ABOUT MLCOMMONS AND ITS INTEREST IN THIS REQUEST FOR INFORMATION

This response to the Australian Government ("The Consultation") on Supporting Responsible AI 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 50 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.

As the Department of Industry, Science and Resources considers appropriate regulatory and policy responses to support safe and responsible AI practices, 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 challenges identified in the June, 2023 Discussion paper on Safe and responsible AI in Australia. Our work is most directly relevant to implementing risk-based approaches in addressing potential AI risks, questions 14-20 of The Consultation. We are building the data and benchmarking infrastructure to facilitate managing risk in a way that we hope will be generally applicable regardless of which specific risk-based approaches that are adopted.

STANDARD BENCHMARKS FACILITATE MANAGING RISK IN AI

The Consultation asks about risk-based approaches to addressing potential AI risks and protecting people's rights, among other concerns. Making progress against these objectives is

[1] Machine learning is one of the key techniques through which AI systems are built.

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.

Standardized metrics and benchmarks are crucial to effective evaluation and measurement of AI, and the Australian Government should make them a key focus of policy efforts.

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 would 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.

In turn, the Australian Government should incorporate the role of standards and benchmarks in evaluating AI systems as part of its approach to mitigating potential risks in AI. Achieving many of the values and policy objectives outlined in the Consultation will require the community of researchers, companies and non-profit organizations working on AI to collaborate on shared infrastructure for these standards and benchmarks.

BENCHMARKING WILL PLAY A KEY ROLE IN RISK MANAGEMENT

The Australian Government should support collaboration with the existing community of practice focused on benchmarking, and MLCommons would be pleased to support such efforts and contribute its expertise.

MLCommons has pioneered a collaborative approach to building useful standards for evaluating AI. Through our development of MLPerf™, we have created a benchmarking suite to characterize the speed of machine learning systems in an open, representative, and reproducible manner. Upon release in 2018, MLPerf™ was quickly adopted by industry, and our community grew rapidly to over 50 partners and members. We've recently announced MLPerf™ Training v3.0, which was expanded to include testing of a large-language model, specifically GPT-3. Since our initial MLPerf™ release, 60 organizations have submitted testing results,

resulting in over 30,000 testing results submitted to MLPerf in a few short years. These numbers speak to the rapid industry adoption we've seen of these shared benchmarks.

Leveraging our experience creating commercially usable datasets, MLCommons is building infrastructure that will enable us to effectively assess other key attributes of AI systems, and in doing so to facilitate management of risks inherent to AI. For example, we have already begun to expand our expertise in developing benchmarks to new areas beyond general hardware performance, including:

- **Automotive**: Last month we announced a partnership with AVCC to develop the automotive industry's first open-source Automotive Benchmark Suite for use by OEMs and automotive suppliers using AI/ML Deep Neural Network technology.[2] Though this work is very early, we anticipate needing to tackle a broader range of benchmarking needs than simply addressing how a given hardware chip performs.
- **Science**: A MLCommons Science working group has been developing benchmarks specific to scientific applications for several years. This spring, the group released four benchmarks focused on cloud masking, geophysics, scanning transmission electron microscopy, and cancer distributed learning environments.[3]
- **Healthcare**: We have a working group focused on robust evaluation medical applications using a federated approach to preserve patient privacy.

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 objectives noted in the Consultation. For example, to help mitigate bias, developers of AI systems benefit from access to diverse, representative datasets. Furthermore, as noted above, standardized datasets are important for testing AI systems; to see if an AI system has biases related to language, for instance, one must have test data that is representative. We believe objectives such as mitigating algorithmic bias and discrimination are heavily dependent on construction of datasets that accurately and comprehensively reflect the world.

MLCommons is actively working to address these challenges. 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.[4] We have built the People's Speech Dataset, which is the world's largest English speech recognition corpus licensed for academic and

---

[2] "Company Launches New Product Line to Meet Growing Demand," PRWeb, May 10, 2023, accessed July 7, 2023, https://www.prweb.com/releases/2023/5/prweb19341862.htm.

[3] Jeyan Thiyagalingam et al., "AI Benchmarking for Science: Efforts from the MLCommons Science Working Group," Accessed July 7, 2023 https://www.researchgate.net/publication/366826798_AI_Benchmarking_for_Science_Efforts_from_the_MLCommons_Science_Working_Group.

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

commercial use.[5] 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.[6]

A second way we've invested in robust datasets is through creation of Dynabench, which is a platform that allows collection of human data dynamically with models in the loop.[7] 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.

IMPORTANCE OF GLOBAL INDUSTRY COORDINATION

MLCommons counts as its members organizations 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.

[5] "People's Speech," MLCommons, accessed July 7, 2023, https://mlcommons.org/en/peoples-speech/.
[6] "Multilingual Spoken Words," MLCommons, accessed July 7, 2023, https://mlcommons.org/en/multilingual-spoken-words/.
[7] "Dynabench", accessed July 7, 2023, https://dynabench.org/.