associated with a latent evidence value, where negative values indicate evidence against disease and positive values indicate evidence for disease. The systems differ only in how they transform that evidence into response probabilities.

The first system is approximately calibrated, meaning that as diagnostic evidence increases, its predicted probability of disease increases in a way that tracks the data-generating process. The second system is overconfident, meaning that it uses a similar categorical decision boundary but maps evidence into more extreme probabilities, assigning values closer to 0 or 1. The third system is disease-biased and shifts probability mass toward disease-present judgments even when the evidence is ambiguous. Thus, the three systems can be understood as having different latent probability-use profiles (calibrated, overconfident, and disease-biased).

Figure 1A illustrates the probability functions for the three systems. All systems receive the same diagnostic evidence, but they

transform that evidence differently. These differences are visible before any final categorical response is sampled; they are properties of the response-generating distribution itself.

Figure 1B shows the corresponding calibration curves, where perfect calibration would place predicted probabilities on the diagonal: Among cases assigned a probability of approximately .80, disease should be present approximately 80% of the time. As can be seen, the calibrated system (green circles) remains close to this reference line, while the overconfident one (orange squares) deviates by assigning probabilities that are too extreme relative to the observed outcome. The disease-biased system (purple triangles) deviates by systematically overpredicting disease presence.

A simple cognitive model can summarize these differences with interpretable parameters (Turner et al., 2014). For illustration, we use a log-odds calibration model,

$$\text{logit}(\gamma) = \alpha \times \text{logit}(p) + \beta,$$

where p is the predicted probability assigned by the agent and γ is the observed outcome frequency among cases receiving similar predicted probabilities. The intercept β captures systematic overprediction or underprediction, whereas the slope α captures the extremity of probability use. An ideal system has $\beta = 0$ and $\alpha = 1$. This model is deliberately simple – its purpose is to show how probability outputs can be translated into latent psychometric parameters, not to exhaust the set of possible computational cognitive models.

Figure 1C shows estimated calibration parameters for the three systems. The calibrated system (green circle) lies close to the ideal point (star shape) while the overconfident system (orange square) and the disease-biased system (purple triangle) lie far away. In response-level evaluation, these systems might appear similar if they often produce the same categorical classification but their probability distributions are indicative of different latent behavioral organizations and a distribution-aware method allows us to distinguish these clearly.

Affordance 2: Control

The first affordance makes it possible to estimate latent parameters from probability distributions. The second affordance concerns what becomes possible once those parameters can be estimated repeatedly under controlled change. In human psychometrics, repeated measurement is central to reliability, but it is also difficult to interpret because with human respondents, measurement is often *reactive*: administering a test can alter the respondent's internal state, memory, motivation, strategy, or relation to the construct being measured (Cronbach et al., 1963; Campbell, 1957).

Artificial agents can sidestep this measurement problem because the same items, prompts, response-space mappings, and model specifications can be reapplied across seeds, decoding settings, deployment contexts, model versions, or experimental interventions. The instrument can be held fixed while the agent or context changes (e.g., through a model update or fine-tuning). Work on model drift and version comparison has shown that deployed AI behavior can change substantially across updates, even when the model nominally remains the same (Chen et al., 2024; Margatina et al., 2023). By holding the measurement framework fixed we can examine how the estimated parameters move and whether a system's behavior is stable, drifting, or changing in theoretically expected ways.

To make the second affordance concrete, consider an extension of the diagnostic example from the Access affordance in which

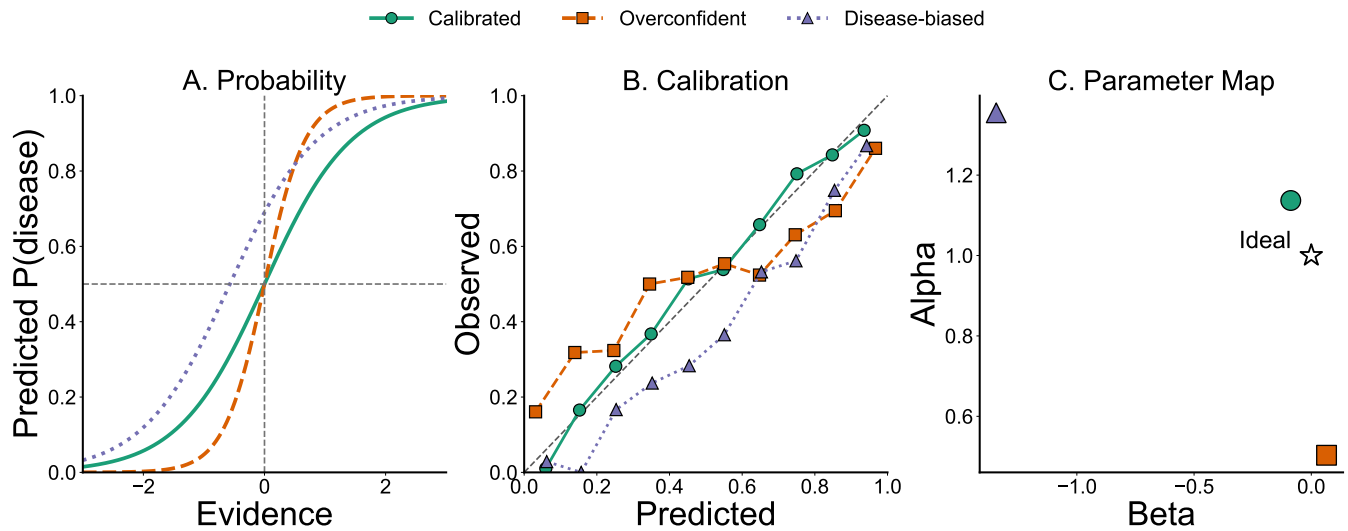


Fig. 1. Distribution-aware evaluation distinguishes latent probability-use patterns that response-level scoring can miss. We simulated three diagnostic classification systems evaluating the same medical cases. **A.** The calibrated system maps diagnostic evidence onto disease probabilities that track the data-generating process. The overconfident system uses a steeper function, assigning more extreme probabilities. The disease-biased system shifts probability mass toward disease-present judgments. **B.** Calibration curves compare predicted probabilities with observed disease frequencies. The diagonal represents perfect calibration. **C.** A simple log-odds calibration model summarizes probability use with two interpretable parameters: slope α , reflecting probability extremity, and intercept β , reflecting systematic overprediction or underprediction. The calibrated system lies near the ideal point ($\beta = 0, \alpha = 1$), whereas the overconfident and disease-biased systems show distinct latent failures in probability use.

the same simulated cases are re-administered across versions of the agent. The first version is overconfident, the second receives a calibration update, and the third receives a later update that introduces a bias toward disease-present judgments. Because the items, response-space mapping, and calibration model are held fixed, changes can be interpreted as movement through a common parameter space rather than as incomparable benchmark scores. Figure 2A shows that the calibration update reduces the steepness of the pre-update probability function, making predicted probabilities less extreme for the same diagnostic evidence. Figure 2B shows the corresponding improvement in calibration: after the update, predicted probabilities more closely track observed disease frequencies. Figure 2C summarizes these changes as movement in calibration-parameter space: The update moves the model from an overconfident region toward the ideal point ($\beta = 0, \alpha = 1$), whereas the later biased update shifts the model away from this point primarily along the intercept dimension. The same instrument distinguishes improvement in one latent property from deterioration in another.

The same logic can then be used to support active intervention. Evaluators can deliberately modify prompts, remove retrieval modules, alter system instructions, compare fine-tuned variants, perturb training data, or ablate architectural components. These interventions can then be evaluated by asking whether the estimated parameters change in theoretically coherent ways. This approach is consistent with the broader causal logic of intervention-based explanation (Pearl, 2009) and with recent work in mechanistic interpretability, where causal mediation, activation patching, causal tracing, and model editing are used to test how changes to internal components affect agent behavior (Vig et al., 2020; Meng et al., 2022; Geiger et al., 2025).

Controlled re-measurement also strengthens construct validation as it can become more experimental, with artificial agents providing the privilege to be directly manipulated and ablated. Parametric models can thus allow evaluators to assess not only whether

an intervention improves or degrades performance, but whether it changes the relevant latent parameter in a theoretically expected direction. Controlled re-measurement and intervention turn machine psychometrics into a dynamic measurement framework, letting us move beyond simple scoring of a model to monitoring a system's latent behavioral organization and how it changes across versions, contexts, and manipulations.

Conclusion

As artificial agents become embedded in complex and critical domains such as medical advising and autonomous research, surface-level evaluation of isolated responses presents clear limitations. Evaluators increasingly need to determine whether a system's behavior is stable, calibrated, and theoretically coherent. Psychometrics and cognitive modeling offer a rigorous foundation for this task by providing established mathematical tools designed to infer latent structure from observable behavior (Epping et al., 2026).

Artificial agents provide an unusual measurement regime. By affording direct access to the probability distributions underlying their outputs and allowing for controlled, repeatable interventions, they enable a highly precise machine psychometrics. This proposed framework unifies behavioral evaluation, mechanistic interpretability, and construct validation, and allows us to approach AI evaluation as a formal measurement problem. Shifting our measurement focus from isolated outputs to the probabilistic structure that generates behavior can provide the robust, interpretable, and dynamic evaluations needed for new generations of artificial agents.

Acknowledgements

This work was supported by a SEED grant to C.D.S and K.S from the College of Arts & Science, at the University of Missouri (Big Idea Initiative).

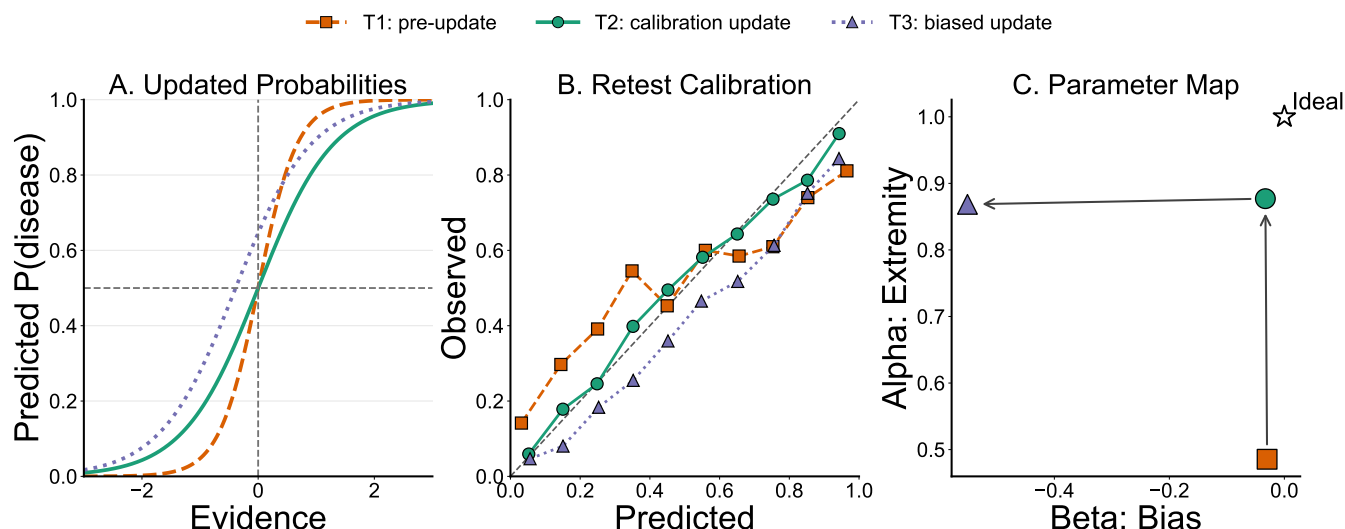


Fig. 2. Controlled remeasurement reveals parameter-level change across model updates. The same simulated diagnostic cases are administered to three versions of a model. **A.** The pre-update model maps evidence into overly extreme probabilities. A calibration update moves the probability function closer to the data-generating process, whereas a later biased update shifts probability mass toward disease-present judgments. **B.** Repeated calibration curves show how the relation between predicted probability and observed disease frequency changes across versions. **C.** A log-odds calibration model summarizes these changes as movement through parameter space. The calibration update moves the model toward the ideal point ($\beta = 0$, $\alpha = 1$), whereas the biased update moves it away along the intercept dimension.

References

- Albashrawi, M. (2025). Generative ai for decision-making: A multidisciplinary perspective. *Journal of Innovation & Knowledge*, 10(4), 100751. Retrieved from <https://www.sciencedirect.com/science/article/pii/S2444569X25000964> doi: <https://doi.org/10.1016/j.jik.2025.100751>
- Bengio, Y., Ducharme, R., Vincent, P., & Janvin, C. (2003, March). A neural probabilistic language model. *J. Mach. Learn. Res.*, 3(null), 1137–1155.
- Bommasani, R., Liang, P., & Lee, T. (2023). Holistic evaluation of language models. *Annals of the New York Academy of Sciences*, 1525(1), 140–146. Retrieved from <https://nyaspubs.onlinelibrary.wiley.com/doi/abs/10.1111/nyas.15007> doi: <https://doi.org/10.1111/nyas.15007>
- Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review*, 111(4), 1061–1071. Retrieved from <https://doi.org/10.1037/0033-295X.111.4.1061> doi: [10.1037/0033-295X.111.4.1061](https://doi.org/10.1037/0033-295X.111.4.1061)
- Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... Amodi, D. (2020). Language models are few-shot learners. In *Proceedings of the 34th international conference on neural information processing systems*. Red Hook, NY, USA: Curran Associates Inc.
- Campbell, D. T. (1957). Factors relevant to the validity of experiments in social settings. *Psychological Bulletin*, 54(4), 297–312. doi: [10.1037/h0040950](https://doi.org/10.1037/h0040950)
- Chen, L., Zaharia, M., & Zou, J. (2024). How is chatgpt's behavior changing over time? *Harvard Data Science Review*, 6(2).
- Chowa, S. S., Alvi, R., Rahman, S. S., Rahman, M. A., Raiaan, M. A. K., Islam, M. R., ... Azam, S. (2026, Jan 06). From language to action: a review of large language models as autonomous agents and tool users. *Artificial Intelligence Review*, 59(2), 71. Retrieved from <https://doi.org/10.1007/s10462-025-11471-9> doi: [10.1007/s10462-025-11471-9](https://doi.org/10.1007/s10462-025-11471-9)
- Cronbach, L. J., & Meehl, P. E. (1956). Construct validity in psychological tests. In H. Feigl & M. Scriven (Eds.), *Minnesota studies in the philosophy of science* (pp. 1–174). , Vol.
- Cronbach, L. J., Rajaratnam, N., & Gleser, G. C. (1963, November). Theory of generalizability: A liberalization of reliability theory. *British Journal of Statistical Psychology*, 16(2), 137–163. Retrieved from <http://dx.doi.org/10.1111/j.2044-8317.1963.tb00206.x> doi: [10.1111/j.2044-8317.1963.tb00206.x](https://doi.org/10.1111/j.2044-8317.1963.tb00206.x)
- Daw, N. D. (2011, 03). Trial-by-trial data analysis using computational models: (tutorial review). In M. R. Delgado, E. A. Phelps, & T. W. Robbins (Eds.), *Decision making, affect, and learning: Attention and performance xxiii*. Oxford University Press. doi: [10.1093/acprof:oso/9780199600434.003.0001](https://doi.org/10.1093/acprof:oso/9780199600434.003.0001)
- Embretson, S. E. (Ed.). (2010). *Measuring psychological constructs: Advances in model-based approaches*. Washington, DC: American Psychological Association. Retrieved from <https://doi.org/10.1037/12074-000> doi: [10.1037/12074-000](https://doi.org/10.1037/12074-000)
- Epping, G. P., Caplin, A., Duhaime, E., Holmes, W. R., Martin, D., & Trueblood, J. S. (2026). Harnessing human uncertainty to train more accurate and aligned ai systems. *Decision Analysis*, 0(0), null. Retrieved from <https://doi.org/10.1287/deca.2025.0395> doi: [10.1287/deca.2025.0395](https://doi.org/10.1287/deca.2025.0395)
- Friend, S. (2019, December). The fictional character of scientific models. In *The scientific imagination* (p. 102–127). Oxford University Press. Retrieved from <http://dx.doi.org/10.1093/oso/9780190212308.003.0005> doi: [10.1093/oso/9780190212308.003.0005](https://doi.org/10.1093/oso/9780190212308.003.0005)
- Geiger, A., Ibeling, D., Zur, A., Chaudhary, M., Chauhan, S., Huang, J., ... Icard, T. (2025). Causal abstraction: A theoretical foundation for mechanistic interpretability. *Journal of Machine Learning Research*, 26(83), 1–64. Retrieved from <http://jmlr.org/papers/v26/23-0058.html>
- Gibson, J. J. (1979). *The ecological approach to visual perception*.

- tion. Boston, MA: Houghton Mifflin. Retrieved from <https://doi.org/10.4324/9781315740218> (Classic Edition (2014), Psychology Press. DOI for reprint: 10.4324/9781315740218) doi: 10.4324/9781315740218
- Gu, J., Jiang, X., Shi, Z., Tan, H., Zhai, X., Xu, C., ... Guo, J. (2026). A survey on llm-as-a-judge. *The Innovation*, 101253. Retrieved from <https://www.sciencedirect.com/science/article/pii/S2666675825004564> doi: <https://doi.org/10.1016/j.xinn.2025.101253>
- Haque, M. A., & Siddique, H. R. (2026). Generative artificial intelligence and large language models in smart healthcare applications: Current status and future perspectives. *Computational Biology and Chemistry*, 120, 108611. Retrieved from <https://www.sciencedirect.com/science/article/pii/S1476927125002725> doi: <https://doi.org/10.1016/j.compbiolchem.2025.108611>
- Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., & Steinhardt, J. (2021). Measuring massive multitask language understanding. In *International conference on learning representations*.
- Holtzman, A., Buys, J., Du, L., Forbes, M., & Choi, Y. (2020). The curious case of neural text degeneration. In *International conference on learning representations*. Retrieved from <https://openreview.net/forum?id=rygGQyrFvH>
- Lai, J., Gan, W., Wu, J., Qi, Z., & Yu, P. S. (2024). Large language models in law: A survey. *AI Open*, 5, 181-196. Retrieved from <https://www.sciencedirect.com/science/article/pii/S2666651024000172> doi: <https://doi.org/10.1016/j.aiopen.2024.09.002>
- Li, X., Li, Y., Qiu, L., Joty, S., & Bing, L. (2024, November). Evaluating psychological safety of large language models. In Y. Al-Onaizan, M. Bansal, & Y.-N. Chen (Eds.), *Proceedings of the 2024 conference on empirical methods in natural language processing* (pp. 1826-1843). Miami, Florida, USA: Association for Computational Linguistics. Retrieved from <https://aclanthology.org/2024.emnlp-main.108/> doi: 10.18653/v1/2024.emnlp-main.108
- Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., ... others (2023). Holistic evaluation of language models. *Annals of the New York Academy of Sciences*, 1525(1), 140-146. Retrieved from <https://doi.org/10.1111/nyas.15007> (Preprint: arXiv:2211.09110 (2022)) doi: 10.1111/nyas.15007
- Margatina, K., Wang, S., Vyas, Y., John, N., Benajiba, Y., & Ballesteros, M. (2023, 01). Dynamic benchmarking of masked language models on temporal concept drift with multiple views. In *Proceedings of the 17th conference of the european chapter of the association for computational linguistics* (p. 2881-2898). doi: 10.18653/v1/2023.eacl-main.211
- Meng, K., Bau, D., Andonian, A., & Belinkov, Y. (2022). Locating and editing factual associations in gpt. In *Proceedings of the 36th international conference on neural information processing systems*. Red Hook, NY, USA: Curran Associates Inc.
- Messick, S. (1994). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *ETS Research Report Series*, 1994(2), i-28. Retrieved from <https://onlinelibrary.wiley.com/doi/abs/10.1002/j.2333-8504.1994.tb01618.x> doi: <https://doi.org/10.1002/j.2333-8504.1994.tb01618.x>
- Miotto, M., Rossberg, N., & Kleinberg, B. (2022, November). Who is GPT-3? an exploration of personality, values and demographics. In D. Bamman, D. Hovy, D. Jurgens, K. Keith, B. O'Connor, & S. Volkova (Eds.), *Proceedings of the fifth workshop on natural language processing and computational social science (nlpcss)* (pp. 218-227). Abu Dhabi, UAE: Association for Computational Linguistics. doi: 10.18653/v1/2022.nlpccs-1.24
- Pearl, J. (2009). *Causality: Models, reasoning, and inference*. Cambridge University Press. Retrieved from <http://dx.doi.org/10.1017/cbo9780511803161> doi: 10.1017/cbo9780511803161
- 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*, 19(5), 808-826. (PMID: 38165766) doi: 10.1177/17456916231214460
- Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, 85, 59-108.
- Raza, M., Jahangir, Z., Riaz, M. B., Saeed, M. J., & Sattar, M. A. (2025, Apr 21). Industrial applications of large language models. *Scientific Reports*, 15(1), 13755. Retrieved from <https://doi.org/10.1038/s41598-025-98483-1> doi: 10.1038/s41598-025-98483-1
- Ribeiro, M. T., Wu, T., Guestrin, C., & Singh, S. (2020, July). Beyond accuracy: Behavioral testing of NLP models with Check-List. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 4902-4912). Online: Association for Computational Linguistics. Retrieved from <https://aclanthology.org/2020.acl-main.442/> doi: 10.18653/v1/2020.acl-main.442
- Serapio-García, G., Safdari, M., Crepy, C., Sun, L., Fitz, S., Romero, P., ... Matarić, M. (2025, Dec 01). A psychometric framework for evaluating and shaping personality traits in large language models. *Nature Machine Intelligence*, 7(12), 1954-1968. doi: 10.1038/s42256-025-01115-6
- Sun, R. (2008). Introduction to computational cognitive modeling. In R. Sun (Ed.), *The cambridge handbook of computational psychology* (p. 3-20). Cambridge University Press.
- Turner, B. M., Steyvers, M., Merkle, E. C., Budescu, D. V., & Wallsten, T. S. (2014). Forecast aggregation via recalibration. *Machine learning*, 95(3), 261-289. Retrieved from <https://doi.org/10.1007/s10994-013-5401-4> doi: 10.1007/s10994-013-5401-4
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. In I. Guyon et al. (Eds.), *Advances in neural information processing systems* (Vol. 30). Curran Associates, Inc. Retrieved from https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf
- Vig, J., Gehrmann, S., Belinkov, Y., Qian, S., Nevo, D., Singer, Y., & Shieber, S. (2020). Investigating gender bias in language models using causal mediation analysis. In *Proceedings of the 34th international conference on neural information processing systems*. Red Hook, NY, USA: Curran Associates Inc.
- Wang, X., Jiang, L., Hernández-Orallo, J., Stillwell, D., Chen, S., Sun, L., ... Xie, X. (2026, April). Evaluating general-

purpose ai with psychometrics. *Commun. ACM*, 69(5), 92–102.
Retrieved from <https://doi.org/10.1145/3769688> doi:
10.1145/3769688

Wilson, R. C., & Collins, A. G. (2019, nov). Ten simple rules

for the computational modeling of behavioral data. *eLife*, 8,
e49547. Retrieved from <https://doi.org/10.7554/eLife.49547> doi: 10.7554/eLife.49547