ML Commons

ABOUT MLCOMMONS AND ITS INTEREST IN THIS REQUEST FOR INFORMATION

This response to the National Institute of Standards and Technology's Request for Information Related to Assignments under Sections 4.1, 4.5 and 11 of the Executive Order Concerning Artificial Intelligence (Sections 4.1, 4.5, and 11) ("The RFI") 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.

In November, 2023, we announced the formation of an AI Safety Working Group, which is an open working group that anyone from the AI community can participate in.[3] The working group will develop a platform and pool of tests from many contributors to support AI safety benchmarks for diverse use cases. The group's initial focus will be developing safety benchmarks for large language models (LLMs), building on groundbreaking work done by researchers at Stanford University's Center for Research on Foundation Models and its Holistic Evaluation of Language Models (HELM). We believe standard AI safety benchmarks will become a vital element of a successful approach to AI safety. This aligns with responsible development and risk-based policy frameworks such as NIST's AI Risk Management Framework.

As NIST works to implement the Executive Order, we believe MLCommons can both inform and support future actions. More specifically, MLCommons can provide a toolkit of useful benchmarks and datasets for policymakers and government agencies that will support addressing many of the issues

[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 Announces the Formation of AI Safety Working Group," MLCommons, October 2023, https://mlcommons.org/2023/10/mlcommons-announces-the-formation-of-ai-safety-working-group/.

identified in the RFI. Our work is most directly relevant to issues of benchmarking as it relates to risk management, safety, security, and other goals in Question 1, particularly Question (1)(a)(2).

MANAGING RISK IN AI WILL REQUIRE STANDARDIZED BENCHMARKS

The RFI rightly asks what is necessary to develop and provide ongoing oversight of AI, in a manner that addresses risk, protects people's rights, and advances equity, among other concerns. Making progress against these objectives is intimately connected to effective evaluation and measurement of how AI systems perform across a range of attributes, including accuracy, safety, bias, security, and energy use.[4]

Standardized metrics and benchmarks are crucial to effective evaluation and measurement of AI, and the National AI strategy should make them a key focus.

A modern AI system requires empirical measurement to understand its characteristics, including safety. Traditional approaches to managing risk in technology systems that rely solely on process compliance will potentially increase friction without delivering intended safety results. Safety-by-process that focuses on documenting and auditing training inputs, without measuring the behavior of the end model, may not ensure safety and can produce unexpected failure modes. For example, requiring use of high-quality training data does not necessarily imply that the dataset is suited to the particular issue one is trying to address; it is possible for a well-documented training process to use high quality training data that inadvertently under-represents a real-world problem class resulting in the model's outputs on problems of that class being incorrect.[5] Instead of managing risk through process compliance alone, modern AI requires a strong emphasis on measurement to mitigate risk.

Testing an AI system for safety 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.[6,7] 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.[8] Measurement for safety 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.

---

[4] "MLPerf Inference v1.0 Results with First Power Measurements," MLCommons, April 2021, https://mlcommons.org/2021/04/mlperf-inference-v1-0-results-with-first-power-measurements/.

[5] Joy Buolamwini and Timnit Gebru, "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification," Proceedings of the 1st Conference on Fairness, Accountability, and Transparency, 2018, https://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf.

[6] 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.

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

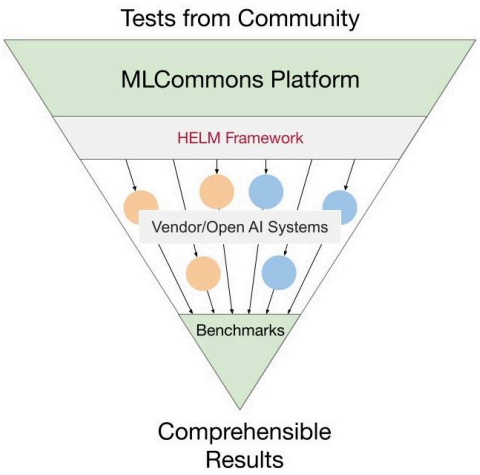
[8] Rishi Bommasani et al., "On the Opportunities and Risks of Foundation Models," arXiv, August 2021, https://arxiv.org/abs/2108.07258.

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 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[9,10,11,12,13] 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

[9] Sven Gowal, et al.,, "Improving Robustness using Generated Data," arXiv, October 2021, https://arxiv.org/abs/2110.09468.
[10] 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/.
[11] 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/.
[12] Barbara Plank, et al., "Linguistically debatable or just plain wrong?," ACL Anthology, 2014, https://aclanthology.org/P14-2083/.
[13] 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.

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. The engine at the heart of the platform is built on the proven Stanford HELM system, with refactoring to make it more robust and increase support for a wide range of tests and AI models. 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.[14] 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.[15,16,17,18] 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.

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.[19] This calibration will require novel methodology for measuring sociotechnical systems, which is often more complex than strictly technical evaluation.[20]

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

[14] 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.
[15] Potential references Carlini, Nicholas, et al. "Quantifying memorization across neural language models." arXiv preprint, February 2022, arXiv:2202.07646.
[16] Douwe Kiela, et al., "Dynabench: Rethinking Benchmarking in NLP," arXiv, April 2021, arXiv:2104.14337.
[17] 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.
[18] 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.
[19] 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.
[20] Abigail Jacobs and Hannah Wallach, "Measurement and Fairness," arXiv, December 2019, https://arxiv.org/pdf/1912.05511.pdf.

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.

IMPORTANCE OF GLOBAL COLLABORATION

MLCommons counts as its members organizations and academic researchers from all across the world. We view this as a distinct strength, in that we are able to develop and share insights, standards and benchmarks that are broadly useful across a large range of organizations. Much like the early development of Internet protocols in the late 20th century, we're at a phase of development in AI where coordination on standards will facilitate greater collaboration and drive innovation. But in the case of building trustworthy AI, the development of open, global standards will also drive greater transparency and oversight of AI systems. Even if the underlying model remains proprietary, in an analogous fashion to proprietary software code, using open benchmarks we will have a way to evaluate any AI system for its alignment to and achievement of a given set of policy objectives.

OUR RECOMMENDATIONS FOR NIST

MLCommons is encouraged by NIST's focus on supporting the deployment of safe, secure and trustworthy AI through better measurement, evaluation and auditing capabilities. We believe there are four key areas where NIST should direct its focus:

1. **Developing an ecosystem approach to AI safety**. It will require a robust ecosystem of providers to develop and operate the full range of safety measures that are needed to mitigate risk in AI. These include a wide range from relatively process-focused auditing functions to highly technical red-teaming capabilities, as well as the creation of human-centered datasets that enable them, all of which will need to adapt at the pace of innovation and be embedded through a wide network of solutions providers to ensure access for all users and operators of AI systems. Standard safety benchmarks will help give this ecosystem direction and facilitate scale, but are only part of a wider comprehensive approach.
2. **Investing heavily in metrology: iterative calibration, data diversity, and ground truth validation**. While industry and research institutions can and should take the lead in development of safety benchmarks, the government is uniquely situated to provide the assurance of trust needed to "test the tests", or do metrology for this domain, effectively. In other

words, NIST can leverage the speed of academic and industry collaboration to build much of the testing technology, but NIST must invest heavily in support needed for calibrating and validating safety benchmarks from an end-user perspective and aligning the tests with our policies as well as human needs.

3. **Supporting research into testing technology and methodology**. Testing technology is fundamental to building AI that meets the wide range of societal safety objectives. Present testing technology is not yet sufficient to perform the kind of scalable, rigorous, repeatable testing that will be needed; for example, even for basic prompt-response interactions, testing technology cannot yet perform at a human level of rigor in a cost-effective way. As AI becomes more advanced, more and more complex interactions will demand that testing technology keeps pace, and that sound scientific methods provide the foundation for the technology. It will also be important to lower the cost of safely operating these systems, as doing so will ensure anyone and everyone is able to implement safety effectively. To the extent NIST can support ongoing research and development of testing technology, methodologies, and communities that develop both, that government support will help to balance the commercial incentives to invest in more complex AI capabilities.

4. **Invest in long-term sustainability of safety benchmarks**. The iterative nature of the safety benchmark development lifecycle mandates repeated calibration with ground truth measures. Over the long-term, safety benchmarks will need to be refreshed in an ongoing way and evolve with datasets and evaluations. Long after academic research interest in development of benchmarks has waned, there will be an ongoing need to evolve safety benchmarks so they maintain relevance to current technologies and societal challenges. Organizations that can support this sustainable approach to benchmarking will need to be stood up for the emerging AI industry to succeed. NIST can work with civil society and other funders to ensure that these organizations get created and resourced to the extent that will be required.