FERMILAB-POSTER-25-0223-CSAID

MLCommons Science Benchmarks

Ben Hawks[1], Nhan Tran[1], Marco Colombo[2], Gregor von Laszewski[3], Reece Shiraishi[4], Anjay Krishnan[5]

Author Affiliations:
1 Fermi National Accelerator Laboratory, USA
2 Discovery Partners Institute, USA
3 University of Virginia, USA
4 Cornell University, USA
5 University of Illinois Urbana-Champaign, USA

Motivation

Benchmarks are a cornerstone of modern machine learning practice, providing standardized evaluations that enable reproducibility, comparison, and scientific progress. Yet, as AI systems become increasingly dynamic, traditional static benchmarking approaches are losing their relevance. Models rapidly evolve in architecture, scale, and capability; datasets shift; and deployment contexts continuously change, creating a moving target for evaluation. We aim to curate a collection of modern, high quality scientific benchmarks across a variety of domains, computing motifs, and workloads which can be used to accurately evaluate a models and computing platforms on scientific tasks.

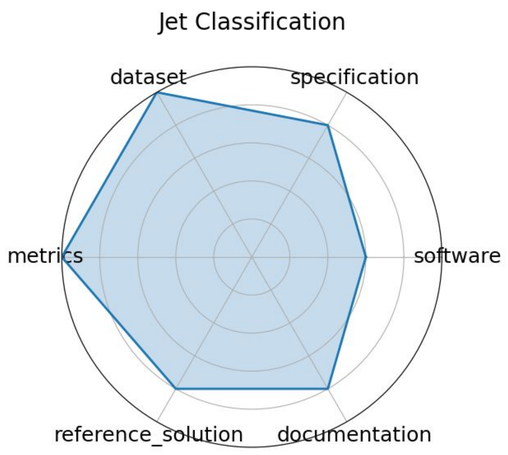

Figure 1: An example radar plot detailing the evaluation criteria for a given benchmark

Figure 2: A preliminary version of the benchmark endorsement badge, given to particularly high quality benchmarks

Benchmark Quality Evaluation System

When curating our benchmark collection, we have a need to be able to distinguish high quality benchmarks within the collection. In order to meet this need, we have developed a system to evaluate a given benchmark across six categories. Each category gets a score from 0-5, with 0 being the lowest and 5 being the highest possible. Additionally, if a benchmark reaches an average score of at least 4.5 across all categories, we award it an “endorsement badge”, such as Fig. 2 identifying it as a particularly high quality benchmark within the collection. The 6 categories we evaluate are as follows:

- Problem Specification & Constraints
- Dataset
- Performance Metric(s)
- Reference Solution
- Documentation
- Reproducible Environment (Software)

Scientific Benchmark Definition

A **benchmark** is a carefully defined standardized version of a scientific application that is used for making quantifiable comparisons of solutions.

A scientific benchmark must contain the following:

- Problem Specification & Constraints
- Dataset
- Performance Metric(s)
- Reference Solution
- Documentation & Reproducible Protocol

Each aspect of a given benchmark must be clearly and completely documented, and must contain all information needed to use, reproduce, evaluate a solution, and (if applicable) submit results.

https://mlcommons-science.github.io/benchmark/

Figure 3: A screenshot of the collection list and an example entry in the MLCommons Science Benchmark collection website

Benchmark Collection

We have began to curate a collection of scientific benchmarks, evaluating each entry’s quality with our 6 category evaluation system. The collection is publicly available at the URL in Fig. 3, and community contributions via pull request are welcome and highly encouraged.

Fermi National Accelerator Laboratory

This manuscript has been authored by FermiForward Discovery Group, LLC under Contract No. 89243024CSC000002 with the U.S. Department of Energy, Office of Science, Office of High Energy Physics.