ML
Commons

ABOUT MLCOMMONS AND ITS INTEREST IN THIS REQUEST FOR COMMENT

This response to the National Telecommunications and Information Administration's Request for Comment on Dual Use Foundation Artificial Intelligence Models with Widely Available Model Weights is submitted on behalf of MLCommons®.

MLCommons is a non-profit consortium that aims to accelerate the benefits of machine learning and artificial intelligence (AI). 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 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 is developing 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, and will be critical to ensuring the safety of both open and proprietary AI systems.

MLCommons counts among members of its AI Safety Working Group organizations that have releasedl AI models without disclosing their weights, and organizations that have released AI models with open weights. As an organization, we believe there is a place in the market for both approaches to flourish.

[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/.

For the purposes of this comment, our focus is on the need to preserve the ability for developers to release widely available open model weights in order to ensure AI Safety for the ecosystem as a whole.

STANDARDIZED BENCHMARKS WILL BE A LINCHPIN OF THE AI SAFETY ECOSYSTEM

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 the problems of that class being incorrect.[4] 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.[5,6] 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.[7] 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.

Managing risk in modern AI is challenging in ways that dramatically differ from traditional software risk management. There are, however, 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 widely adopted with success. No automobile can be deemed perfectly safe in all possible circumstances, but there are general expectations that automobiles must 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

[4] 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.

[5] 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.

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

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

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.

Most critically for the purpose of the NTIA's inquiry into widely available open weights, standardized benchmarks will be a critical component for mitigating the risk of models both with and without widely available open weights. Benchmarks provide an equivalent approach to measuring the safety of AI model outputs, regardless of what is known about the model weights. But building reliable benchmarks will also depend on including models with widely available open weights in the development process.

AI SAFETY BENCHMARKING BENEFITS BY INCORPORATING MODELS WITH OPEN WEIGHTS

Models with open weights have played a central role in developing widely trusted benchmarks that have been used to evaluate and measure AI models, and in doing so have helped drive progress in AI. GLUE, BigBench, Harness, HELM and openCLIP Benchmark are all examples of widely used benchmarks that have helped researchers and developers measure progress in the development of AI models.[8,9,10,11] In order to create these and similar benchmarks, researchers benefited from the ability to freely reuse existing models as “reference” models which generated the benchmark results against which other models' results are compared. Critically, the use of reference models with open weights enable the broader AI community to trust the results of these benchmarks because they are able to interrogate the reference model's functioning. In a sense, using models with open weights as reference models has acted as a trust-building mechanism for evaluation of AI models generally.

Increasingly, we see safety benchmarks being operationalized with the use of a separate, “evaluator” model. When an AI model is tested against a benchmark, it is fed a set of test data that it uses to produce results, and these results are checked for safety. This safety check can be done by human review, but it can also be done with the use of an evaluator model. Human review will always be a part of evaluating AI safety, but it has limitations, specifically it doesn't scale in a very cost effective way. Evaluator AI models can be used to automatically review test results in a far more cost effective way.

It may prove ideal to use multiple evaluator models. Some of these will have open weights, building trust with the wider community that any changes in how results are evaluated can be understood by interrogating the open evaluator model. But relying only on models with open weights runs the risk that developers will over-fit to the evaluator model. On the other hand, if only a closed weights model is used in evaluation, results may change in unpredictable and unintelligible ways, undermining trust in the evaluation. To balance these competing incentives, we believe it can be ideal to use multiple evaluator models simultaneously, some with open weights and others with closed weights.

[8] "GLUE Benchmark," GLUE Benchmark, accessed March 26, 2024, https://gluebenchmark.com/.

[9] "Big Bench Datasets," Hugging Face Datasets, accessed March 26, 2024, https://huggingface.co/datasets/bigbench.

[10] "Large Language Model Evaluation," EleutherAI, accessed March 26, 2024, https://www.eleuther.ai/projects/large-language-model-evaluation.

[11] "CLIP Benchmark," GitHub, accessed March 26, 2024, https://github.com/LAION-AI/CLIP_benchmark.

The AI Safety Working Group led by MLCommons is now in the process of building operational safety benchmarks, leveraging the HELM framework from Stanford’s Center for Research on Foundation Models. Our goal in operationalizing safety benchmarks is to inform AI development by empowering users of AI systems and the regulators who oversee them with simple, easy to understand, standardized measures of AI system performance against a variety of safety goals. The most challenging aspect of doing this work is not technically defining the benchmarks, but is instead building trust in the process of how these benchmarks get defined. We believe that defining standard safety benchmarks will benefit tremendously from the inherent trust-building mechanism of incorporating AI models with open weights both as reference and evaluator models.

Models with widely available open weights allow the entire AI safety community – including auditors, regulators, civil society, users of AI systems, and developers of AI systems – to engage with the benchmark development process. Together with open data and model code, open weights enable the community to clearly and completely understand what a given safety benchmark is measuring, eliminating any confounding opacity around how a model was trained or optimized.