ABOUT MLCOMMONS AND ITS INTEREST IN THIS NOTICE OF INQUIRY

This response to the Copyright Office's Notice of Inquiry on Copyright and Artificial Intelligence 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 and startups that are actively developing and deploying artificial intelligence products for their customers. Critically, our founding membership included 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 machine learning (ML) 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. 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.

Most recently, we announced the formation of an AI Safety Working Group.[2] 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 language models (LMs), building on groundbreaking work done by researchers at Stanford University's Center for Research on Foundation Models and its Holistic Evaluation of Language Models.[3] 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 the voluntary commitments on safety, security, and trust that several tech companies made to the White House in July 2023[4] and NIST's AI Risk Management Framework[5].

[1] Machine learning is one of the key techniques through which AI systems are built.
[2] "MLCommons announces the formation of AI Safety Working Group", MLCommons, October 26, 2023, https://mlcommons.org/en/news/formation-ai-safety-working-group/ (accessed October 27, 2023).
[3] Percy Liang, et al., "Holistic Evaluation of Language Models," (2021), https://arxiv.org/abs/2211.09110 (accessed October 27, 2023).
[4] "Fact Sheet: Biden-Harris Administration Secures Voluntary Commitments from Leading AI Companies to Manage the Risks Posed by AI," White House, July 21, 2023, https://www.whitehouse.gov/briefing-room/statements-releases/2023/07/21/fact-sheet-biden-harris-administration-secures-voluntary-commitments-from-leading-artificial-intelligence-companies-to-manage-the-risks-posed-by-ai/ (accessed October 27, 2023).
[5] "AI Risk Management Framework," NIST, https://www.nist.gov/itl/ai-risk-management-framework (accessed October 27, 2023).

As the Copyright Office continues its study of copyright law and policy issues raised by generative AI, in particular the use of copyrighted works to train AI models, we hope you will take into consideration that approaches to measuring AI for safety risks are dependent on both training and testing data.

MANAGING AI RISK DEPENDS ON STANDARD MEASUREMENT

Effective evaluation and measurement of how AI systems perform across a range of attributes, including accuracy, safety, speed, bias, security and energy use, is how we will make progress protecting people's rights, advancing equity, and addressing the risk of harm with AI. Standardized metrics and benchmarks like those we will be developing in our AI Safety Working Group are crucial to effective evaluation and measurement of AI.

Modern AI is not merely code that is written to deterministically execute commands, instead it is fundamentally data-centric. AI should be understood conceptually as models that are trained on underlying datasets to produce predictive outputs given a range of input variables. The capabilities of the model are determined by the training data, and an assessment of the capabilities in a given context is determined by the dataset used to test the model. As a result, it becomes crucial in evaluating and characterizing AI systems to use standardized benchmarks based on comprehensive and challenging test datasets.

For example, consider evaluating whether autonomous vehicles are safe to drive in the snow. To compare two vehicles' capabilities to safely drive in the snow, both need to be characterized using the same underlying test dataset. If you use varied datasets, different vehicles would be effectively measured against different requirements. Further, it would be vital that the test dataset include the full range conditions, such as all different times of day and weather conditions, as well as difficult corner-cases such as whether there is debris on the road.

OPEN, HIGH-QUALITY DATASETS FOR TESTING AND TRAINING AI ARE CRUCIAL

Open, high-quality datasets also have a crucial role to play in advancing the measurement of AI systems and AI research more broadly. The ability to train on widely-available, published data without having to secure specific copyright licenses is crucial for several reasons. In our own study of nearly 2,000 research publications over the past five years, we've found widespread use and adoption of open data sets, suggesting that progress in AI is entirely dependent on availability of open training data.[6] First, datasets like Common Crawl provide access to a wide diversity of information, which can help address (though not fully resolve) concerns around bias in the data. Second, access to the data is free, which keeps barriers to entry low for organizations such as ours as well as many academic researchers. And third, because the

[6] Vijay Janapa Reddi, et al., "Data Engineering For Everyone" (2021), https://arxiv.org/abs/2102.11447 (accessed October 27, 2023).

dataset is widely available for anyone to build on, it allows for anyone (e.g., researchers, customers, vendors, regulators) to reproduce and compare their results to ours in a standardized way.

MLCommons is also actively working to build open data sets that others can use in training AI models, with the objective of providing data sets that can reduce bias in the eventual AI model, and improve its robustness. For example, our DollarStreet dataset for computer vision applications was manually built and labeled to ensure the thousands of images of household items were representative of a wide range of communities and socioeconomic households from around the world.[7] We have built the People's Speech Dataset, which is the world's largest English speech recognition corpus licensed for academic and commercial use, already cited by at least thirty-eight academic papers.[8,9] We have also built a Multilingual Spoken Words Corpus that represents spoken words in 50 languages collectively spoken by over 5 billion people; this was the first open dataset reflecting spoken words in 45 of those languages.[10]

A second way we've invested in robust datasets is through Dynabench, which is a platform that allows collection of human data dynamically with models in the loop.[11] People can be tasked with finding examples of data that fool a state-of-the-art AI model, or models can help people find interesting examples that would fool the model. We believe this approach allows rapid iteration of models by yielding data that can be used to further train even better state-of-the-art models. As part of Dynabench, we've also launched a DataPerf benchmark that evaluates the quality of training and test data, as well as the algorithms for constructing or optimizing datasets. By enabling construction and optimization of test sets, we believe platforms like Dynabench can play a critical role in evaluating future AI systems for bias and advancing equity.

[7] "Dollar Street," MLCommons, accessed July 7, 2023, https://mlcommons.org/en/dollar-street/

[8] "People's Speech," MLCommons, accessed July 7, 2023, https://mlcommons.org/en/peoples-speech/.

[9] https://scholar.google.com/citations?view_op=view_citation&hl=en&user=jLyty0sAAAAJ&citation_for_view=jLyty0sAAAAJ:_Qo2XoVZTnwC (accessed October 30, 2023)

[10] "Multilingual Spoken Words," MLCommons, accessed July 7, 2023, https://mlcommons.org/en/multilingual-spoken-words/.

[11] "Dynabench", accessed July 7, 2023, https://dynabench.org/.