Beyond benchmarks: The undone science of model validation

Ismail Harrando and Alexander T. Kindel
médialab, Sciences Po
ismail.harrando@sciencespo.fr
alexander.kindel@sciencespo.fr

Machine learning research evaluates models primarily in terms of their predictive accuracy. This priority is reflected in the field's emphasis on *benchmarks*: standardized sets of prediction problems, meant to represent concrete tasks we imagine models might help us perform (e.g., "question answering"). The benchmarking paradigm is the subject of intense debate within machine learning research; we offer a perspective on this debate focused on its consequences for *applications* of machine learning methods to scientific problems outside of computer science.

Benchmarking has both benefits and drawbacks for machine learning research. Its primary aim is to accelerate the field's search for effective models by constructing common modeling problems and comparing models that aim to solve the same ones. Evaluated on this point alone, benchmarking is a remarkably successful method: by providing a common signal for navigating a vast model space, benchmarks have helped to direct research efforts toward an expanding variety of informative targets and improved the performance of automated systems in an increasingly complex range of task domains (Donoho 2017). At the same time, defining success in machine learning as high scores on benchmarks risks prioritizing incremental performance gains over deeper insight into the phenomena being modeled. Publication requirements, prestige incentives, and career structures in machine learning all encourage framing problems as prediction tasks and reaching for state-of-the-art benchmark performance (a.k.a. "SOTA-chasing"; see Church & Kordoni 2022; Bender et al. 2021; Raji et al. 2021).

This paper examines benchmarking from an applied-scientific perspective. We focus in particular on the role of benchmarking in the adaptation of models for use in scientific problem-solving. We observe that predictive validation has become the primary basis for evaluating essentially any scientific use of models considered to be machine learning. Researchers rely on benchmark scores to evaluate the suitability of models for their empirical context or theoretical topic; conceive of relative predictive accuracy as a criterion for confirming or rejecting theories; and propose new benchmarks representing their areas of interest as prediction tasks. Comparatively little effort has been devoted to evaluating uses of models outside of this framework. Furthermore, very little of this activity is intended for the original purpose of benchmarking (i.e., searching for models); benchmarks are instead presented as tools enabling researchers with machine learning skills to improve the openness of science and coordinate research efforts toward widely shared priorities. We view this expanded epistemological role as characteristic of the benchmarking paradigm in applied settings.

We posit that much of the secondary interest in benchmarking reflects a need for a yet-undone computer science that would focus on the validation of machine learning models beyond their predictive accuracy (e.g., studying their invariance properties when used as measurement models). Our argument builds on a substantial body of work analyzing the epistemological role of prediction in computer science and machine learning, most centrally critiques of benchmarking and performance-driven evaluation (Raji et al. 2021; see also Langley 2011; Bender et al. 2021; Cheng et al. 2024; Felin & Holweg 2024; Fodor 2025; Chang et al. 2025). We concur with the concerns raised in this literature regarding construct validity, incentive structures, and the limits of benchmark-centered progress. We add to these critiques by considering the role that benchmarking practices play in the expanded community of users of machine learning results. We argue that the benchmarking paradigm makes it much easier to "use machine learning to do something" in applied settings by (1) licensing scientifically responsible model selection under essentially arbitrary decision-making conditions; (2) providing effective but minimal demonstrations of effective control over data through tutorials and example code; and (3) enabling inductive and exploratory modes of quantitative data analysis. This action-oriented form of interdisciplinarity leads to a situation in which many researchers can use machine learning models, but very little is known about the increasingly complex and increasingly widely-shared set of modeling decisions that constitute them.

To illustrate these patterns, we discuss the growing use of machine learning methods in computational social science (CSS). A strand of research in this area suggests social scientists develop benchmarks to evaluate competing models of social phenomena (Salganik et al. 2020; Pankowska et al. 2024). This development reflects a broader shift in parts of the social sciences toward viewing predictive performance as a primary criterion for theory evaluation and comparison (Watts 2014; Cranmer and Desmarais 2017). A related trend emphasizes inferential frameworks that trade statistical power for agnosticism about the underlying quantitative modeling strategy, other than asserting high predictive accuracy (e.g., Grimmer et al. 2021; Egami et al. 2023). The prediction-driven perspective on CSS represented by these trends has grown in importance as researchers increasingly turn to language models to derive measures from collections of texts. Beyond benchmark scores and the ubiquitous advice to spot-check examples of assigned classes or scores, methods for evaluating how LLMs perform these tasks do not yet exist. There are important exceptions to this general pattern—we note in particular the growing body of research documenting cultural, political, and socio-demographic biases in large language models (Gallegos 2024; Dehdashtian 2024). Yet in most areas of computational social science, researchers lack tools that would make it simple to (for example) compare how two models represent variation in data, or to determine how much of a specific kind of modeling error would result in a substantively meaningful change in an estimated quantity of interest. We envision a science of model validation that answers these questions by studying the relationship between specific architectural choices and kinds of data particular to applied settings of interest in the social sciences. Such a science would complement benchmarks by providing more interpretive infrastructure for specific models that are becoming standard tools.

References

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021, March). On the dangers of stochastic parrots: Can language models be too big?🦜. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency* (pp. 610-623).

Chang, Y., *et al.* (2024). A survey on evaluation of large language models. *ACM transactions on intelligent systems and technology,* 15(3), 1-45.

Cheng, Y., Chang, Y., & Wu, Y. (2025). A survey on data contamination for large language models. arXiv preprint arXiv:2502.14425.

Chollet, F. (2019). On the measure of intelligence. *arXiv preprint arXiv:1911.01547*.

Cranmer, S. J., & Desmarais, B. A. (2017). What can we learn from predictive modeling?. *Political Analysis*, 25(2), 145-166.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52(4), 281.

Church, K. W., & Kordoni, V. (2022). Emerging Trends: SOTA-Chasing. Natural Language Engineering, 28(2), 249-269.

Dehdashtian, S., et al. (2024). Fairness and Bias Mitigation in Computer Vision: A Survey. arXiv preprint arXiv:2408.02464.

Donoho, D. (2017). 50 years of data science. *Journal of Computational and Graphical Statistics*, 26:4, 745-766.

Egami, N., Hinck, M., Stewart, B., & Wei, H. (2023). Using imperfect surrogates for downstream inference: Design-based supervised learning for social science applications of large language models. *Advances in Neural Information Processing Systems*, *36*, 68589-68601.

Felin, T., & Holweg, M. (2024). Theory is all you need: AI, human cognition, and causal reasoning. Strategy Science, 9(4), 346-371.

Fodor, J. (2025). Line goes up? inherent limitations of benchmarks for evaluating large language models. arXiv preprint arXiv:2502.14318.

Gallegos, I. O., Rossi, R. A., Barrow, J., Tanjim, M. M., Kim, S., Dernoncourt, F., Yu, T., Zhang, R., & Ahmed, N. K. (2024). Bias and fairness in large language models: A survey. *Computational Linguistics*, 50(3), 1097-1179.

Grimmer J., Roberts M. E., Stewart. 2021 B. M. (2021). Machine Learning for Social Science: An Agnostic Approach. *Annual Review Political Science. 24*, 395-419.

Hossain M. M., Kovatchev V., Dutta P., Kao T., Wei E., & Blanco E. (2020). An analysis of natural language inference benchmarks through the lens of negation. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 9106-9118.

Langley, P. (2011). The changing science of machine learning. *Machine learning*, 82(3), 275.

Mitchell, M. (2021). Why AI is harder than we think. arXiv preprint arXiv:2104.12871.

Pankowska, P., Mendrik, A., Emery, T., & Garcia-Bernardo, J. (2024). The potential of benchmark challenges in the social sciences. *Social Science Information*, *63*(4), 498-519.

Raji, I. D., Bender, E. M., Paullada, A., Denton, E., & Hanna, A. (2021). AI and the everything in the whole wide world benchmark. *arXiv preprint arXiv:2111.15366*.

Riabi A., Mouilleron V., Mahamdi M., Antoun W., and Seddah D. (2025). Beyond Dataset Creation: Critical View of Annotation Variation and Bias Probing of a Dataset for Online Radical Content Detection. *Proceedings of the 31st International Conference on Computational Linguistics*, 8640–8663.

Salganik, M. J., et al. (2020). Measuring the predictability of life outcomes with a scientific mass collaboration. *Proceedings of the National Academy of Sciences*, 117(15), 8398-8403.

Watts, D. J. (2014). Common sense and sociological explanations. *American Journal of Sociology*, *120*(2), 313-351.