We Keep Evaluating LLMs, But Are We Learning Anything?

Dimas Cassimiro Nascimento, Daliton da Silva

[1]Universidade Federal do Agreste de Pernambuco (UFAPE)
Garanhuns - PE – Brazil

{dimas.cassimiro,daliton.silva}@ufape.edu.br

Abstract. *Large language model (LLM) research has become increasingly dominated by benchmark-oriented evaluations that measure performance on predefined tasks and datasets. While such evaluations are useful for tracking progress, they often provide limited scientific insight into how these systems reason, generalize, fail, or support real-world applications. In this article, we argue that LLM research should move beyond incremental leaderboard improvements and adopt a broader scientific agenda inspired by the historical evolution of machine learning research. Drawing analogies with established research areas in machine learning, we propose a set of guidelines for conducting more rigorous and impactful LLM research. These guidelines include investigating internal reasoning mechanisms, developing calibration and uncertainty estimation methods, addressing bias and fairness systematically, studying generalization to novel domains, exploring benchmark-free evaluation paradigms, improving methodological transparency, and, more ambitiously, using LLMs as instruments for scientific discovery rather than merely as objects of evaluation.*

1. Introduction

The release of each new large language model (LLM) is typically accompanied by many articles evaluating it on a growing list of benchmarks. The formula is familiar: select a model (or several), choose a task or dataset, run inference, report accuracy, and conclude that the model either excels or falls short. While such evaluations have their place, they increasingly resemble the early days of machine learning research, where papers would report a single accuracy number on datasets and consider that alone a sufficient scientific contribution [Lipton and Steinhardt 2019]. Recent years have seen an explosion of benchmark-oriented studies covering reasoning, coding, mathematics, multilinguality, factuality, and safety [Srivastava et al. 2023, Chang et al. 2024]. Although these benchmarks have been valuable for tracking rapid progress in LLM capabilities, they have also encouraged a research culture centered on leaderboard improvements rather than on deeper understanding of model behavior, limitations, and scientific utility.

Overview of Related Works. Earlier phases of machine learning research also emphasized empirical comparisons on standardized datasets, often prioritizing marginal gains in predictive performance over theoretical or methodological advances. Over time, however, the field evolved toward richer and more ambitious questions. Modern machine learning research investigates topics such as optimization dynamics [Bengio et al. 2017], robustness to adversarial and distributional shifts [Madry et al. 2017, Ovadia et al. 2019], fairness and bias mitigation [Barocas and Selbst 2016, Mehrabi et al. 2021], interpretability [Wiggins and Tejani 2022], and transfer learning across domains [Weiss et al. 2016].

These research directions seek not merely to measure performance, but to understand why models behave as they do, when they can be trusted, and how they can be improved in principled ways.

A similar maturation process is now beginning to emerge in LLM research. Beyond benchmark evaluations, recent studies have explored mechanistic interpretability [Meng et al. 2022], reasoning architectures [Yao et al. 2023], tool use and autonomous agents [Yao et al. 2022], calibration and uncertainty-aware reasoning [Tan et al. 2025], and the use of LLMs for scientific discovery [Li et al. 2025]. Other works have highlighted critical limitations of benchmark-centric evaluation, including data contamination, memorization effects, and poor correspondence between benchmark performance and real-world reliability [Shi et al. 2024]. Collectively, these studies suggest that the central scientific questions surrounding LLMs are no longer simply whether a model performs well on a benchmark, but rather how these systems reason, generalize, fail, adapt, and support human scientific activity.

In this article, we argue that LLM research should move decisively beyond static benchmark comparisons and adopt a broader scientific agenda inspired by the evolution of machine learning research. Rather than treating LLMs as opaque systems to be ranked on leaderboards, we advocate investigating them as objects of scientific inquiry and as instruments for discovery. To support this perspective, we identify a set of promising research directions, including mechanistic studies of reasoning, calibration and uncertainty estimation, bias and fairness analysis, generalization to novel domains, multi-agent collaboration, benchmark-free evaluation methodologies, and the use of LLMs to accelerate scientific discovery. By drawing explicit parallels with established machine learning research areas, we aim to provide both a conceptual framework and a practical set of guidelines for conducting more impactful, rigorous, and scientifically meaningful LLM research.

2. The Benchmarking Trap

The proliferation of LLM benchmarks has created a research culture that prioritizes incremental accuracy improvements over deeper understanding. A typical paper might evaluate GPT-4, Claude, and Llama on a question-answering dataset, report scores, and conclude with vague statements about model capabilities. This approach suffers from several limitations. First, benchmarks suffer from data contamination, where evaluation examples inadvertently appear in training data. This makes it difficult to know whether a model is genuinely reasoning or simply recalling memorized content. Second, benchmark performance often fails to predict real-world behavior, especially in high-stakes domains where reliability and calibration matter more than raw accuracy. Third, this style of research treats LLMs as black boxes, generating little insight into why they succeed or fail. Real-world deployment [McIntosh et al. 2025, Springer et al. 2025] involves messy, subjective aspects like creativity and adaptability that static benchmarks cannot capture.

3. Learning from Machine Learning: An Analogy

To see a path forward, we can look at how machine learning research matured beyond simple benchmarking. In the 2000s and early 2010s, many papers would report accuracy on a dataset and stop. Today, the field has diversified into richer research areas. Table 1

draws parallels between these established areas and promising directions for LLM research.

Tabela 1. Analogies between traditional ML research and more interesting LLM research directions.

Traditional ML Research Area	Traditional ML Question	LLM Research Analogy
Model Improvement	How can we design better architectures?	How can we design better LLM architectures (e.g., recurrent or reasoning-focused designs)? [Yao et al. 2023]
Hyperparameter Tuning	How do hyperparameters affect convergence?	How do prompts, decoding strategies, and in-context examples affect LLM behavior?
Calibration	How well-calibrated are model probabilities?	Are LLM confidence scores reliable? How can we improve calibration? [Tan et al. 2025]
Bias Reduction	How do models perpetuate societal biases?	How do LLMs encode and amplify cultural or scientific biases? [Fehring et al. 2025]
Generalization & OOD	How do models perform on unseen distributions?	How do LLMs generalize to novel scientific domains or tasks? [Wang et al. 2023]
Ensemble Methods	Can combining models improve performance?	Can multi-agent LLM systems outperform single models?
Interpretability	What do models learn internally?	How do LLMs represent concepts like novelty, analogy, or scientific reasoning?
Scientific Discovery	Can ML accelerate science?	Can LLMs generate novel hypotheses or automate scientific reasoning? [Li et al. 2025]

4. Guidelines for More Interesting LLM Research

Based on the analogies presented in Table 1, we propose a set of guidelines for conducting LLM research that moves beyond simple benchmark comparisons. These guidelines are inspired by the evolution of machine learning research, which gradually shifted from isolated performance evaluations toward deeper investigations into model behavior, reliability, interpretability, and scientific utility. Rather than viewing LLMs merely as systems to be ranked on leaderboards, researchers should approach them as complex computational artifacts whose mechanisms and limitations deserve systematic investigation. Importantly, the goal is not to abandon evaluation altogether. Benchmarks remain valuable for tracking progress and identifying broad trends, but they should serve as a starting point rather than the endpoint of inquiry. The most impactful LLM research will likely come from studies that explain *why* models behave in certain ways, identify conditions under which they fail, and explore how they can support scientific reasoning and decision-making.

4.1. Study Model Mechanisms, Not Just Outcomes

LLM research should move beyond reporting final outputs and investigate the mechanisms that produce them. Benchmark-centric evaluations typically treat models as black boxes, measuring performance without examining how reasoning, abstraction, or representation emerge internally. While useful for comparison, this approach provides limited scientific understanding and contributes little to the development of more interpretable and controllable systems. Researchers should therefore prioritize studies that analyze internal representations, attention dynamics, reasoning trajectories, and architectural properties. Questions such as how concepts are encoded, how reasoning unfolds across layers, and why specific failure modes occur are substantially more informative than isolated accuracy metrics. Recent work on recurrent reasoning architectures and mechanistic interpretability illustrates this direction [Yao et al. 2023, Meng et al. 2022]. Similarly, studies connecting internal representations to analogy and reasoning abilities suggest that mechanistic analyses can reveal principles that generalize across tasks and models.

4.2. Develop Calibration and Uncertainty Methods

Future LLM research should place greater emphasis on calibration and uncertainty estimation rather than focusing exclusively on predictive accuracy. In many real-world scenarios, especially in medicine, law, and scientific research, an incorrect but highly confident answer may be more harmful than an explicit acknowledgment of uncertainty. Yet current evaluations rarely examine whether LLM confidence aligns with actual correctness. Researchers should investigate methods that allow models to estimate, communicate, and reason about uncertainty more reliably. This includes studying confidence calibration, robustness under distribution shifts, self-consistency strategies, and uncertainty-aware decision-making pipelines. Prior works [Guo et al. 2017, Tan et al. 2025] demonstrate the importance of integrating uncertainty estimates into evaluation and reasoning processes. More broadly, LLM studies should ask not only whether a model is correct, but whether it knows when it is likely to be wrong.

4.3. Address Bias and Fairness Systematically

Bias and fairness should become central concerns in LLM research rather than secondary evaluation criteria. Because LLMs are trained on large-scale web data, they inevitably inherit social, cultural, linguistic, and scientific biases present in those corpora. Benchmark scores alone often conceal these issues, masking systematic disparities across languages, demographic groups, or cultural contexts. Researchers should therefore adopt systematic methodologies for identifying, documenting, and mitigating bias [Mehrabi et al. 2021] throughout the entire modeling pipeline. This includes investigating how biases emerge during pre-training and fine-tuning, how prompting strategies affect outputs, and how fairness interventions impact performance.

4.4. Study Generalization to Novel Domains

A central goal of future LLM research should be understanding whether models genuinely generalize beyond their training distributions. Strong performance on familiar benchmarks does not necessarily imply robust reasoning in novel scientific, professional, or

interdisciplinary settings. Many benchmark tasks remain close to distributions already represented in training corpora, making it difficult to distinguish abstraction from memorization. Researchers should therefore prioritize studies involving out-of-distribution reasoning, domain transfer, and adaptation to unfamiliar tasks. Instead of asking whether a model reproduces benchmark answers, future work should investigate whether it can operate reliably under epistemic novelty and learn transferable conceptual structures. Research on out-of-distribution robustness [Ovadia et al. 2019] and emerging scientific foundation models such as Walrus and AION-1 [Wang et al. 2023] suggests that future systems may integrate textual, mathematical, and scientific representations to achieve broader forms of generalization beyond linguistic pattern matching.

4.5. Use LLMs as Tools for Scientific Discovery

Perhaps the most transformative direction for LLM research is treating these systems as instruments for scientific discovery rather than merely as objects of evaluation. In this paradigm, the key objective is not simply improving benchmark performance, but understanding whether LLMs can support hypothesis generation, experimental design, literature synthesis, and scientific reasoning. Researchers should therefore investigate how LLMs can augment human creativity and accelerate scientific workflows. This includes studying multi-agent collaboration, automated knowledge integration, tool-augmented reasoning, and hypothesis exploration systems. Such approaches position LLMs not merely as predictive systems, but as collaborative research tools capable of contributing to the production of new knowledge. The central challenge is to design evaluation methods that capture this broader scientific role, including the novelty, usefulness, reliability, and verifiability of the contributions generated with LLM support.

4.6. Develop Benchmark-Free Evaluation Methods

Future LLM research should reduce its dependence on static benchmarks and explore evaluation paradigms that better capture dynamic reasoning, adaptability, and real-world behavior. Traditional benchmarks are increasingly affected by limitations such as data contamination, benchmark saturation, and narrow task formulations, making them insufficient for understanding how models behave in open and evolving environments. Researchers should therefore investigate alternative forms of evaluation based on interaction, adversarial testing, open-ended reasoning, and adaptive task generation. Benchmark-free approaches such as LLM-Crowdsourced and broader capability-oriented evaluations [Srivastava et al. 2023] illustrate how dynamic evaluation frameworks can reveal properties that static datasets often overlook, including memorization effects, and reasoning consistency. Rather than replacing benchmarks entirely, these approaches should complement traditional evaluation with richer analyses of model behavior.

4.7. Report Methodological Details Transparently

Methodological transparency should become a fundamental requirement in LLM research. Because LLM outputs can vary substantially depending on prompts, decoding parameters, model versions, and inference settings, insufficient reporting makes replication difficult and limits the scientific value of published results. Benchmark scores alone are often meaningless without a clear description of the experimental setup that produced them. Researchers should therefore adopt rigorous reporting practices that document

prompts, model configurations, sampling strategies, validation procedures, and evaluation protocols in sufficient detail to enable reproducibility. Frameworks such as COREQ+LLM [Fehring et al. 2025] highlight the importance of explicitly describing how LLMs are integrated into research workflows and how outputs are validated. Greater methodological transparency not only improves reproducibility, but also enables cumulative scientific progress by allowing future studies to critically assess previous findings.

5. Conclusions and Future Work

In this article, we outlined several promising directions capable of moving the field of LLM-based research beyond benchmark-centric evaluation, including mechanistic investigations of internal reasoning processes, calibration and uncertainty estimation, fairness and bias mitigation, analyses of out-of-distribution generalization, benchmark-free evaluation paradigms, and transparent methodological reporting. Most importantly, we highlighted the growing potential of LLMs as instruments for scientific discovery, capable of assisting hypothesis generation, literature synthesis, and interdisciplinary reasoning.

These directions share a common characteristic: they treat LLMs not merely as products to be ranked, but as objects of scientific inquiry and as tools for inquiry themselves. Rather than asking only whether a model achieves high accuracy on curated datasets, they seek to understand how these systems reason, why they fail, when they can be trusted, and how they can augment human scientific activity. As the field matures, the most impactful contributions will likely come from researchers who move beyond asking *"How accurate is this model"*? to asking *"What can these models teach us about intelligence, language, reasoning, and scientific discovery"*?

Referências

Barocas, S. and Selbst, A. D. (2016). Big data's disparate impact. *California Law Review*, 104(3):671–732.

Bengio, Y., Goodfellow, I., Courville, A., et al. (2017). *Deep learning*, volume 1. MIT press Cambridge, MA, USA.

Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., Chen, H., Yi, X., Wang, C., Wang, Y., et al. (2024). A survey on evaluation of large language models. *ACM transactions on intelligent systems and technology*, 15(3):1–45.

Fehring, L., Frings, J., Rust, P., Kempny, C., Thürmann, P. A., and Meister, S. (2025). Extension of the consolidated criteria for reporting qualitative research guideline to large language models (coreq+ llm): Protocol for a multiphase study. *JMIR Research Protocols*, 14(1):e78682.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. (2017). On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR.

Li, Y. et al. (2025). Multi agent large language models for biomedical hypothesis generation in drug combination discovery. *iScience*, 28(12):113984.

Lipton, Z. C. and Steinhardt, J. (2019). Troubling trends in machine learning scholarship: Some ml papers suffer from flaws that could mislead the public and stymie future research. *Queue*, 17(1):45–77.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. (2017). Towards deep learning models resistant to adversarial attacks. *stat*, 1050:9.

McIntosh, T. R., Susnjak, T., Arachchilage, N., Liu, T., Xu, D., Watters, P., and Halgamuge, M. N. (2025). Inadequacies of large language model benchmarks in the era of generative artificial intelligence. *IEEE Transactions on Artificial Intelligence*.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys*, 54(6):1–35.

Meng, K., Bau, D., Andonian, A., and Belinkov, Y. (2022). Locating and editing factual associations in gpt. *Advances in neural information processing systems*, 35:17359–17372.

Ovadia, Y., Fertig, E., Ren, J., et al. (2019). Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in Neural Information Processing Systems (NeurIPS)*, 32.

Shi, W., Ajith, A., Xia, M., Huang, Y., Liu, D., Blevins, T., Chen, D., and Zettlemoyer, L. (2024). Detecting pretraining data from large language models. In *International Conference on Learning Representations*, volume 2024, pages 51826–51843.

Springer, J. M., Goyal, S., Wen, K., Kumar, T., Yue, X., Malladi, S., Neubig, G., and Raghunathan, A. (2025). Overtrained language models are harder to fine-tune. *arXiv preprint arXiv:2503.19206*.

Srivastava, A., Rastogi, A., Rao, A., Shoeb, A. A. M., Abid, A., Fisch, A., Brown, A. R., Santoro, A., Gupta, A., Garriga-Alonso, A., et al. (2023). Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on machine learning research*.

Tan, H., Zhan, S., Jia, F., Zheng, H.-T., and Chan, W. K. V. (2025). A hierarchical framework for measuring scientific paper innovation via large language models. *Information Sciences*, page 122787.

Wang, H., Fu, T., Du, Y., Gao, W., Huang, K., Liu, Z., Chandak, P., Liu, S., Van Katwyk, P., Deac, A., et al. (2023). Scientific discovery in the age of artificial intelligence. *Nature*, 620(7972):47–60.

Weiss, K., Khoshgoftaar, T. M., and Wang, D. (2016). A survey of transfer learning. *Journal of Big data*, 3(1):9.

Wiggins, W. F. and Tejani, A. S. (2022). On the opportunities and risks of foundation models for natural language processing in radiology. *Radiology: Artificial Intelligence*, 4(4):e220119.

Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T., Cao, Y., and Narasimhan, K. (2023). Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36:11809–11822.

Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., and Cao, Y. (2022). React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*.