PUBLIC SUBMISSION

Received: May 29, 2025
Tracking No. mb9-u0gh-783d
Comments Due: May 28, 2025 **Submission Type:** Web

Docket: NSF-2025-OGC-0001
NITRD_FRDOC_0001

Comment On: NSF-2025-OGC-0001-0001
Request for Information: Development of a 2025 National Artificial Intelligence Research and Development Strategic Plan

Document: NSF-2025-OGC-0001-DRAFT-0257
Comment on FR Doc # 2025-07332

Submitter Information

Organization: MLCommons

General Comment

See attached file(s)

Attachments

MLCommonsNSF-2025-OGC0001

ABOUT MLCOMMONS AND ITS INTEREST IN THIS REQUEST FOR INFORMATION

This response to the National Science Foundation's Request for Information on the Development of a 2025 National Artificial Intelligence (AI) Research and Development (R&D) Strategic Plan is submitted on behalf of MLCommons®.

MLCommons is a non-profit consortium that aims to accelerate the benefits of machine learning and artificial intelligence. Our members and partners include over 125 organizations from around the world, many of which are leading technology companies, startups, academics, and nonprofits that are actively researching, developing, and deploying artificial intelligence products for customers. Critically, our founding membership includes academic researchers at the forefront of machine learning research, and the research community continues to be core to our membership helping to lead many of our working groups. MLCommons acts as a neutral nexus for commercial and non-commercial actors to collaborate on tools that advance the field.

We create, operate and maintain community assets, especially benchmarks and datasets, that facilitate developing and evaluating artificial intelligence (AI) systems in pursuit of our mission to "make artificial intelligence better for everyone."[1] The original project that brought MLCommons into being is a benchmarking suite called MLPerf®, which provides unbiased evaluations of training and inference speed for AI hardware and software.[2] These measurements enable a fair comparison of competing systems, accelerate ML progress through fair and useful measurement, enforce reproducibility to ensure reliable results, and do so in an open and collaborative way to keep benchmarking affordable for all participants. We have also developed and released a number of open datasets for AI training, including images of everyday objects from around the world and spoken words across dozens of languages.

STANDARD SAFETY AND SECURITY BENCHMARKS

In December, 2024, we launched AILuminate, a first-of-its-kind benchmark to measure the safety of large language models.[3] AILuminate provides a series of safety grades[4] for the most widely-used LLMs by assessing LLM responses to over 24,000 test prompts across twelve categories of hazards. The AILuminate Assessment Standard is publicly available and details the methodology for assessment.[5] None of the LLMs evaluated were given any advance knowledge of the evaluation prompts, nor access to the evaluator model used to assess responses. This independence provides a methodological rigor uncommon in standard academic research and ensures an empirical analysis that can be trusted by industry and academia alike. This benchmark was developed by the MLCommons AI Risk and Reliability

[1] Machine learning is one of the key techniques through which AI systems are built.
[2] Peter Mattson, et al, "MLPerf: An Industry Standard Benchmark Suite for Machine Learning Performance," IEEE Xplore, accessed February 1, 2024, https://ieeexplore.ieee.org/abstract/document/9001257.
[3] "MLCommons Launches AILuminate, First-of-its-Kind Benchmark to Measure the Safety of Large Language Models," MLCommons, December 2024, https://mlcommons.org/2024/12/mlcommons-ailuminate-v1-0-release/
[4] AILuminate benchmark for general purpose AI chat models. https://ailuminate.mlcommons.org/benchmarks/
[5] The AILumiante Assessment Standard is accessible at https://drive.google.com/file/d/1jVYoSGJHtDo1zQLTzU7QXDkRMZIberdo/view

working group[6] — a team of leading AI researchers from institutions including Stanford University, Columbia University, and TU Eindhoven, civil society representatives, and technical experts from Google, Intel, NVIDIA, Meta, Microsoft, Qualcomm Technologies, Inc., and other industry giants committed to a standardized approach to AI safety.

The working group has released AILuminate in French and Chinese, and is working on expanding coverage to Hindi. The working group is also working toward a benchmark that covers multi-modal language models and agentic AI.[7,8] We plan to continue making ongoing updates as AI technologies continue to advance.

MAKING AI DATA SETS DISCOVERABLE, PORTABLE AND INTEROPERABLE

MLCommons also has efforts in data standards and data set curation. Training and evaluating AI systems depends on rigorous, standardized test datasets, and to this end we have invested in creating an open ecosystem of datasets for AI training and evaluation. We have developed and released a number of open datasets, including images of everyday objects from around the world and spoken words across dozens of languages.[9] We have also invested in a metadata standard, called “Croissant,” in order to simplify how data is used by ML tools and frameworks.[10] Croissant makes datasets more discoverable, portable and interoperable, thereby addressing significant challenges in ML data management and responsible AI.

Croissant describes datasets’ attributes, the resources they contain, and their structure and semantics in a way that streamlines their usage and sharing within the ML community while fostering responsible AI practices. Croissant does not require changing the underlying data representation, and can therefore be easily added to existing datasets, and adopted by dataset repositories. Croissant has been successfully integrated into three dataset repositories: Hugging Face datasets, Kaggle datasets, and OpenML, yielding over 400,000 datasets in the Croissant format. A fourth repository platform, Harvard's Dataverse, has added support for Croissant in its beta channel.

The Croissant vocabulary is an extension to schema.org, a machine-readable standard to describe structured data, used by over 50M datasets on the Web, which allows the datasets to be discoverable through dataset search engines. Croissant enables datasets to be loaded into different ML platforms without the need for reformatting. Popular ML frameworks like TensorFlow, JAX and PyTorch can already load Croissant datasets via the TensorFlow Datasets library. Additionally, by providing operationalized documentation, Croissant users can easily understand the best practices for contributing to and utilizing the data.

---

[6] MLCommons AI Risk & Reliability. https://mlcommons.org/working-groups/ai-risk-reliability/ai-risk-reliability/
[7] “MLCommons Releases AILuminate LLM v1.1, Adding French Language Capabilities to Industry-Leading AI Safety Benchmark.” February 2025. https://mlcommons.org/2025/02/ailumiate-v1-1-fr/
[8] “MLCommons Announces Expansion of Industry-Leading AILuminate Benchmark.” May 29, 2025. https://mlcommons.org/2025/05/nasscom/
[9] See information on our Cognata, People’s Speech, Dollar Street, and Multilingual Spoken Words data sets at https://mlcommons.org/datasets/.
[10] MLCommons, "Croissant Working Group," accessed July 15, 2024, https://mlcommons.org/working-groups/data/croissant/.

MANAGING RISK IN AI WILL REQUIRE STANDARDIZED BENCHMARKS

A modern AI system requires empirical measurement to understand its characteristics, including misuse risks. Traditional approaches to managing risk in technology systems that rely solely on process compliance will potentially increase friction without delivering intended safety results. Instead of managing risk through process compliance alone, modern AI requires a strong emphasis on measurement to mitigate risk.

Evaluating an AI system for misuse risk is unlike testing conventional software code that is intended to produce discrete and objectively verifiable behavior. Defining a test set for an AI model that has effective coverage of the potential input space is a nascent measurement science.[11,12] This is because the latest iteration of said models, known as language models, are able to directly interact in natural language with an exponentially large number of possible input sentences, making full coverage intractable.[13] Measurement for misuse risk is also challenging because of the many aspects of responsible development that need to be evaluated including avoiding physical harms, resistance to malicious uses, fairness, misinformation, and privacy. Each of these requires dedicated tests and evaluation resources, as well as robust input from a wide range of stakeholders and experts. Unlike the more objective measurement of hardware speed or model performance, these varied aspects of safety contain an inherent subjectivity and ambiguity. Finally, AI model deployment is iterative and requires ongoing monitoring and measurement.

While managing risk in modern AI is challenging in ways that dramatically differ from traditional software risk management, there are lessons to be learned in how other industries approach risk management and safety. In complex systems that necessarily interact with the unpredictability of the physical world, such as automobiles or planes, standardized approaches to safety testing have been adopted with success. No automobile can be deemed perfectly safe in all possible circumstances, but we expect automobiles to meet standard safety benchmarks.

We believe in mirroring this approach to create standardized safety benchmarks in AI. Such benchmarks will create a common direction for research efforts across companies and academic institutions, and raise the bar for safety across the industry. Furthermore, if built with care, the benchmarks can produce safety analyses that are comprehensible to purchasers, policy makers, and the public.

STANDARDIZED BENCHMARKS DEMAND IMPROVEMENTS IN THE STATE OF THE ART

In order for standardized benchmarks to be successful, they will need to be rigorous enough to substantially reduce risk and yet be delivered at a moderate enough cost to be widely adopted

---

[11] Amershi, Saleema, et al. "Software engineering for machine learning: A case study." 2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP). IEEE, 2019.

[12] Sculley, David, et al. "Hidden technical debt in machine learning systems." Advances in neural information processing systems 28 (2015).

[13] Rishi Bommasani et al., "On the Opportunities and Risks of Foundation Models," arXiv, August 2021, https://arxiv.org/abs/2108.07258.

across the industry, from startups to large corporations. At present, AI safety testing faces a tension between (1) rigorous but expensive approaches that depend on manual prompting and/or rating and (2) more scalable (cost-effective and repeatable) automated approaches that lack the rigor of humans-in-the-loop. More research will be required to design robust algorithms for evaluating AI output that accurately reflect human perceptions so that robust AI safety testing can be scaled effectively and widely adopted. There are many examples of ongoing research in this area that are redefining fundamental assumptions in how we measure truth, confidence, and trust in generative AI systems[14,15,16,17,18] but they need support and rapid integration of resulting innovations for industrial use.

MLCommons is developing an approach to standard benchmarks intended to support rapid evolution of AI safety benchmarking technology. We are building a platform that can accept and manage tests from academic and industry partners, and support multiple evaluation methodologies including algorithms, evaluation models, and human raters. Our projection is that a hybrid approach using humans to shore up the limitations of automatic models is likely for the near term, but we are engineering our platform in a modular manner to accept more substantial innovations. We are also developing the necessary statistical and cost models to support test data generation and response evaluation at industry scale.

BENCHMARKS WILL NEED CONSTANT EVALUATION AND CALIBRATION

Standardized benchmarks will also require ongoing research and novel test data creation to ensure they remain durable and applicable to evolving AI models.[19] Current academic research and public leaderboards tend to focus on static test datasets for AI, but these datasets quickly fail as evaluation resources because, whether intentionally or unintentionally, models become trained and overfit to perform well against the static dataset.[20,21,22,23] Even the models that are used to rate AI outputs can be subject to overfitting unless constantly improved. A constant improvement cycle for safety benchmarks will be needed to prevent overfitting and keep pace with AI technology development, and will demand evolution of both prompts and evaluation methodologies. Both conventional policy development and standards processes need to design for this evolution – which must iterate faster than policies or standards are typically revised.

---

[14] Sven Gowal, et al.,, "Improving Robustness using Generated Data," arXiv, October 2021, https://arxiv.org/abs/2110.09468.
[15] Shira Wein, et al., "Follow the leader(board) with confidence: Estimating p-values from a single test set with item and response variance," ACL Anthology, 2023, https://aclanthology.org/2023.findings-acl.196/.
[16] Daniel Deutsch, et al., "A Statistical Analysis of Summarization Evaluation Metrics Using Resampling Methods," Transactions of the Association for Computational Linguistics (TACL), 2021, https://aclanthology.org/2021.tacl-1.67/.
[17] Barbara Plank, et al., "Linguistically debatable or just plain wrong?," ACL Anthology, 2014, https://aclanthology.org/P14-2083/.
[18] Lora Aroyo and Christ Welty, "Truth Is a Lie: Crowd Truth and the Seven Myths of Human Annotation," AI Magazine, https://ojs.aaai.org/aimagazine/index.php/aimagazine/article/view/2564.
[19] Prabha Kannan, "How Trustworthy Are Large Language Models Like GPT?," Stanford HAI News, Aug 23, 2023, https://hai.stanford.edu/news/how-trustworthy-are-large-language-models-gpt.
[20] Potential references Carlini, Nicholas, et al. "Quantifying memorization across neural language models." arXiv preprint, February 2022, arXiv:2202.07646.
[21] Douwe Kiela, et al., "Dynabench: Rethinking Benchmarking in NLP," arXiv, April 2021, arXiv:2104.14337.
[22] Tirumala, Kushal, et al. "Memorization without overfitting: Analyzing the training dynamics of large language models." Advances in Neural Information Processing Systems 35 (2022): 38274-38290.
[23] Bordt, Sebastian, Harsha Nori, and Rich Caruana. "Elephants Never Forget: Testing Language Models for Memorization of Tabular Data." NeurIPS 2023 Second Table Representation Learning Workshop. 2023.

Further, we will need calibration to ensure that the benchmarks truly measure the impact of AI on the user in the context of real-world use cases and applications. AI output evaluations are necessarily subjective, and may be done by either humans or algorithms that imperfectly emulate humans (both sources of measurement error). As a result, the tests will need to be iteratively calibrated with human involvement to correlate test prompts and output evaluations with actual user experience as closely as possible.[24] This calibration will require novel methodology for measuring sociotechnical systems, which is often more complex than strictly technical evaluation.[25]

BENCHMARKING IS AS MUCH ORGANIZATIONAL AS IT IS TECHNOLOGICAL

In creating and operating the MLPerf family of benchmarks over the last five years, we have observed that AI benchmarks require a combination of technological innovation and organizational commitment. Cutting edge test data and evaluation methodologies do not work unless supported by less glamorous software infrastructure to manage submissions and results, fair governance and policies to resolve disputes, and a community of experts to build, maintain, and improve the technology.

We are committed to working toward a future in which industry standard AI safety benchmarks exist for the most common AI applications, and in which these benchmarks are relied upon for evaluating safety by both vendors and purchasers. We believe MLCommons as an institution is equipped to take on responsibility for building and operating benchmarks that are not susceptible to over-fitting. We aim to build dynamic benchmarks that are connected to social science research and updated accordingly to accurately represent societal preferences. The benchmarks and technology platform we are building will provide a robust model that industry can engage with, akin to the certification model found in other mature, high-productivity, low-risk industries.

[24] Victor Dibia, et al., "Aligning Offline Metrics and Human Judgments of Value for Code Generation Models," ACL Anthology, 2023, https://aclanthology.org/2023.findings-acl.540.pdf.

[25] Abigail Jacobs and Hannah Wallach, "Measurement and Fairness," arXiv, December 2019, https://arxiv.org/pdf/1912.05511.pdf.