frontiers

1

Supplementary Material

Gregor von Laszewski,[1] **Wesley Brewer,**[2] **Jeyan Thiyagalingam,**[3] **Juri Papay,**[3] **Armstrong Foundjem,**[4] **Piotr Luszczek,**[5] **Murali Emani,**[6] **Shirley V. Moore,**[7] **Vijay Janapa Reddi,**[8] **Matthew D. Sinclair,**[9] **Sebastian Lobentanzer,**[10] **Sujata Goswami,**[11] **Benjamin Hawks,**[12] **Marco Colombo,**[13] **Nhan Tran,**[12] **Christine R. Kirkpatrick,**[14] **Abdulkareem Alsudais,**[15] **Gregg Barrett,**[16] **Tianhao Li,**[17] **Kirsten Morehouse,**[18] **Shivaram Venkataraman,**[9] **Rutwik Jain,**[9] **Kartik Mathur,**[20] **Victor Lu,**[21] **Tejinder Singh,**[22] **Khojasteh Z. Mirza,**[23] **Kongtao Chen,**[24] **Sasidhar Kunapuli,**[25] **Gavin Farrell,**[26] **Renato Umeton,**[27] **Geoffrey C. Fox**[1]

[1] *Biocomplexity Institute, University of Virginia, Charlottesville, VA, USA*
[2] *Oak Ridge National Laboratory, Oak Ridge, TN, USA*
[3] *Rutherford Appleton Laboratory, STFC, Harwell Campus, UK*
[4] *DEEL, Polytechnique Montreal, Montreal, Canada*
[5] *LLSC, MIT Lincoln Laboratory, Lexington, MA, USA*
[6] *Argonne National Laboratory, Lemont, IL, USA*
[7] *Computer Science Department, UTEP, El Paso, TX, USA*
[8] *Harvard University, Boston, MA, USA*
[9] *Computer Sciences Department, Univ. of Wisconsin–Madison, Madison, WI, USA*
[10] *Helmholtz Center Munich, Munich, Germany*
[11] *ALS, LBNL, Berkeley, CA, USA*
[12] *Fermilab, Batavia, IL, USA*
[13] *Discovery Partners Institute, UIUC, Chicago, IL, USA*
[14] *SDSC, UC San Diego, San Diego, CA, USA*
[15] *Prince Sattam bin Abdulaziz University, Saudi Arabia*
[16] *Cirrus AI, Johannesburg, South Africa*
[17] *Duke University, Durham, NC, USA*
[18] *Harvard University, Cambridge, MA, USA*
[20] *Microsoft, Vancouver, BC, Canada*
[21] *Independent Researcher, Tampa, FL, USA*
[22] *Office of the CTO, Dell Technologies, Santa Clara, CA, USA*
[23] *Cornell Tech, Cornell University, New York, NY, USA*
[24] *Google, Mountain View, CA, USA*
[25] *Independent Researcher, San Jose, CA, USA*
[26] *University of Padua, Padua, Italy*
[27] *St. Jude Children's Research Hospital, Memphis, TN*

Correspondence*:
Gregor von Laszewski
laszewski@gmail.com

1 SUPPLEMENTARY DATA

Table 2 MLCommons Benchmarks

Benchmark Name	Model	Task	Application Domain / Use Case	Model Type / Architecture	Metrics / KPIs	Hardware	Notes / Description
MLPerf Inference: Datacenter							
deepseek-r1	DeepSeek R1 (671B params)	Reasoning / Code Generation	Knowledge & Reasoning, Complex Problem Solving, Step-by-Step Planning	Large Language Model (LLM), Reasoning LLM, High context/output length (up to 20K tokens)	Accuracy: Exact Match, Code Evaluation; Latency: TTFT (Time to First Token), TPOT (Time Per Output Token)	Data Center GPUs (NVIDIA H100/H200) with massive VRAM, optimized for 671B parameters.	The model's large output length emphasizes its use in complex reasoning chains. Requires powerful systems (e.g., multiple H100 GPUs).
dlrm-v2-99	DLRM-v2	Recommendation	Personalized product/content recommendation (e.g., e-commerce, social media feeds)	Deep Learning Recommendation Model (DLRM), Sparse/Dense Architecture	Throughput: Queries Per Second (QPS); Latency: 99th Percentile Latency	Data Center CPUs and GPUs (NVIDIA B200/GB200/B300), prioritizing high I/O and memory bandwidth for massive embedding tables.	Tests high-throughput, low-latency deployment for online services with a 99% latency constraint.
dlrm-v2-99.9	DLRM-v2	Recommendation	Personalized product/content recommendation (e-commerce, social media feeds)	Deep Learning Recommendation Model (DLRM), Sparse/Dense Architecture	Throughput: Queries Per Second (QPS); Latency: 99.9th Percentile Latency	Data Center CPUs and GPUs (NVIDIA H200), often using higher precision to ensure quality target is met.	Tests high-throughput, very low-latency deployment for critical online services with a strict 99.9% latency constraint.
llama2-70b-99	Llama 2 (70B params)	Large Language Model (LLM) Inference	General text generation, chat, summarization, and understanding	LLM, Transformer-based	Throughput: Tokens Per Second (TPS); Latency: TTFT, TPOT (99th Percentile)	Data Center GPUs (e.g., AMD MI300X/MI325X, NVIDIA B200/GB200/H100/H200/L40S, Intel Gaudi, MS-Intel Arc Pro B60) in multi-GPU configurations, focused on high throughput and low latency.	Represents a larger LLM workload, measuring performance under a 99% latency constraint.

Continued on next page

Table 2 MLCommons Benchmarks (Cont.)

Benchmark Name	Model	Task	Application Domain / Use Case	Model Type / Architecture	Metrics / KPIs	Hardware	Notes / Description
llama2-70b-99.9	Llama 2 (70B params)	Large Language Model (LLM) Inference	General text generation, chat, summarization, and understanding	LLM, Transformer-based	Throughput: Tokens Per Second (TPS); Latency: TTFT, TPOT (99.9th Percentile)	Data Center GPUs (AMD MI300X/MI325X, NVIDIA B200/GB200/H100/H200/L40S, Intel Gaudi, MS-Intel Arc Pro B60), often testing the limits of precision vs. speed trade-offs.	Represents a larger LLM workload, measuring performance under a stricter 99.9% latency constraint.
llama3.1-8b-datacenter	Llama 3.1 (8B params)	Summarization / Text Generation	Low-cost, high-volume LLM services, interactive code assistants	LLM, Transformer-based	Accuracy: ROUGE metrics (1, 2, L); Latency: TTFT ≤2s, TPOT ≤100ms (Server)	Single-node systems or smaller GPU clusters, used to lower the entry barrier for the MLPerf Training suite.	Benchmarks a smaller LLM for efficient deployment in both Data Center and Edge scenarios.
llama3.1-405b	Llama 3.1 (405B params)	Large Language Model (LLM) Inference	Generative AI, high-capability models	LLM, Transformer-based	Throughput: Output Tokens per second; Latency: TTFT, TPOT	Large-scale AI Clusters and Supercomputers (requires hundreds of GPUs (NVIDIA B200/GB200/GB300/H100/H200) with high-speed interconnects).	One of the largest LLMs in the suite, demonstrating the need for advanced parallelism (tensor, pipeline) on high-end systems (e.g., NVIDIA H200).
mixtral-8x7b	Mixtral (46.7B total params)	Large Language Model (LLM) Inference	generative AI, multilingual tasks	Mixture-of-Experts (MoE) LLM (activates ≈13B params per token)	Throughput: Tokens Per Second (TPS); Latency	Data Center GPUs (AMD MI300X/MI325X, NVIDIA H200/RTX PRO 6000), optimizing MoE architecture for low active compute per token.	Showcases the efficiency of MoE architecture, offering high quality with lower active compute cost than dense models.

Continued on next page

Table 2 MLCommons Benchmarks (Cont.)

Benchmark Name	Model	Task	Application Domain / Use Case	Model Type / Architecture	Metrics / KPIs	Hardware	Notes / Description
retinanet	Retinanet-ResNext50	Object Detection	Identifying and localizing objects in images	Object Detection Model, often with ResNext backbone and FPN	Accuracy: mAP (mean Average Precision); Throughput: Samples Per Second	Data Center and Edge GPUs (NVIDIA GeForce RTX 4090/H200/L4-PCIe/L40S), measuring both throughput and latency under a 100ms constraint.	A standard computer vision benchmark using the OpenImages dataset.
rgat	Relational Graph Attention Network	Node Classification	Graph data analysis, social network processing, knowledge graphs	Graph Neural Network (GNN), Graph Attention Network (GAT) variant	Accuracy (on node classification); Throughput: Samples Per Second	Data Center GPUs (NVIDIA B200), specifically testing performance on irregular, graph-structured data.	Addresses graph-structured data and multi-relational graphs, testing system efficiency for complex graph workloads.
stable-diffusion-xl	Stable Diffusion XL (SDXL)	Text-to-Image Generation	Generative AI for creating high-quality images from text prompts	Diffusion Model (Latent Diffusion)	Throughput: Images Per Second; Latency	Data Center and Professional GPUs (AMD MI325X,NVIDIA B200/H100/H200/L4-PCI/L40S/NVIDIA RTX PRO 6000), focusing on the speed of image generation (samples/second).	Represents the Text-to-Image Generative AI domain, measuring the speed of image synthesis.

Continued on next page

Table 2 MLCommons Benchmarks (Cont.)

Benchmark Name	Model	Task	Application Domain / Use Case	Model Type / Architecture	Metrics / KPIs	Hardware	Notes / Description
whisper	Whisper-Large-V3	Automatic Speech Recognition (ASR)	Converting spoken audio to text	Encoder-Decoder Transformer, Speech-to-Text Model	Accuracy: WER (Word Error Rate), Word Accuracy (Acc); Latency	Data Center GPUs (NVIDIA B200/GB200/GeForce RTX 4090/H100/H200/L4-PCIe/L40S), measuring performance on a complex sequence-to-sequence model for speech.	An ASR benchmark on multilingual audio, measuring both encoder (audio feature) and decoder (token generation) performance.
MLPerf HPC							
CosmoFlow	CosmoFlow 3D CNN	Regression	Astrophysics, Cosmology (predicting properties of the universe from simulation data)	3D Convolutional Neural Network (3D CNN)	Time to Quality (TTQ) (e.g., Time to reach validation MAE ≤ 0.124)	Supercomputers & Large HPC Clusters (e.g., Fugaku, Perlmutter). Stresses distributed training, 3D data handling, and fast data I/O for massive volumetric datasets ($\approx$ 5 TB) GPUs used for running this benchmark: NVIDIA A100/V100.	Uses massive 3D volumetric data ($\approx$ 5.1 TB). Stresses memory bandwidth and interconnect.

Continued on next page

Table 2 MLCommons Benchmarks (Cont.)

Benchmark Name	Model	Task	Application Domain / Use Case	Model Type / Architecture	Metrics / KPIs	Hardware	Notes / Description
DeepCAM	DeepCAM Encoder-Decoder	Semantic Segmentation	Climate Science, Extreme Weather Prediction (identifying atmospheric rivers, tropical cyclones)	Convolutional Encoder-Decoder (e.g., U-Net or DeepLab-like)	Time to Quality (TTQ) (e.g., Time to reach validation IoU ≥ 0.82)	Supercomputers & Large HPC Clusters. Stresses large-scale image processing, high-dimensional data (many channels), and efficient communication on systems with thousands of GPUs (A100/P100/V100).	Trained on massive, high-resolution 2D image data ($\approx$ 8.8 TB). Stresses I/O and communication efficiency.
OpenCatalyst	DimeNet++	Regression	Computational Chemistry, Materials Science (discovering new catalysts for energy storage)	Graph Neural Network (GNN)	Time to Quality (TTQ) (Time to reach target energy/force prediction error)	Supercomputers & Large HPC Clusters. Stresses performance on graph-structured data (atomic systems) and complex GNN operations that require high GPU utilization. GPUs used for running this benchmark: NVIDIA A100/P100/V100.	Models atoms and bonds as a graph structure. Benchmarks complex, irregular GNN workloads at scale.
MLPerf Training							
BERT (Bidirectional Encoder Representations from Transformers)	NLP - Question Answering	General NLP, Text Understanding	Transformer (Encoder)	Time to Quality (TTQ) (F1 Score on SQuAD)	Data Center GPUs, Accelerators	CPU, Single GPU (e.g., NVIDIA A100/H100), or moderate clusters.	A foundational benchmark for Natural Language Processing tasks.

Continued on next page

Table 2 MLCommons Benchmarks (Cont.)

Benchmark Name	Model	Task	Application Domain / Use Case	Model Type / Architecture	Metrics / KPIs	Hardware	Notes / Description
DLRM-dcnv2 (Deep Learning Recommendation Model - DCNv2)	Recommendation Systems	E-commerce, Content Streaming, Personalized Ads	Deep Learning Recommendation Model w/ DCNv2	Time to Quality (TTQ) (AUC on Criteo 4TB)	Data Center GPUs, Specialized Accelerators	Large-scale GPU clusters with high-speed interconnects (e.g., InfiniBand) for distributed training. This benchmark was running on GPUs: NVIDIA B300/B200/GB200/H200/H100/H200.	Stresses memory bandwidth and communication for massive embedding tables.
llama2-70b-lora	LLM Fine-Tuning	Customizing LLMs for specific enterprise tasks	Transformer with LoRA	Time to Quality (TTQ) (ROUGE Score)	Multi-GPU servers, Mid-size GPU clusters	High-end Multi-GPU servers or small clusters (e.g., systems with AMD MI300X/MI325X/MI350X/MI355X, NVIDIA B200/B300/H100/H200).	Measures the efficiency of Low-Rank Adaptation (LoRA) on a ≈ 70B parameter model.
llama3.1-405b	LLM Pretraining	Generative AI, Foundational Model Development	Transformer-based LLM (≈ 405B params)	Time to Quality (TTQ) (Log Perplexity)	Large-scale, Multi-node GPU clusters	Single Node or small GPU systems (e.g., a few GPUs per node) to keep the benchmark accessible.Benchmark running on GPUs: NVIDIA B200/B300/H200.	The largest, most compute-intensive benchmark for pretraining state-of-the-art LLMs.
RetinaNet	Object Detection	Autonomous Vehicles, Surveillance, Image Analysis	One-stage Object Detector (ResNet, FPN)	Time to Quality (TTQ) (mAP on COCO)	Data Center GPUs, Cloud Instances	Single or multi-GPU systems (NVIDIA B200/H200/RTX Pro 6000), often used in both Datacenter and Edge devices for inference.	Measures performance for a core computer vision task: localizing and classifying objects.

Continued on next page

Table 2 MLCommons Benchmarks (Cont.)

Benchmark Name	Model	Task	Application Domain / Use Case	Model Type / Architecture	Metrics / KPIs	Hardware	Notes / Description
RGAT (Relational Graph Attention Network)	GNN - Node Classification	Drug Discovery, Social Network Analysis, Fraud Detection	Relational Graph Attention Network (R-GAT)	Time to Quality (TTQ) (Accuracy on IGBH)	Systems optimized for high-bandwidth interconnects	GPU-based systems (NVIDIA B200/B300/H100), optimized for workloads with complex, sparse data structures like graphs.	Focuses on the irregular memory access and communication patterns of Graph Neural Networks.
Flux1 (stable-diffusion)	Text-to-Image Generation	Generative AI, Digital Art, Content Creation	Latent Diffusion Model (U-Net, Transformer)	Time to Quality (TTQ) (FID and CLIP Scores)	Multi-GPU servers, Cloud Instances	High-performance Single or Multi-GPU systems (especially for fast inference or training). This benchmark was running on: NVIDIA B200/GB200/GB300.	Benchmarks the training of a major generative model in the AI industry.
MLPerf Inference: Edge							
3D U-Net (99%)	3D U-Net	Medical Image Segmentation	Healthcare, Volumetric Imaging (e.g., MRI/CT)	3D Convolutional Encoder-Decoder CNN	Accuracy (Dice Score), Latency, Throughput (QPS)	Data Center GPUs (e.g., NVIDIA A100/H100), high-performance computing (HPC) systems, specialized accelerators.	99% of reference accuracy target. Typically runs in Offline scenario for batch processing of medical scans.
3D U-Net (99.9%)	3D U-Net	Medical Image Segmentation	Healthcare, High-Fidelity Imaging	3D Convolutional Encoder-Decoder CNN	Accuracy (Dice Score), Latency, Throughput (QPS)	Data Center GPUs (e.g., NVIDIA A100/H100), high-performance computing (HPC) systems, specialized accelerators.	99% of reference accuracy target. Represents a stricter quality constraint, often requiring higher-precision compute (e.g., FP16 vs. INT8).

Continued on next page

Table 2 MLCommons Benchmarks (Cont.)

Benchmark Name	Model	Task	Application Domain / Use Case	Model Type / Architecture	Metrics / KPIs	Hardware	Notes / Description
llama3.1-8b-edge	Llama 3.1 (8B params)	Text Generation / Summarization	Edge AI, On-device LLMs, Interactive Assistants	Quantized Transformer (Decoder-only LLM)	Tokens Per Second (TPS), Latency (TTFT, TPOT), Power	Edge devices, mobile SoCs (System-on-Chips), smaller GPUs (MS-Intel Arc Pro B60), high-end CPUs.	Benchmarks a modern, smaller LLM variant optimized for performance and low-latency on resource-constrained Edge devices.
resnet	ResNet50-v1.5	Image Classification	Vision, Quality Control, Surveillance	CNN (Residual Network)	Accuracy (Top-1), Latency, Throughput (QPS)	Data Center GPUs (NVIDIA GeForce RTX 4090/RTX-2000E), Edge devices, Mobile SoCs, CPUs, specialized accelerators.	The foundational computer vision benchmark, often used as a baseline for measuring performance and efficiency across all MLPerf tiers.
retinanet	RetinaNet-ResNext50	Object Detection	Autonomous Vehicles, Advanced Security Systems	One-stage Object Detection (often with FPN)	Accuracy (mAP - mean Average Precision), Latency, Throughput (SPS)	Data Center GPUs (NVIDIA GeForce RTX 4090/4000/2000E), Edge devices, specialized detection accelerators.	Measures the system's ability to find and localize multiple objects in images. Uses the OpenImages dataset.
stable-diffusion-xl	Stable Diffusion XL (SDXL)	Text-to-Image Generation	Generative AI, Digital Content Creation	Diffusion Model (Latent Diffusion with U-Net)	Images Per Second, Latency (Time to generate an image)	Data Center GPUs (e.g., NVIDIA H100/H200, AMD MI300 series), powerful consumer-grade GPUs.	Represents the high-compute generative AI domain. Measures the speed of synthesizing high-resolution images from text prompts.

Continued on next page

Table 2 MLCommons Benchmarks (Cont.)

Benchmark Name	Model	Task	Application Domain / Use Case	Model Type / Architecture	Metrics / KPIs	Hardware	Notes / Description
whisper	Whisper-Large-V3	Automatic Speech Recognition (ASR)	Speech-to-Text Services, Live Transcription	Encoder-Decoder Transformer	Accuracy (WER - Word Error Rate, Word Acc), Tokens Per Second	Data Center GPUs (NVIDIA GeForce RTX 4090), Edge/Client devices for real-time transcription.	A modern, high-accuracy ASR benchmark, using a Transformer architecture that handles both audio encoding and token generation.
MLPerf Inference: Mobile							
MLPerf Mobile/Edge	MobileNetV4-Conv-L	Image Classification, Object Detection	Edge/Mobile AI, low-latency on-device vision tasks.	CNN / MobileNet Family (V4)	Latency (ms), Throughput (Inferences/sec), Top-1/Top-5 Accuracy, Average Precision (AP).	Mobile SoCs, Specialized Mobile Accelerators (e.g., Apple Neural Engine, Edge TPUs, dedicated DSPs)	The largest convolutional-only variant of MobileNetV4. Optimized via Neural Architecture Search (NAS) for better latency-accuracy trade-offs on mobile and embedded hardware.
MLPerf Mobile/Edge	Mobile SSD Variants	Object Detection	Edge/Mobile AI, real-time detection on resource-constrained devices.	Single Shot Detector (SSD) + Mobile Backbone	Average Precision (AP) (e.g., COCO AP), Latency (ms), FPS.	Mobile SoCs (CPU, GPU, NPU/DSP), Edge AI Accelerators	Refers to models like SSD-MobileNet V1/V2/V3 which are standard mobile benchmarks.
MLPerf Edge	SSD-MobileNet	Object Detection (Small)	Edge/Mobile AI, detection for systems with tight latency/power budgets.	Single Shot Detector (SSD) + MobileNet Backbone	Average Precision (AP), Latency (ms).	Mobile SoCs (CPU, GPU, NPU/DSP), Edge AI Accelerators	A specific variant that is an original, primary benchmark for MLPerf Inference: Edge.

Continued on next page

Table 2 MLCommons Benchmarks (Cont.)

Benchmark Name	Model	Task	Application Domain / Use Case	Model Type / Architecture	Metrics / KPIs	Hardware	Notes / Description
MLPerf Mobile/Edge	MobileNet V1–V4	Image Classification, Feature Extractor	Efficient Vision Models, low-power and low-latency inference.	CNN (V1: Depthwise Separable Convs, V2: Inverted Residuals, V3: Squeeze-and-Excitation, V4: UIB/Mobile MQA)	MACs/FLOPs, Latency (ms).	Mobile SoCs (CPU, GPU, NPU/DSP, e.g., Qualcomm Snapdragon, Apple A-series), Microcontrollers (MCUs), Edge AI Accelerators (e.g., Google Edge TPU)	A progression of architectures from Google, all focused on minimal computational cost while maintaining high accuracy, crucial for all MLPerf Edge divisions.
MLPerf Mobile	MobileNet V4	Image Classification, Object Detection	Universally Efficient AI, aiming for state-of-the-art accuracy-latency trade-offs.	Hybrid (Convolutional + Attention - Mobile MQA)	Latency (ms)	Mobile SoCs (CPU, GPU, NPU/DSP, e.g., Qualcomm Snapdragon)	The latest generation, featuring the Universal Inverted Bottleneck (UIB) and Mobile MQA.
MLPerf Mobile	MOSAIC	Image Segmentation	Mobile Image Segmentation, on-device image processing.	U-Net variant with a MobileNet-style backbone.	Mean Intersection over Union (mIoU), Latency (ms).	Mobile SoCs (CPU, GPU, NPU)	A common model used for segmentation tasks in the MLPerf Mobile suite.
MLPerf Mobile	MobileDETs	Object Detection	Edge/Mobile AI, high-speed detection for mobile chips.	Model Family derived from Neural Architecture Search (NAS)	Average Precision (AP), Latency (ms).	Mobile SoCs (NPU/DSP emphasized), Edge AI Accelerators	A family of detectors specifically optimized for latency on mobile SoCs.
MLPerf TinyMobile	BERT-Tiny DistilBERT	Natural Language Processing (NLP) Tasks (e.g., Q&A)	MobileEdge NLP, faster, smaller language understanding on local devices.	Transformer Distillation Models	Latency (ms), F1 Score (SQuAD), GLUE Score.	CPUs, GPUs, Edge AI Accelerators, Mobile SoCs (optimized for low-latency)	Smaller, compressed versions of BERT achieved through knowledge distillation for resource-constrained environments.

Continued on next page

Table 2 MLCommons Benchmarks (Cont.)

Benchmark Name	Model	Task	Application Domain / Use Case	Model Type / Architecture	Metrics / KPIs	Hardware	Notes / Description
MLPerf Mobile	Mobile-BERT	Natural Language Processing (NLP) Tasks	Edge/Mobile NLP, task-agnostic BERT for resource-limited devices.	Compressed Transformer (Bottleneck structures, Knowledge Distillation)	Latency (ms), F1 Score (SQuAD), GLUE Score.	CPUs, GPUs, Edge AI Accelerators, Mobile SoCs (optimized for low-latency)	Achieves competitive results to BERT-Base with much higher speed and smaller size.
MLPerf Mobile	EDSR F32B5	Image Super-Resolution (SR)	Image Enhancement, upscaling low-resolution images for improved quality.	Enhanced Deep Super-Resolution (EDSR) Network	Latency (ms), PSNR, SSIM.	GPUs, Custom Hardware/FPGAs, specialized ISP (Image Signal Processor) components.	A common, high-quality reference model for measuring performance on image enhancementquality tasks.
MLPerf Mobile	Stable Diffusion	Text-to-Image Generation	Generative AI, creating high-resolution images from text prompts.	Latent Diffusion Model (LDM) (U-Net, VAE, CLIP Text Encoder)	Images/Query Per Second (Throughput), Latency (Time-to-Image), FID/CLIP Scores.	High-end GPUs (e.g., NVIDIA A100/H100, RTX series), high-power Workstations and Data Center Accelerators.	A critical benchmark for measuring performance on large, complex generative workloads.
MLPerf Inference: Tiny							
MLPerf Tiny v 0.5	Keyword Spotting Model	Audio Classification	TinyML/MCU, always-on voice assistant, device wake-word detection.	Small CNN (e.g., DS-CNN) or RNN.	Latency (ms), Energy (Joules), Area Under the ROC Curve (AUC).	Microcontrollers (MCUs) (e.g., Arm Cortex-M4M7), Digital Signal Processors (DSPs), Tiny Neural Network Accelerators.	Detects a specific word (e.g., "Hey Google") from a stream of audio, running on a highly constrained power budget.
MLPerf Tiny v 0.5	Visual Wake Words (VWW) Model	Image Classification (Binary)	TinyML/MCU, low-power sensing, person detection, motion-activated cameras.	Small CNN (e.g., MobileNet V1/V2 variant).	Latency (ms), Energy (Joules), AUC.	MCUs, low-power vision processors, small-scale embedded systems.	Determines if a person is present in the image (person/not-person). Much simpler and smaller than general ImageNet classification.

Continued on next page

Table 2 MLCommons Benchmarks (Cont.)

Benchmark Name	Model	Task	Application Domain / Use Case	Model Type / Architecture	Metrics / KPIs	Hardware	Notes / Description
MLPerf Tiny v 0.5	Image Classification Model	Image Classification (Multi-class)	TinyML/MCU, general object recognition on ultra-low-power sensors.	Very small CNN (e.g., ResNet-8 or Micro-CNN).	Latency (ms), Energy (Joules), Top-1 Accuracy (e.g., on CIFAR-10).	MCUs with limited RAM and Flash storage.	A more complex classification task than VWW, but still constrained to a very small model size.
MLPerf Tiny v 0.5	Anomaly Detection (AD) Model	Time Series Anomaly Detection	TinyML/MCU, industrial predictive maintenance, system health monitoring.	Small Autoencoder or similar lightweight model.	Latency (ms), Energy (Joules), AUC.	MCUs, industrial IoT sensors, devices monitoring vibration or sound.	Learns a baseline of normal sensor data (e.g., machine vibrations) and flags deviations as anomalies.
MLPerf Client							
MLPerf Client	Llama 2 7B Chat	Code analysis, Content generation, Creative writing, Summarization (various lengths).	General-purpose AI, Dialogue/Chatbots, Client-side LLM inference on PCs.	Transformer, Decoder-Only, Instruction-Tuned (SFT + RLHF), 7 Billion parameters.	Time-to-First Token (TTFT), Tokens/Second (Throughput).	Client GPUs (e.g., AMD Radeon, Intel Arc), Integrated NPUs (e.g., Intel Core Ultra, AMD Ryzen AI), Data Center GPUs (e.g., NVIDIA A100/H100) for server-side inference.	A foundational model in the benchmark for measuring core client-side LLM performance.
MLPerf Client	Llama 3.1 8B Instruct (8B parameters)	Generative AI workloads: Code analysis, Content generation, Creative writing, Summarization.	General-purpose AI, Instruction Following, Client-side LLM inference on PCs.	Transformer, Decoder-Only, Instruction-Tuned, 8 Billion parameters.	Time-to-First Token (TTFT) (Latency), Tokens/Second (Throughput).	Client PCs and Data Center/Cloud-based GPUs (optimized for both low-latency "Time to First Token" and high-throughput "Tokens Per Second").	An updated and highly capable open-weight model, demonstrating improved performance and alignment over Llama 2.

Continued on next page

Table 2 MLCommons Benchmarks (Cont.)

Benchmark Name	Model	Task	Application Domain / Use Case	Model Type / Architecture	Metrics / KPIs	Hardware	Notes / Description
MLPerf Client	Phi 3.5 Mini Instruct	Reasoning (Math, Code, Logic), Long Context Query & Summarization (up to 128K tokens).	Memory/Compute Constrained Environments, Low-Latency Applications, On-device deployment (AI PCs, mobile).	Dense Decoder-Only Transformer, Instruction-Tuned, 3.8 Billion parameters.	Time-to-First Token (TTFT) (Latency), Tokens/Second (Throughput).	Client GPUs, NPUs, and potentially high-end mobile/edge processors (optimized for on-device deployment).	A highly efficient and lightweight model optimized for speed and strong reasoning despite its small size.
MLPerf Client	Phi 4 Reasoning 14B	Complex Reasoning (multi-step math, scientific, coding, planning), Generating detailed chain-of-thought traces.	Agentic applications, High-accuracy problem-solving, Applications requiring explainability.	Dense Decoder-Only Transformer, Reasoning-Focused SFT (and possible RLHF for Plus variant), 14 Billion parameters.	Time-to-First Token (TTFT) (Latency), Tokens/Second (Throughput), Accuracy on reasoning tasks.	High-performance Client PCs (Workstations) and Data Center GPUs (due to its larger size and focus on complex, token-intensive reasoning).	Included as an experimental model in the benchmark, specifically designed to emphasize logical and complex problem-solving.
MLPerf Storage							
MLPerf Storage	ResNet-50	I/O Workload for Image Classification Training	General-purpose computer vision, low-latency image processing.	Convolutional Neural Network (CNN)	Max Supported Accelerators, Aggregate Throughput (MiB/s), Accelerator Utilization ($\geq$ 90% required).	Data Center GPUs (NVIDIA A100/H100), Edge AI Accelerators, and high-end CPUs (widely used across all MLPerf divisions: Data Center, Edge, Tiny).	High IOPS Demand. Characterized by highly concurrent, random reads of many small data samples ($\approx$ 150 KB each), stressing metadata and IOPS capability.

Continued on next page

Table 2 MLCommons Benchmarks (Cont.)

Benchmark Name	Model	Task	Application Domain / Use Case	Model Type / Architecture	Metrics / KPIs	Hardware	Notes / Description
MLPerf Storage	3D U-Net	I/O Workload for Medical Image Segmentation Training	Healthcare/Radiology, Medical Image Analysis, 3D data processing.	3D U-Net (3D CNN)	Max Supported Accelerators, Aggregate Throughput (GiB/s), Accelerator Utilization ($\geq$ 90% required).	High-end Data Center GPUs (NVIDIA A100/H100) and specialized high-throughput storage systems (MLPerf Storage benchmark focus).	High Bandwidth Demand. Characterized by concurrent random reads of very large data files ($\approx$ 140 MB each), stressing sustained data throughput.
MLPerf Storage	CosmoFlow	I/O Workload for Scientific Parameter Prediction Training	Scientific High-Performance Computing (HPC), Astrophysics.	3D Convolutional Neural Network (3D CNN)	Max Supported Accelerators, Aggregate Throughput (GiB/s), Accelerator Utilization ($\geq$ 70% required).	Supercomputers & HPC Clusters: Requires massive scale distributed training across hundreds or thousands of GPUs (e.g., utilizing NVIDIA H100s, Intel Gaudi, and specialized high-speed interconnects like InfiniBand).	CPU-Intensive Workload. Uses medium-sized samples ($\approx$ 2 MB), but the client-side processing is more CPU-heavy, leading to a slightly lower required accelerator utilization threshold.
MLPerf Automotive							
MLPerf Automotive	SSD-ResNet50	2D Object Recognition and Segmentation	ADAS / Collision Avoidance, Lane Departure	Single Shot Detector (SSD) with ResNet-50 Backbone	Latency, Throughput, *mAP* (Accuracy)	Edge AI Accelerators, Embedded GPUs, and Automotive System-on-Chips (SoCs).	Baseline benchmark for camera-based detection on high-res (8MP) images. Used in v0.5.

Continued on next page

Table 2 MLCommons Benchmarks (Cont.)

Benchmark Name	Model	Task	Application Domain / Use Case	Model Type / Architecture	Metrics / KPIs	Hardware	Notes / Description
MLPerf Automotive	BEVFormer Tiny	Camera-based 3D Object Detection	Autonomous Driving (L2+ to L4), Environmental Perception	Bird's Eye View (BEV) Transformer-based Network	Latency, Throughput, mAP (Accuracy)	High-compute Automotive SoCs, next-generation AI accelerators (specifically targeting transformer and multi-sensor fusion capabilities).	Represents state-of-the-art camera-only 3D perception. Used in MLPerf Auto v0.5.
MLPerf Automotive	DeepLabV3+ / PointPainting	Semantic Segmentation (as a component of 3D Detection)	Lidar-Camera Sensor Fusion, 3D Perception	DeepLabV3+ (for Segmentation) + PointPillars (for 3D Detection)	Latency ($p99.9$ percentile), Throughput, Accuracy)	Safety-critical Automotive SoCs, purpose-built AI processors for ADAS/AV, often requiring high-reliability and low-latency performance.	DeepLabV3+ is the 2D segmentation part of the PointPainting sensor fusion pipeline. Used in MLPerf Inference v5.0 Automotive.
MLPerf Training:HPC							
MLPerf Training:HPC	CosmoFlow	Prediction of Cosmological Parameters ($\Omega_m, \sigma_8, n_s, H$)	Astrophysics, Cosmology, Scientific Simulation Parameter Prediction	3D Convolutional Neural Network (3D CNN)	Time-to-Train (Total time to reach a target quality metric), Aggregate Throughput (Models trained per unit of time in weak scaling).	Supercomputers & Large Clusters (e.g., NVIDIA Selene, Perlmutter, Fugaku), utilizing thousands of interconnected High-Performance GPUs (e.g., NVIDIA A100/H100) and high-speed parallel file systems.	Trained on 3D volumetric data (dark matter distributions) from N-body simulations. The large, volumetric data introduces significant I/O challenges and stresses high-bandwidth interconnects and storage.

Continued on next page

Table 2 MLCommons Benchmarks (Cont.)

Benchmark Name	Model	Task	Application Domain / Use Case	Model Type / Architecture	Metrics / KPIs	Hardware	Notes / Description
MLPerf Training:HPC	DeepCAM	Semantic Segmentation of Extreme Weather Events (e.g., atmospheric rivers, tropical cyclones)	Climate Science, Weather Forecasting, Earth System Modeling	Convolutional Encoder-Decoder (U-Net variant)	Time-to-Train (Total time to reach a target quality metric), Aggregate Throughput (Models trained per unit of time in weak scaling).	Supercomputers & Large Clusters, demanding high I/O bandwidth to handle the massive 8.8 TB climate datasets and requiring excellent strong-scaling performance. This benchmark was running on NVIDIA V100/A100.	Trained on massive, high-resolution, multi-channel images (e.g., 768×1152 pixels with 16 channels). Features high computational intensity and large memory footprint per sample.
MLPerf Training:HPC	OpenCatalyst	Prediction of energy and forces for molecular systems (AI for materials science)	Catalyst Discovery, Computational Chemistry, Materials Science, Energy Storage	Graph Neural Network (GNN), specifically DimeNet++	Time-to-Train (Total time to reach a target quality metric), Aggregate Throughput (Models trained per unit of time in weak scaling).	Supercomputers & Large Clusters, typically emphasizing the performance of GNNs, which stress different aspects of the system, like memory access patterns and graph-specific operations. This benchmark was running on NVIDIA V100/A100.	Predicts quantum mechanical properties of catalyst systems. Stresses complex data structures (graphs) and large-scale parallel processing. Uses the massive OC20 dataset.
MLCommons Science							

Continued on next page

Table 2 MLCommons Benchmarks (Cont.)

Benchmark Name	Model	Task	Application Domain / Use Case	Model Type / Architecture	Metrics / KPIs	Hardware	Notes / Description
MLCommons Science	Cloud Mask	image processing / segmentation	Earth Observation, Segmentation model for the pixel classification in satellite images	U-Net deep neural network	training and inference timing and scalability on the training across a number of GPUs;runtime of training and inference.	HPC Clusters & High-Performance GPUs (e.g., NVIDIA A100/V100) running distributed training frameworks like PyTorch or TensorFlow, often benchmarked for large-scale data I/O.	Focuses on identifying and isolating cloud cover in high-resolution satellite imagery for subsequent analysis.
MLCommons Science	STEMDL	A universal classifier for the space group of solid-state materials.	Scientific Machine Learning (General benchmark suite)	CNN: ResNet, VGG, DenseNet	top1 accuracy and F1 score (Macro)	HPC Systems of all sizes, used for general performance comparison across different hardware architectures and scaling tests. This benchmrak was running NVIDIA A100/V100.	The goals of this benchmark are to: (1) explore the suitability of machine learning algorithms in the advanced analysis of Convergent beam electron diffraction (CBED) and (2) produce a machine learning algorithm capable of overcoming intrinsic difficulties posed by scientific datasets.

Continued on next page

Table 2 MLCommons Benchmarks (Cont.)

Benchmark Name	Model	Task	Application Domain / Use Case	Model Type / Architecture	Metrics / KPIs	Hardware	Notes / Description
MLCommons Science	CANDLE UNO	Cancer Drug Response Prediction	Life Sciences / Personalized Medicine	Neural Networks(MLP)	TTT, Prediction Accuracy	HPC Systems (e.g., Summit, Polaris) and Cloud Environments, stressing both compute performance and workflow management for parameter sweep tasks. This benchmark was running NVIDIA A100.	Benchmarks deep learning models for predicting the response of various cancer cell lines to different therapeutic compounds.
MLCommons Science	Earthquake	TEvolOp Earthquake Forecasting Model	Earthquake Science	Neural Networks(MLP)-recurrent neural networks and transformers	Nash Sutcliffe efficiency	HPC & Big Data Systems, requiring efficient handling of large, continuous time-series datasets and high-throughput data processing. This benchmark was running on NVIDIA V100.	Benchmarks deep learning models for predicting the response of various cancer cell lines to different therapeutic compounds.
MLCommons AlgoPerf							
AlgoPerf	Criteo 1TB	Click-Through Rate (CTR) Prediction	Large-scale Recommender Systems, Digital Advertising	DLRM-Small (Deep Learning Recommendation Model)	Time-to-Result (Time to reach a target AUC)	Datacenter CPUs/GPUs with high memory bandwidth (HBM) due to massive embedding tables, and highly optimized network I/O.	Stresses memory access and sparse feature embedding computations due to the large, sparse Criteo 1TB dataset. Represents a common commercial workload.

Continued on next page

Table 2 MLCommons Benchmarks (Cont.)

Benchmark Name	Model	Task	Application Domain / Use Case	Model Type / Architecture	Metrics / KPIs	Hardware	Notes / Description
AlgoPerf	FastMRI	k-space MRI Reconstruction	Medical Imaging, Healthcare Diagnostics	U-Net (Convolutional Encoder-Decoder)	Time-to-Result (Time to reach a target PSNR / SSIM)	High-Performance GPUs and dedicated AI accelerators, as the model must run with high accuracy and low latency for clinical use.	Focuses on accelerating the image formation process from raw MRI data. U-Net is a standard model for semantic segmentation and image-to-image translation tasks.
AlgoPerf	ImageNet	Image Classification	General-purpose Computer Vision	ResNet-50 and Vision Transformer (ViT) variants	Time-to-Result (Time to reach a target Top-1 Accuracy	General-Purpose GPUs (Training/Inference), Edge Devices, and Mobile SoCs, as it is a widely-used test across all compute scales.	The quintessential computer vision workload. Includes two major architecture types (CNN and Transformer) to test algorithm generalizability.
AlgoPerf	LibriSpeech	Speech Recognition / ASR (Automatic Speech Recognition)	Voice Assistants, Transcription Services	Conformer and DeepSpeech variants	Time-to-Result (Time to reach a target Word Error Rate (WER))	Datacenter/Cloud GPUs (for large-scale ASR), Edge/Mobile Processors (for on-device assistants).	Tests algorithms on sequential data. Conformer is a hybrid CNN/Transformer architecture common in modern ASR.
AlgoPerf	OGBG	Graph Property Prediction	Scientific Machine Learning, Drug Discovery, Social Networks	GNN (Graph Neural Network)	Time-to-Result (Time to reach a target ROC-AUC)	Datacenter CPUs/GPUs with high-speed interconnects due to the irregular, sparse nature of graph-structured data.	Uses the Open Graph Benchmark (OGB) dataset. This workload stresses algorithms in domains that rely on non-Euclidean data structures.

Continued on next page

Table 2 MLCommons Benchmarks (Cont.)

Benchmark Name	Model	Task	Application Domain / Use Case	Model Type / Architecture	Metrics / KPIs	Hardware	Notes / Description
AlgoPerf	WMT	Machine Translation (En-De)	Natural Language Processing (NLP), Global Communication	Transformer (Base Architecture)	Time-to-Result (Time to reach a target BLEU Score)	Datacenter CPUs/GPUs with specialized Tensor Cores for efficient processing of the Transformer's self-attention mechanism.	A standard, large-scale sequence-to-sequence task, famous for being the original domain of the Transformer architecture.
MLCommons AILuminate							
AILuminate Safety v1.0	System Under Test (SUT) (Any LLM-based general-purpose chat system)	Assess Baseline AI Safety and Reliability	Pre-deployment Validation, Regulatory Compliance, Vendor Comparison	LLMs and AI Chat Systems (Text-to-Text), potentially with guardrails/filters	Overall Safety Grade (5-tier scale: Poor to Excellent), Violation Rate (% of unsafe responses), Per-Hazard Performance	The AI System itself (typically hosted in a Datacenter/Cloud) is the system under test (SUT). The evaluation is performed by a separate, specialized Safety Evaluator Model (often a tuned LLM ensemble).	Assesses safety against 12 Hazard Categories (e.g., Violent Crimes, Hate, Suicide & Self-Harm). Uses a tuned ensemble of safety evaluator models for grading. Focuses on single-turn, content-only hazards.
AILuminate Jailbreak Benchmark v0.5	System Under Test (SUT) (Any LLM-based general-purpose chat system)	Quantify Resilience to Adversarial "Jailbreak" Attacks	AI Security, Robustness Testing, Defense Mechanism Comparison	LLMs (Text-to-Text) and Vision-Language Models (VLMs) (Text+Image-to-Text)	Resilience Gap (Drop in safety performance from baseline to under-attack), Jailbreak Success Rate	The AI System (SUT) is tested in a Datacenter/Cloud environment. The benchmark focuses on the input (adversarial prompts) and the system's subsequent failure rate under attack conditions.	v0.5 is an initial release establishing the framework. It specifically measures the degradation of safety when a system is subjected to prompts designed to bypass its safety filters ("jailbreaks").

Table 3 Large Language Model Benchmark Details

Benchmark Name	Model	Task	Application Domain / Use Case	Model Type / Architecture	Metrics / KPIs	Notes / Description
Commercial/Proprietary LLMs (API/Systems)						
LLM Inference	Claude 3.5 Haiku 20241022	Generative AI	General Purpose, Light Reasoning	Large Transformer (Proprietary)	TTFT, TPOT, Throughput, MMLU (Quality)	A faster, smaller version in the Claude 3.5 family.
LLM Inference	Claude 3.5 Sonnet 20241022	Generative AI	Complex Reasoning, Data Processing	Large Transformer (Proprietary)	TTFT, TPOT, Throughput, MMLU (Quality)	Mid-tier model focusing on balance of speed and intelligence.
LLM Inference	Mistral Large 2402 Moderated	Generative AI	Enterprise Chatbots, Content Moderation	MoE/Dense Transformer (Proprietary)	TTFT, TPOT, Throughput, Safety Index	Flagged as moderated; emphasis on safety and reliable output.
LLM Inference	Amazon Nova Lite v1.0	Generative AI	AWS Services, Embedded Use Cases	Large Transformer (Proprietary)	Latency, Throughput, Cost/Token	Lightweight, cloud-optimized model.
LLM Inference	Gemini 1.5 Pro (API, with option)	Generative AI / Multimodal	Long Context, Multi-Source Reasoning	MoE/Dense Transformer (Proprietary, Multimodal)	TTFT, Throughput, Latency, RAG/Context Recall	Known for its massive context window.
LLM Inference	Gemini 2.0 Flash 001	Generative AI / Multimodal	High-Speed Chat, Real-time Tasks	Dense Transformer (Proprietary, Multimodal)	p99 Latency, Throughput	Focuses on speed and efficiency for low-latency needs.
LLM Inference	Gemini 2.0 Flash Lite	Generative AI	Edge/Client-Side Inference	Dense Transformer (Proprietary, Small)	Energy Efficiency, Latency	Highly optimized for resource-constrained environments.
LLM Inference	GPT-4o	Generative AI / Multimodal	Real-time Conversation, Vision Integration	Dense Transformer (Proprietary, Multimodal)	TTFT, TPOT, Low-Latency Response	All-in-one model for low-latency multimodal interactions.
LLM Inference	GPT-4o mini	Generative AI	Quick, Cost-Effective Tasks	Dense Transformer (Proprietary, Small)	Cost/Token, Throughput	Optimized for efficiency and scaling simple tasks.

Continued on next page

Table 3 – Continued from previous page

Benchmark Name	Model	Task	Application Domain / Use Case	Model Type / Architecture	Metrics / KPIs	Notes / Description
LLM Inference	Minustral 8B 24.10 (API)	Generative AI	General Text Generation	MoE/Dense Transformer (Proprietary)	Latency, Throughput	Represents a competitive, smaller model in a commercial API.
Open-Source/Bare Models (Used for Training or Deployment)						
LLM Inference	Minustral 8B 24.10 Moderation	Generative AI	General Text Generation, Safety Research	MoE/Dense Transformer (Open-weights)	Latency, Safety Compliance	Open-weight version with a focus on safety.
LLM Inference	Gemma 2 9b	Generative AI	Fine-tuning, Edge Deployment	Dense Transformer (Open-weights)	Perplexity, MMLU, Throughput	Smaller model from the Gemma family, good for fine-tuning.
LLM Inference	Phi 3.5 MoE Instruct	Generative AI	Instruction Following, Small Scale Reasoning	MoE (Open-weights, Small)	MMLU, HumanEval (Code)	Instruction-tuned, likely using a small Mixture-of-Experts.
LLM Inference	Phi 4	Generative AI	Research, Prototyping	Dense Transformer (Open-weights, Small)	Perplexity, BLEU (Generation)	Successor in the Phi family, typically very small.
LLM Inference	Athene V2 Chat Hf	Generative AI	Open Chatbot Deployment	Dense Transformer (Open-weights, Fine-tuned)	TTFT, TPOT, Chat Metrics	An instruction-tuned model from the Hugging Face ecosystem.
LLM Inference	Aya Expanse 8B Hf	Generative AI	Multilingual Tasks, Text Translation	Dense Transformer (Open-weights)	BLEU (Translation), Accuracy	Focused on broad language coverage.
LLM Inference	Cohere C4Ai Command A 03 2025 Hf	Generative AI	Enterprise RAG, Instruction Following	Dense Transformer (Open-weights)	Contextual Recall, RAG Latency	Cohere model variant used in the Hugging Face ecosystem.

Continued on next page

Table 3 – Continued from previous page

Benchmark Name	Model	Task	Application Domain / Use Case	Model Type / Architecture	Metrics / KPIs	Notes / Description
LLM Inference	Llama 3.1 405B Instruct	Generative AI	State-of-the-Art Reasoning, Long Context	Dense Transformer (Open-weights)	TTFT, Throughput, MMLU	An extremely large, cutting-edge open-weight model (used in MLPerf).
LLM Inference	Llama 3.1 8b Instruct FP8	Generative AI	Edge/Quantized Deployment	Dense Transformer (Quantized)	Inference Accuracy, Memory Footprint	Highly optimized for efficient computation using 8-bit precision.
LLM Inference	Llama 3 1 Tulu 3 8B Hf	Generative AI	General Chat, Fine-tuning Research	Dense Transformer (Open-weights, Fine-tuned)	Alpaca Eval, Human Preference	A variant of Llama tuned for instruction following.
LLM Inference	Mistralai Mistral Large 2402	Generative AI	Complex Reasoning, RAG	MoE/Dense Transformer (Open-weights)	TTFT, TPOT, MMLU	Open-weight version of Mistral's flagship model.
LLM Inference	Olmo 2 0325 32b Instruct	Generative AI	Research, Reproducible AI	Dense Transformer (Open-weights)	Perplexity, Training Speed	High-parameter model focused on openness and research.
LLM Inference	Olmo 2 1124 13B Instruct Hf	Generative AI	Instruction Following, General Chat	Dense Transformer (Open-weights)	TTFT, Throughput	Smaller, instruction-tuned version of the Olmo family.
LLM Inference	Phi 3.5 Mini Instruct	Generative AI	Mobile/Edge Inference, Simple Tasks	Dense Transformer (Open-weights, Small)	Latency, MMLU	Ultra-small model optimized for fast responses.
LLM Inference	Qwen1 5 110B Chat Hf	Generative AI	Multi-Language Chat, High Accuracy	Dense Transformer (Open-weights)	C-Eval, MMLU, Throughput	High-parameter model known for strong Chinese/general performance.
LLM Inference	Yi 1 5 34B Chat Hf	Generative AI	General Purpose, Instruction Following	Dense Transformer (Open-weights)	MMLU, C-Eval, Latency	Mid-to-large size model focusing on quality chat performance.

Continued on next page

Table 3 – Continued from previous page

Benchmark Name	Model	Task	Application Domain / Use Case	Model Type / Architecture	Metrics / KPIs	Notes / Description
LLM Inference	Ai21Labs Ai21 Jamba Large 1.5 Azure	Generative AI	Cloud Deployment, Enterprise Apps	Hybrid MoE/Dense Transformer	Throughput, Latency	A large model known for its hybrid architecture, deployed via Azure.
LLM Inference	Google Gemma 3 27B It Hf Nebius	Generative AI	Cloud Deployment, Fine-tuning	Dense Transformer (Open-weights, Fine-tuned)	TTFT, TPOT, Cloud Efficiency	Gemma model deployed on the Nebius cloud platform.
LLM Inference	Llama 3.3 70B Instruct Turbo Together	Generative AI	Fast, High-Quality Instruction Following	Dense Transformer (Open-weights)	Latency, Throughput, Cost	A large model optimized for speed via the Together API.
LLM Inference	Mistral Large 24.11	Generative AI	Enterprise AI, High Performance	MoE/Dense Transformer (Open-weights)	Throughput, MMLU, Reasoning	A very recent high-performance model.
LLM Inference	Qwq 32B Hf	Generative AI	General Purpose, Instruction Following	Dense Transformer (Open-weights)	Latency, Throughput	A mid-sized model in the open-weight ecosystem.
LLM Inference	OLMo 7b 0724 Instruct	Generative AI	Research, Instruction Following	Dense Transformer (Open-weights)	Perplexity, Speed	Smaller, instruction-tuned model for general tasks.

Table 3 Ontology Table for Selected AI Science Benchmarks.
(For detailed view of the Radar Charts, see [1].)

Ratings	Name	Domain	Models	Metrics	Citation
	ClimateLearn - Weather Forcasting	Climate & Earth Science	CNN baselines, ResNet variants	RMSE, Anomaly correlation	[2]
	ClimateLearn - Downscaling	Climate & Earth Science	CNN baselines, ResNet variants	RMSE, Anomaly correlation	[2]
	ClimateLearn - Climate Projection	Climate & Earth Science	CNN baselines, ResNet variants	RMSE, Anomaly correlation	[2]
	MLCommons Science - CloudMask	Climate & Earth Science	CNN, GNN, Transformer	MAE, Accuracy, Speedup vs simulation	[3]
	MLCommons Science - Earthquake	Climate & Earth Science	CNN, GNN, Transformer	MAE, Accuracy, Speedup vs simulation	[3]
	MLCommons Science - Candle UNO	Biology & Medicine	CNN, GNN, Transformer	MAE, Accuracy, Speedup vs simulation	[3]
	MLCommons Science - STEMDL	Materials Science	CNN, GNN, Transformer	MAE, Accuracy, Speedup vs simulation	[3]
	ARC-Challenge (Advanced Reasoning Challenge)	Computational Science & AI	GPT-4, Claude	Accuracy	[4]
	MOLGEN	Chemistry	MolGen	Validity%, Novelty%, QED, Docking score, penalized logP	[5]
	Open Graph Benchmark (OGB) - Biology	Biology & Medicine	GCN, GraphSAGE, GAT	Accuracy, ROC-AUC	[6]
	LLMs for Crop Science	Climate & Earth Science	GPT-3.5, GPT-4, Claude-3-opus, Qwen-max, LLama3-8B, InternLM2-7B, Qwen1.5-7B	Accuracy, F1 score	[7]
	SciCode	Computational Science & AI	Claude3.5-Sonnet	Solve rate (%)	[8]
	CaloChallenge 2022	High Energy Physics	VAE variants, GAN variants, Normalizing flows, Diffusion models	Histogram similarity, Classifier AUC, Generation latency	[9]
	PDEBench	Computational Science & AI, Climate & Earth Science, Mathematics	FNO, U-Net, PINN, Gradient-Based inverse methods	RMSE, boundary RMSE, Fourier RMSE	[10]

Continued on next page

Table 3 Ontology Table for Selected AI Science Benchmarks (cont.).

Ratings	Name	Domain	Models	Metrics	Citation
	Urban Data Layer (UDL) - PM2.5 Concentration Prediction	Climate & Earth Science	Baseline regression/classification pipelines	Task-specific accuracy or RMSE	[11]
	Urban Data Layer (UDL) - Built-up Area Classification	Climate & Earth Science	Baseline regression/classification pipelines	Task-specific accuracy or RMSE	[11]
	Urban Data Layer (UDL) - Administrative Boundaries Identification	Climate & Earth Science	Baseline regression/classification pipelines	Task-specific accuracy or RMSE	[11]
	Urban Data Layer (UDL) - El Nino Anomaly Detection	Climate & Earth Science	Baseline regression/classification pipelines	Task-specific accuracy or RMSE	[11]
	SPIQA (LLM)	Computational Science & AI	LLaVA, MiniGPT-4, Owl-LLM adapter variants	Accuracy, F1 score	[12]
	MLCommons Medical AI - Pancreas Segmentation (DFCI)	Biology & Medicine	MedPerf-validated CNNs, GaNDLF workflows	ROC AUC, Accuracy, Fairness metrics	[13]
	MLCommons Medical AI - Brain Tumor Segmentation (BraTS)	Biology & Medicine	MedPerf-validated CNNs, GaNDLF workflows	ROC AUC, Accuracy, Fairness metrics	[13]
	MLCommons Medical AI - Surgical Workflow Phase Recognition (SurgMLCube)	Biology & Medicine	MedPerf-validated CNNs, GaNDLF workflows	ROC AUC, Accuracy, Fairness metrics	[13]
	SeafloorAI	Climate & Earth Science	SegFormer, ViLT-style multimodal models	Segmentation pixel accuracy, QA accuracy	[14]
	SeafloorGenAI	Climate & Earth Science	SegFormer, ViLT-style multimodal models	Segmentation pixel accuracy, QA accuracy	[14]
	GeSS - Track Pileup	High Energy Physics	GCN, EGNN, DimeNet++	Accuracy, RMSE, OOD robustness delta	[15]
	GeSS - Track Signal	High Energy Physics	GCN, EGNN, DimeNet++	Accuracy, RMSE, OOD robustness delta	[15]
	GeSS - DrugOOD	Biology & Medicine	GCN, EGNN, DimeNet++	Accuracy, RMSE, OOD robustness delta	[15]
	GeSS - QMOF	Materials Science	GCN, EGNN, DimeNet++	Accuracy, RMSE, OOD robustness delta	[15]
	OCP (Open Catalyst Project)	Chemistry, Materials Science	CGCNN, SchNet, DimeNet++, GemNet-OC	MAE (energy), MAE (force)	[16, 17]

Continued on next page

Table 3 Ontology Table for Selected AI Science Benchmarks (cont.).

Ratings	Name	Domain	Models	Metrics	Citation
	Jet Classification	High Energy Physics	Keras DNN, QKeras quantized DNN	Accuracy, AUC	[18]
	Irregular Sensor Data Compression	High Energy Physics	Autoencoder, Quantized autoencoder	MSE, Compression ratio	[18]
	MLPerf HPC - Cosmoflow	High Energy Physics	CosmoFlow, DeepCAM, OpenCatalyst	Training time, Accuracy, GPU utilization	[19]
	MLPerf HPC - DeepCAM	Climate & Earth Science	DeepCAM	Training time, Accuracy, GPU utilization	[19]
	MLPerf HPC - Open Catalyst Project DimeNet++	Chemistry	DeepCAM	Training time, Accuracy, GPU utilization	[19]
	MLPerf HPC - OpenFold	Biology & Medicine	DeepCAM	Training time, Accuracy, GPU utilization	[19]
	HDR ML Anomaly Challenge - Gravitational Waves	High Energy Physics	Deep latent CNNs, Autoencoders	ROC-AUC, Precision/Recall	[20]
	SuperCon3D - Property Prediction	Materials Science	SODNet, DiffCSP-SC	MAE (Tc), Validity of generated structures	[21]
	SuperCon3D - Inverse Crystal Structure Generation	Materials Science	SODNet, DiffCSP-SC	MAE (Tc), Validity of generated structures	[21]
	BaisBench (Biological AI Scientist Benchmark) - Question Answering	Biology & Medicine	LLM-based AI scientist agents	Annotation accuracy, QA accuracy	[22]
	BaisBench (Biological AI Scientist Benchmark) - Cell Type Annotation	Biology & Medicine	LLM-based AI scientist agents	Annotation accuracy, QA accuracy	[22]
	The Well	Biology & Medicine, Computational Science & AI, High Energy Physics	FNO baselines, U-Net baselines	Dataset size, Domain breadth	[23]
	MMLU (Massive Multitask Language Understanding)	Computational Science & AI	GPT-4o, Gemini 1.5 Pro, o1, DeepSeek-R1	Accuracy	[24]
	SatImgNet	Climate & Earth Science	CLIP, BLIP, ALBEF	Accuracy	[25]
	GPQA Diamond	Biology & Medicine, Chemistry, High Energy Physics	o1, DeepSeek-R1	Accuracy	[26]
	PRM800K	Mathematics	GPT-4	Accuracy	[27]

Continued on next page

Table 3 Ontology Table for Selected AI Science Benchmarks (cont.).

Ratings	Name	Domain	Models	Metrics	Citation
	FEABench (Finite Element Analysis Benchmark): Evaluating Language Models on Multiphysics Reasoning Ability	Mathematics	FEniCS, deal.II	Solve time, Error norm	[28]
	Neural Architecture Codesign for Fast Physics Applications	High Energy Physics	NAC-based BraggNN, NAC-optimized Deep Sets (jet)	Accuracy, Latency, Resource utilization	[29]
	Delta Squared-DFT	Chemistry, Materials Science	Delta Squared-ML correction networks, Kernel ridge regression	Mean Absolute Error (eV), Energy ranking accuracy	[30]
	HDR ML Anomaly Challenge - Sea Level Rise	Climate & Earth Science	CNNs, RNNs, Transformers	ROC-AUC, Precision/Recall	[20]
	Vocal Call Locator (VCL)	Biology & Medicine	CNN-based SSL models	Localization error (cm), Recall/Precision	[31]
	MassSpecGym - De novo molecule generation	Chemistry	Graph-based generative models, Retrieval baselines	Structure accuracy, Retrieval precision, Simulation MSE	[32]
	MassSpecGym - Molecule Retrieval	Chemistry	Graph-based generative models, Retrieval baselines	Structure accuracy, Retrieval precision, Simulation MSE	[32]
	MassSpecGym - Spectrum Simulation	Chemistry	Graph-based generative models, Retrieval baselines	Structure accuracy, Retrieval precision, Simulation MSE	[32]
	SPIQA (Scientific Paper Image Question Answering)	Computational Science & AI	Chain-of-Thought models, Multimodal QA systems	Accuracy, F1 score	[33]
	GPQA: A Graduate-Level Google-Proof Question and Answer Benchmark	Biology & Medicine, High Energy Physics, Chemistry	GPT-4 baseline	Accuracy	[34]
	MedQA	Biology & Medicine	Neural reader, Retrieval-based QA systems	Accuracy	[35]
	Single Qubit Readout on QICK System	Computational Science & AI	hls4ml quantized NN	Accuracy, Latency	[36]
	CFDBench (Fluid Dynamics)	Mathematics	FNO, DeepONet, U-Net	L2 error, MAE	[37]
	CURIE (Scientific Long-Context Understanding, Reasoning and Information Extraction)	Materials Science, High Energy Physics, Biology & Medicine, Chemistry, Climate & Earth Science	unkown	Accuracy	[38]

Continued on next page

Table 3 Ontology Table for Selected AI Science Benchmarks (cont.).

Ratings	Name	Domain	Models	Metrics	Citation
	Smart Pixels for LHC	High Energy Physics	2-layer pixel NN	Data rejection rate, Power per pixel	[39]
	LHC New Physics Dataset	High Energy Physics	Autoencoder, Variational autoencoder, Isolation forest	ROC-AUC, Detection efficiency	[40]
	Quantum Computing Benchmarks (QML)	Computational Science & AI	IBM Q, IonQ, AQT@LBNL	Fidelity, Success probability	[41]
	Ultrafast jet classification at the HL-LHC	High Energy Physics	MLP, Deep Sets, Interaction Network	Accuracy, Latency, Resource utilization	[42]
	HEDM (BraggNN)	Materials Science	BraggNN	Localization accuracy, Inference time	[43]
	4D-STEM	Materials Science	CNN models (prototype)	Classification accuracy, Throughput	[44]
	Beam Control	High Energy Physics	DDPG, PPO (planned)	Stability, Control loss	[18, 45]
	Intelligent experiments through real-time AI	High Energy Physics	Bipartite Graph Network with Set Transformers (BGN-ST), GarNet (edge-classifier)	Accuracy (charm and beauty detection), Latency (micros), Resource utilization (LUT/FF/BRAM/DSP)	[46]
	HDR ML Anomaly Challenge - Butterfly	Biology & Medicine	CNN-based detectors	Classification accuracy, F1 score	[20]
	DUNE	High Energy Physics	CNN, LSTM (planned)	Detection efficiency, Latency	[47]
	FrontierMath	Mathematics	unknown	Accuracy	[48]
	AIME (American Invitational Mathematics Examination)	Mathematics	unknown	Accuracy	[49]
	Quench detection	High Energy Physics	Autoencoder, RL agents (in development)	ROC-AUC, Detection latency	[50]
	Materials Project	Materials Science	Automatminer, Crystal Graph Neural Networks	MAE, R^2	[51]
	In-Situ High-Speed Computer Vision	High Energy Physics	CNN	Accuracy, FPS	[52]

Table 4 Summary of Example Profiling Tools Useful for Deep Learning and AI Workloads

Tool / Category	Vendor / Maintainer	Level / Primary Use Case	Key Features and Capabilities
Framework Profilers			
PyTorch Profiler [53]	Meta	Framework-level	Records CPU/GPU activities, memory usage, and operator timings; integrates with TensorBoard and Perfetto; useful for training optimization and layer timing.
TensorBoard / TensorFlow Profiler [53]	Google	Framework-level	Visualizes input pipelines, GPU kernels, and op-level timings; includes memory and device utilization tracing; supports bottleneck analysis.
torch.utils.bottleneck [54]	Meta	Framework-level	Combines autograd and Python profilers for quick bottleneck diagnostics.
JAX Profiler [55]	Google	Framework-level	Works with TensorBoard to trace XLA compilation, HLO graphs, and TPU/GPU runtime performance.
NVIDIA DLProf [56]	NVIDIA	Framework-level (GPU-focused)	High-level view of deep learning layers and operations; integrates with TensorBoard DLProf plugin.
Hardware / System Profilers			
Nsight Systems [57]	NVIDIA	System-level	Timeline visualization of CPU–GPU interactions, kernel launch overheads, multi-process analysis, and NCCL tracing.
Nsight Compute [58]	NVIDIA	Kernel-level	Detailed GPU kernel performance metrics: memory throughput, Tensor Core utilization, occupancy, and roofline analysis.
nvprof (deprecated) [59]	NVIDIA	GPU-level	Legacy command-line CUDA profiler, replaced by Nsight tools.
VTune Profiler [60]	Intel	CPU/System-level	Hotspot analysis, vectorization, threading efficiency, and CPU performance bottlenecks.
omnitrace / rocprof / rocm-smi [61]	AMD	GPU-level	Profiling and monitoring for AMD GPUs: kernel execution metrics, power, and temperature.
HPCToolkit [62]	Rice University	System-level (CPU+GPU)	Hierarchical performance profiling, time attribution to calling context, supports CUDA and HIP.
TAU [63]	University of Oregon	System-level (CPU+GPU+MPI)	Multi-level performance analysis, MPI integration, supports heterogeneous systems.
Perfetto [64]	Google (Open Source)	System-level	High-resolution trace visualization, interoperable with PyTorch/TensorFlow profiler exports.

Continued on next page

Table 4 Summary of Example Profiling Tools Useful for Deep Learning and AI Workloads (Cont.)

Tool / Category	**Vendor / Maintainer**	**Level / Primary Use Case**	**Key Features and Capabilities**
PAPI [65]	University of Tennessee	Hardware counter interface	Provides access to CPU/GPU performance counters for integration with other profiling tools or custom instrumentation.
Compiler / Graph Profilers			
XLA Profiler [66]	Google	Compiler-level (XLA)	Profiles XLA-compiled operations and execution times; supports JAX/TF and TPU/GPU workloads.
TorchDynamo / TorchInductor Debug Tools [67]	Meta	Compiler-level (PyTorch 2.x)	Analyzes graph fusion, compiler optimizations, and operator performance of compiled PyTorch models.
Triton Profiler [68]	OpenAI	Kernel-level (Custom Kernels)	Reports kernel execution time, register usage, and occupancy for custom Triton GPU kernels.
Communication / Distributed Profilers			
NCCL Profiler [69]	NVIDIA	Communication-level	Profiles NCCL collective communication operations (e.g., all-reduce, broadcast); timeline visualization of multi-GPU communication.
AWS SageMaker Debugger / Azure Profiler [70, 71]	AWS / Microsoft	Cloud-level	Distributed GPU/CPU monitoring, training metric collection, and profiling at cloud scale.
Weights & Biases, Comet, MLflow [72, 73, 74]	Multiple Vendors	Experiment / Cloud-level	Logs performance traces, GPU utilization, integrates with PyTorch and TensorFlow profilers for real-time monitoring.
System & Memory Profilers			
Torch / TensorFlow Memory Tools [75]	Meta / Google	Framework-level (Memory)	Reports GPU memory allocation, fragmentation, and utilization trends for debugging memory bottlenecks.
Python Profilers (cProfile, py-spy) [76]	Python Community	CPU-level	Measures Python-level overhead and I/O performance; used for diagnosing data preprocessing bottlenecks.

REFERENCES

1.G. von Laszewski, B. Hawks, M. Colombo, R. Shiraishi, A. Krishnan, N. Tran, and G. C. Fox, "Mlcommons science working group ai benchmarks collection," GitHub, Jun. 2025, online Collection: =https://mlcommons-science.github.io/benchmark/. [Online]. Available: https://mlcommons-science.github.io/benchmark/benchmarks.pdf

2.T. Nguyen, J. Jewik, H. Bansal, P. Sharma, and A. Grover, "Climatelearn: Benchmarking machine learning for weather and climate modeling," arXiv, 2023. [Online]. Available: https://arxiv.org/abs/2307.01909

3.J. Thiyagalingam, G. von Laszewski, J. Yin, M. Emani, J. Papay, G. Barrett, P. Luszczek, A. Tsaris, C. Kirkpatrick, F. Wang, T. Gibbs, V. Vishwanath, M. Shankar, G. Fox, and T. Hey, "Ai benchmarking for science: Efforts from the mlcommons science working group," in *High Performance Computing. ISC High Performance 2022 International Workshops*, H. Anzt, A. Bienz, P. Luszczek, and M. Baboulin, Eds. Cham: Springer International Publishing, 2022, pp. 47–64.

4.P. Clark, I. Cowhey, O. Etzioni, T. Khot, A. Sabharwal, C. Schoenick, and O. Tafjord, "Think you have solved question answering? try arc, the ai2 reasoning challenge," arXiv:1803.05457v1, 2018.

5.Y. Fang, N. Zhang, Z. Chen, L. Guo, X. Fan, and H. Chen, "Domain-agnostic molecular generation with chemical feedback," arXiv, 2024. [Online]. Available: https://arxiv.org/abs/2301.11259

6.W. Hu, M. Fey, M. Zitnik, Y. Dong, H. Ren, B. Liu, M. Catasta, and J. Leskovec, "Open Graph Benchmark: Datasets for Machine Learning on Graphs," arXiv, 2021. [Online]. Available: https://arxiv.org/abs/2005.00687

7.H. Zhang, J. Sun, R. Chen, W. Liu, Z. Yuan, X. Zheng, Z. Wang, Z. Yang, H. Yan, H.-S. Zhong, X. Wang, W. Ouyang, F. Yang, and N. Dong, "Empowering and assessing the utility of large language models in crop science," in *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. [Online]. Available: https://openreview.net/forum?id=hMj6jZ6JWU

8.M. Tian, L. Gao, S. D. Zhang, X. Chen, C. Fan, X. Guo, R. Haas, P. Ji, K. Krongchon, Y. Li, S. Liu, D. Luo, Y. Ma, H. Tong, K. Trinh, C. Tian, Z. Wang, B. Wu, Y. Xiong, S. Yin, M. Zhu, K. Lieret, Y. Lu, G. Liu, Y. Du, T. Tao, O. Press, J. Callan, E. Huerta, and H. Peng, "Scicode: A research coding benchmark curated by scientists," arXiv, 2024. [Online]. Available: https://arxiv.org/abs/2407.13168

9.C. Krause, M. F. Giannelli, G. Kasieczka, B. Nachman, D. Salamani, D. Shih, A. Zaborowska, O. Amram, K. Borras, M. R. Buckley, E. Buhmann, T. Buss, R. P. D. C. Cardoso, A. L. Caterini, N. Chernyavskaya, F. A. G. Corchia, J. C. Cresswell, S. Diefenbacher, E. Dreyer, V. Ekambaram, E. Eren, F. Ernst, L. Favaro, M. Franchini, F. Gaede, E. Gross, S.-C. Hsu, K. Jaruskova, B. Käch, J. Kalagnanam, R. Kansal, T. Kim, D. Kobylianskii, A. Korol, W. Korcari, D. Krücker, K. Krüger, M. Letizia, S. Li, Q. Liu, X. Liu, G. Loaiza-Ganem, T. Madula, P. McKeown, I.-A. Melzer-Pellmann, V. Mikuni, N. Nguyen, A. Ore, S. P. Schweitzer, I. Pang, K. Pedro, T. Plehn, W. Pokorski, H. Qu, P. Raikwar, J. A. Raine, H. Reyes-Gonzalez, L. Rinaldi, B. L. Ross, M. A. W. Scham, S. Schnake, C. Shimmin, E. Shlizerman, N. Soybelman, M. Srivatsa, K. Tsolaki, S. Vallecorsa, K. Yeo, and R. Zhang, "Calochallenge 2022: A community challenge for fast calorimeter simulation," arXiv, 2024. [Online]. Available: https://arxiv.org/abs/2410.21611

10.M. Takamoto, T. Praditia, R. Leiteritz, D. MacKinlay, F. Alesiani, D. Pflüger, and M. Niepert, "Pdebench: An extensive benchmark for scientific machine learning," arXiv, 2024. [Online]. Available: https://arxiv.org/abs/2210.07182

11.Y. Wang, T. Wang, Y. Zhang, H. Zhang, H. Zheng, G. Zheng, and L. Kong, "Urbandatalayer: A unified data pipeline for urban science," in *Advances in Neural Information Processing Systems*, A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, Eds., vol. 37. Curran Associates, Inc., 2024, pp. 7296–7310. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2024/file/0db7f135f6991e8cec5e516ecc66bfba-Paper-Datasets_and_Benchmarks_Track.pdf

12.S. Pramanick, R. Chellappa, and S. Venugopalan, "Spiqa: A dataset for multimodal question answering on scientific papers," arXiv, 2025. [Online]. Available: https://arxiv.org/abs/2407.09413

13.A. Karargyris, R. Umeton, M. J. Sheller, A. Aristizabal, J. George, A. Wuest, S. Pati, H. Kassem, M. Zenk, U. Baid, P. Narayana Moorthy, A. Chowdhury, J. Guo, S. Nalawade, J. Rosenthal, D. Kanter, M. Xenochristou, D. J. Beutel, V. Chung, T. Bergquist, J. Eddy, A. Abid, L. Tunstall, O. Sanseviero, D. Dimitriadis, Y. Qian, X. Xu, Y. Liu, R. S. M. Goh, S. Bala, V. Bittorf, S. R. Puchala, B. Ricciuti, S. Samineni, E. Sengupta, A. Chaudhari, C. Coleman, B. Desinghu, G. Diamos, D. Dutta, D. Feddema, G. Fursin, X. Huang, S. Kashyap, N. Lane, I. Mallick, P. Mascagni, V. Mehta, C. F. Moraes, V. Natarajan, N. Nikolov, N. Padoy, G. Pekhimenko, V. J. Reddi, G. A. Reina, P. Ribalta, A. Singh, J. J. Thiagarajan, J. Albrecht, T. Wolf, G. Miller, H. Fu, P. Shah, D. Xu, P. Yadav, D. Talby, M. M. Awad, J. P. Howard, M. Rosenthal, L. Marchionni, M. Loda, J. M. Johnson, S. Bakas, P. Mattson, F. Consortium, B.-. Consortium, and A. Consortium, "Federated benchmarking of medical artificial intelligence with medperf," *Nature Machine Intelligence*, vol. 5, no. 7, pp. 799–810, Jul. 2023. [Online]. Available: https://doi.org/10.1038/s42256-023-00652-2

14.K. X. Nguyen, F. Qiao, A. Trembanis, and X. Peng, "Seafloorai: A large-scale vision-language dataset for seafloor geological survey," arXiv, 2024. [Online]. Available: https://arxiv.org/abs/2411.00172

15.D. Zou, S. Liu, S. Miao, V. Fung, S. Chang, and P. Li, "Gess: Benchmarking geometric deep learning under scientific applications with distribution shifts," in *Advances in Neural Information Processing Systems*, A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, Eds., vol. 37. Curran Associates, Inc., 2024, pp. 92 499–92 528. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2024/file/a8063075b00168dc39bc81683619f1a8-Paper-Datasets_and_Benchmarks_Track.pdf

16.L. Chanussot, A. Das, S. Goyal, T. Lavril, M. Shuaibi, M. Riviere, K. Tran, J. Heras-Domingo, C. Ho, W. Hu, A. Palizhati, A. Sriram, B. Wood, J. Yoon, D. Parikh, C. L. Zitnick, and Z. Ulissi, "The open catalyst 2020 (oc20) dataset and community challenges," *ACS Catalysis*, vol. 11, no. 10, pp. 6059–6072, 2021. [Online]. Available: https://pubs.acs.org/doi/10.1021/acscatal.0c04525

17.R. Tran, J. Lan, M. Shuaibi, B. M. Wood, S. Goyal, A. Das, J. Heras-Domingo, A. Kolluru, A. Rizvi, N. Shoghi, A. Sriram, F. Therrien, J. Abed, O. Voznyy, E. H. Sargent, Z. Ulissi, and C. L. Zitnick, "The open catalyst 2022 (oc22) dataset and challenges for oxide electrocatalysts," *ACS Catalysis*, vol. 13, no. 5, pp. 3066–3084, 2023. [Online]. Available: https://pubs.acs.org/doi/10.1021/acscatal.2c05426

18.J. Duarte, N. Tran, B. Hawks, C. Herwig, J. Muhizi, S. Prakash, and V. J. Reddi, "Fastml science benchmarks: Accelerating real-time scientific edge machine learning," arXiv, 2022. [Online]. Available: https://arxiv.org/abs/2207.07958

19.S. Farrell, M. Emani, J. Balma, L. Drescher, A. Drozd, A. Fink, G. Fox, D. Kanter, T. Kurth, P. Mattson, D. Mu, A. Ruhela, K. Sato, K. Shirahata, T. Tabaru, A. Tsaris, J. Balewski, B. Cumming, T. Danjo, J. Domke, T. Fukai, N. Fukumoto, T. Fukushi, B. Gerofi, T. Honda, T. Imamura, A. Kasagi, K. Kawakami, S. Kudo, A. Kuroda, M. Martinasso, S. Matsuoka, H. Mendonça, K. Minami, P. Ram, T. Sawada, M. Shankar, T. S. John, A. Tabuchi, V. Vishwanath, M. Wahib, M. Yamazaki, and J. Yin, "Mlperf hpc: A holistic benchmark suite for scientific machine learning on hpc systems," arXiv, 2021. [Online]. Available: https://arxiv.org/abs/2110.11466

20.E. G. Campolongo, Y.-T. Chou, E. Govorkova, W. Bhimji, W.-L. Chao, C. Harris, S.-C. Hsu, H. Lapp, M. S. Neubauer, J. Namayanja, A. Subramanian, P. Harris, A. Anand, D. E. Carlyn, S. Ghosh, C. Lawrence, E. Moreno, R. Raikman, J. Wu, Z. Zhang, B. Adhi, M. A. Gharehtoragh, S. A. Monsalve, M. Babicz, F. Baig, N. Banerji, W. Bardon, T. Barna, T. Berger-Wolf, A. B. Dieng, M. Brachman, Q. Buat, D. C. Y. Hui, P. Cao, F. Cerino, Y.-C. Chang, S. Chaulagain, A.-K. Chen, D. Chen, E. Chen, C.-J. Chou, Z.-C. Ciou, M. Cochran-Branson, A. C. O. Choi, M. Coughlin, M. Cremonesi, M. Dadarlat, P. Darch, M. Desai, D. Diaz, S. Dillmann, J. Duarte, I. Duporge, U. Ekka, S. E. Heravi, H. Fang, R. Flynn, G. Fox, E. Freed, H. Gao, J. Gao, J. Gonski, M. Graham, A. Hashemi, S. Hauck, J. Hazelden, J. H. Peterson, D. Hoang, W. Hu, M. Huennefeld, D. Hyde, V. Janeja, N. Jaroenchai, H. Jia, Y. Kang, M. Kholiavchenko, E. E. Khoda, S. Kim, A. Kumar, B.-C. Lai, T. Le, C.-W. Lee, J. Lee, S. Lee, S. van der Lee, C. Lewis, H. Li, H. Li, H. Liao, M. Liu, X. Liu, X. Liu, V. Loncar, F. Lyu, I. Makarov, A. M. C.-Y. Mao, A. Michels, A. Migala, F. Mokhtar, M. Morlighem, M. Namgung, A. Novak, A. Novick, A. Orsborn, A. Padmanabhan, J.-C. Pan, S. Pandya, Z. Pei, A. Peixoto, G. Percivall, A. P. Leung, S. Purushotham, Z. Que, M. Quinnan, A. Ranjan, D. Rankin, C. Reissel, B. Riedel, D. Rubenstein, A. Sasli, E. Shlizerman, A. Singh, K. Singh, E. R. Sokol, A. Sorensen, Y. Su, M. Taheri, V. Thakkar, A. M. Thomas, E. Toberer, C. Tsai, R. Vandewalle, A. Verma, R. C. Venterea, H. Wang, J. Wang, S. Wang, S. Wang, G. Watts, J. Weitz, A. Wildridge, R. Williams, S. Wolf, Y. Xu, J. Yan, J. Yu, Y. Zhang, H. Zhao, Y. Zhao, and Y. Zhong, "Building machine learning challenges for anomaly detection in science," arXiv, 2025. [Online]. Available: https://arxiv.org/abs/2503.02112

21.P. Chen, L. Peng, R. Jiao, Q. Mo, Z. Wang, W. Huang, Y. Liu, and Y. Lu, "Learning superconductivity from ordered and disordered material structures," in *Advances in Neural Information Processing Systems*, A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, Eds., vol. 37. Curran Associates, Inc., 2024, pp. 108 902–108 928. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2024/file/c4e3b55ed4ac9ba52d7df11f8bddbbf4-Paper-Datasets_and_Benchmarks_Track.pdf

22.E. Luo, J. Jia, Y. Xiong, X. Li, X. Guo, B. Yu, L. Wei, and X. Zhang, "Benchmarking ai scientists in omics data-driven biological research," arXiv, 2025. [Online]. Available: https://arxiv.org/abs/2505.08341

23.R. Ohana, M. McCabe, L. Meyer, R. Morel, F. J. Agocs, M. Beneitez, M. Berger, B. Burkhart, S. B. Dalziel, D. B. Fielding, D. Fortunato, J. A. Goldberg, K. Hirashima, Y.-F. Jiang, R. R. Kerswell, S. Maddu, J. Miller, P. Mukhopadhyay, S. S. Nixon, J. Shen, R. Watteaux, B. R.-S. Blancard, F. Rozet, L. H. Parker, M. Cranmer, and S. Ho, "The well: a large-scale collection of diverse physics simulations for machine learning," in *Advances in Neural Information Processing Systems*, A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, Eds., vol. 37. Curran Associates, Inc., 2024, pp.

44 989–45 037. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2024/file/4f9a5acd91ac76569f2fe291b1f4772b-Paper-Datasets_and_Benchmarks_Track.pdf

24.D. Hendrycks, C. Burns, and S. Kadavath, "Measuring Massive Multitask Language Understanding," 2021. [Online]. Available: https://arxiv.org/abs/2009.03300

25.J. Roberts, K. Han, and S. Albanie, "Satin: A multi-task metadataset for classifying satellite imagery using vision-language models," *ICCV Workshop: Towards the Next Generation of Computer Vision Datasets*, 3 2023.

26.D. Rein, B. L. Hou, and A. C. Stickland, "Gpqa: A graduate-level google-proof q and a benchmark," 2023. [Online]. Available: https://arxiv.org/abs/2311.12022

27.H. Lightman, V. Kosaraju, Y. Burda, H. Edwards, B. Baker, T. Lee, J. Leike, J. Schulman, I. Sutskever, and K. Cobbe, "Let's verify step by step," *arXiv preprint arXiv:2305.20050*, 2023.

28.N. Mudur, H. Cui, S. Venugopalan, P. Raccuglia, M. P. Brenner, and P. Norgaard, "Feabench: Evaluating language models on multiphysics reasoning ability," arXiv, 2025. [Online]. Available: https://arxiv.org/abs/2504.06260

29.J. Weitz, D. Demler, L. McDermott, N. Tran, and J. Duarte, "Neural architecture codesign for fast physics applications," arXiv, 2025. [Online]. Available: https://arxiv.org/abs/2501.05515

30.K. Khrabrov, A. Ber, A. Tsypin, K. Ushenin, E. Rumiantsev, A. Telepov, D. Protasov, I. Shenbin, A. Alekseev, M. Shirokikh, S. Nikolenko, E. Tutubalina, and A. Kadurin, "Delta-squared dft: A universal quantum chemistry dataset of drug-like molecules and a benchmark for neural network potentials," arXiv, 2024. [Online]. Available: https://arxiv.org/abs/2406.14347

31.R. E. Peterson, A. Tanelus, C. Ick, B. Mimica, N. Francis, V. J. Ivan, A. Choudhri, A. L. Falkner, M. Murthy, D. M. Schneider, D. H. Sanes, and A. H. Williams, "Vocal call locator benchmark (vcl) for localizing rodent vocalizations from multi-channel audio," in *Advances in Neural Information Processing Systems*, A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, Eds., vol. 37. Curran Associates, Inc., 2024, pp. 106 370–106 382. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2024/file/c00d37d6b04d73b870b963a4d70051c1-Paper-Datasets_and_Benchmarks_Track.pdf

32.R. Bushuiev, A. Bushuiev, N. F. de Jonge, A. Young, F. Kretschmer, R. Samusevich, J. Heirman, F. Wang, L. Zhang, K. Dührkop, M. Ludwig, N. A. Haupt, A. Kalia, C. Brungs, R. Schmid, R. Greiner, B. Wang, D. S. Wishart, L.-P. Liu, J. Rousu, W. Bittremieux, H. Rost, T. D. Mak, S. Hassoun, F. Huber, J. J. van der Hooft, M. A. Stravs, S. Böcker, J. Sivic, and T. Pluskal, "Massspecgym: A benchmark for the discovery and identification of molecules," in *Advances in Neural Information Processing Systems*, A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, Eds., vol. 37. Curran Associates, Inc., 2024, pp. 110 010–110 027. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2024/file/c6c31413d5c53b7d1c343c1498734b0f-Paper-Datasets_and_Benchmarks_Track.pdf

33.X. Zhong, Y. Gao, and S. Gururangan, "Spiqa: Scientific paper image question answering," 2024. [Online]. Available: https://arxiv.org/abs/2407.09413

34.D. Rein, B. L. Hou, A. C. Stickland, J. Petty, R. Y. Pang, J. Dirani, J. Michael, and S. R. Bowman, "Gpqa: A graduate-level google-proof q and a benchmark," arXiv, 2023. [Online]. Available: https://arxiv.org/abs/2311.12022

35.D. Jin, E. Pan, N. Oufattole, W.-H. Weng, H. Fang, and P. Szolovits, "What disease does this patient have? a large-scale open domain question answering dataset from medical exams," arXiv, 2020. [Online]. Available: https://arxiv.org/abs/2009.13081

36.G. D. Guglielmo, B. Du, J. Campos, A. Boltasseva, A. V. Dixit, F. Fahim, Z. Kudyshev, S. Lopez, R. Ma, G. N. Perdue, N. Tran, O. Yesilyurt, and D. Bowring, "End-to-end workflow for machine learning-based qubit readout with qick and hls4ml," arXiv, 2025. [Online]. Available: https://arxiv.org/abs/2501.14663

37.Y. Luo, Y. Chen, and Z. Zhang, "Cfdbench: A large-scale benchmark for machine learning methods in fluid dynamics," 2024. [Online]. Available: https://arxiv.org/abs/2310.05963

38.H. Cui, Z. Shamsi, G. Cheon, X. Ma, S. Li, M. Tikhanovskaya, P. Norgaard, N. Mudur, M. Plomecka, P. Raccuglia, Y. Bahri, V. V. Albert, P. Srinivasan, H. Pan, P. Faist, B. Rohr, E. D. Cubuk, M. Aykol, A. Merchant, M. J. Statt, D. Morris, D. Purves, E. Kleeman, R. Alcantara, M. Abraham, M. Mohammad, E. P. VanLee, C. Jiang, E. Dorfman, E.-A. Kim, M. P. Brenner, V. Jain, S. Ponda, and S. Venugopalan, "Curie: Evaluating llms on multitask scientific long context understanding and reasoning," arXiv, 2025. [Online]. Available: https://arxiv.org/abs/2503.13517

39.B. Parpillon, C. Syal, J. Yoo, J. Dickinson, M. Swartz, G. D. Guglielmo, A. Bean, D. Berry, M. B. Valentin, K. DiPetrillo, A. Badea, L. Gray, P. Maksimovic, C. Mills, M. S. Neubauer, G. Pradhan, N. Tran, D. Wen, and F. Fahim, "Smart pixels: In-pixel ai for on-sensor data filtering," arXiv, 2024. [Online]. Available: https://arxiv.org/abs/2406.14860

40.T. Aarrestad, E. Govorkova, J. Ngadiuba, E. Puljak, M. Pierini, and K. A. Wozniak, "Unsupervised New Physics detection at 40 MHz: Training Dataset," 2021. [Online]. Available: https://zenodo.org/record/5046389

41.J. Bowles, S. Ahmed, and M. Schuld, "Better than classical? the subtle art of benchmarking quantum machine learning models," arXiv, 2024. [Online]. Available: https://arxiv.org/abs/2403.07059

42.P. Odagiu, Z. Que, J. Duarte, J. Haller, G. Kasieczka, A. Lobanov, V. Loncar, W. Luk, J. Ngadiuba, M. Pierini, P. Rincke, A. Seksaria, S. Summers, A. Sznajder, A. Tapper, and T. K. Aarrestad, "Ultrafast jet classification on fpgas for the hl-lhc," arXiv, 2024. [Online]. Available: https://arxiv.org/abs/2402.01876

43.Z. Liu, H. Sharma, J.-S. Park, P. Kenesei, A. Miceli, J. Almer, R. Kettimuthu, and I. Foster, "Braggnn: Fast x-ray bragg peak analysis using deep learning," arXiv, 2021. [Online]. Available: https://arxiv.org/abs/2008.08198

44.S. Qin, J. Agar, and N. Tran, "Extremely noisy 4d-tem strain mapping using cycle consistent spatial transforming autoencoders," in *AI for Accelerated Materials Design - NeurIPS 2023 Workshop*, 2023. [Online]. Available: https://openreview.net/forum?id=7yt3N0o0W9

45.D. Kafkes and J. S. John, "Boostr: A dataset for accelerator control systems," arXiv, 2021. [Online]. Available: https://arxiv.org/abs/2101.08359

46.J. Kvapil, G. Borca-Tasciuc, H. Bossi, K. Chen, Y. Chen, Y. C. Morales, H. D. Costa, C. D. Silva, C. Dean, J. Durham, S. Fu, C. Hao, P. Harris, O. Hen, H. Jheng, Y. Lee, P. Li, X. Li, Y. Lin, M. X. Liu, V. Loncar, J. P. Mitrevski, A. Olvera, M. L. Purschke, J. S. Renck, G. Roland, J. Schambach, Z. Shi, N. Tran, N. Wuerfel, B. Xu, D. Yu, and H. Zhang, "Intelligent experiments through real-time ai: Fast data processing and autonomous detector control for sphenix and future eic detectors," arXiv, 2025. [Online]. Available: https://arxiv.org/abs/2501.04845

47.A. A. Abud, B. Abi, R. Acciarri, M. A. Acero, G. Adamov, D. Adams, M. Adinolfi, A. Aduszkiewicz, Z. Ahmad, J. Ahmed, T. Alion, S. A. Monsalve, M. Alrashed, C. Alt, A. Alton, P. Amedo, J. Anderson, C. Andreopoulos, M. P. Andrews, F. Andrianala, S. Andringa, N. Anfimov, A. Ankowski, M. Antonova, S. Antusch, A. Aranda-Fernandez,

A. Ariga, L. O. Arnold, M. A. Arroyave, J. Asaadi, A. Aurisano, V. Aushev, D. Autiero, M. Ayala-Torres, F. Azfar, H. Back, J. J. Back, C. Backhouse, P. Baesso, I. Bagaturia, L. Bagby, S. Balasubramanian, P. Baldi, B. Baller, B. Bambah, F. Barao, G. Barenboim, G. J. Barker, W. Barkhouse, C. Barnes, G. Barr, J. B. Monarca, N. Barros, J. L. Barrow, A. Basharina-Freshville, A. Bashyal, V. Basque, E. Belchior, J. B. R. Battat, F. Battisti, F. Bay, J. L. B. Alba, J. F. Beacom, E. Bechetoille, B. Behera, L. Bellantoni, G. Bellettini, V. Bellini, O. Beltramello, D. Belver, N. Benekos, F. B. Neves, S. Berkman, P. Bernardini, R. M. Berner, H. Berns, S. Bertolucci, M. Betancourt, A. B. Rodríguez, M. Bhattacharjee, S. Bhuller, B. Bhuyan, S. Biagi, J. Bian, M. Biassoni, K. Biery, B. Bilki, M. Bishai, A. Bitadze, A. Blake, F. D. M. Blaszczyk, G. C. Blazey, E. Blucher, J. Boissevain, S. Bolognesi, T. Bolton, L. Bomben, M. Bonesini, M. Bongrand, F. Bonini, A. Booth, C. Booth, S. Bordoni, A. Borkum, T. Boschi, N. Bostan, P. Bour, C. Bourgeois, S. B. Boyd, D. Boyden, J. Bracinik, D. Braga, D. Brailsford, A. Brandt, J. Bremer, C. Brew, E. Brianne, S. J. Brice, C. Brizzolari, C. Bromberg, G. Brooijmans, J. Brooke, A. Bross, G. Brunetti, M. Brunetti, N. Buchanan, H. Budd, D. Caiulo, P. Calafiura, J. Calcutt, M. Calin, S. Calvez, E. Calvo, A. Caminata, M. Campanelli, K. Cankocak, D. Caratelli, G. Carini, B. Carlus, P. Carniti, I. C. Terrazas, H. Carranza, T. Carroll, J. F. C. no Forero, A. Castillo, C. Castromonte, E. Catano-Mur, C. Cattadori, F. Cavalier, F. Cavanna, S. Centro, G. Cerati, A. Cervelli, A. C. Villanueva, M. Chalifour, A. Chappell, E. Chardonnet, N. Charitonidis, A. Chatterjee, S. Chattopadhyay, H. Chen, M. Chen, Y. Chen, Z. Chen, D. Cherdack, C. Chi, S. Childress, A. Chiriacescu, G. Chisnall, K. Cho, S. Choate, D. Chokheli, S. Choubey, A. Christensen, D. Christian, G. Christodoulou, A. Chukanov, E. Church, P. Clarke, T. E. Coan, A. G. Cocco, J. A. B. Coelho, E. Conley, R. Conley, J. M. Conrad, M. Convery, S. Copello, L. Corwin, L. Cremaldi, L. Cremonesi, J. I. Crespo-Anadón, E. Cristaldo, R. Cross, A. Cudd, C. Cuesta, Y. Cui, D. Cussans, M. Dabrowski, O. Dalager, H. da Motta, L. D. S. Peres, C. David, Q. David, G. S. Davies, S. Davini, J. Dawson, K. De, R. M. D. Almeida, P. Debbins, I. D. Bonis, M. P. Decowski, A. de Gouvêa, P. C. D. Holanda, I. L. D. I. Astiz, A. Deisting, P. D. Jong, A. Delbart, D. Delepine, M. Delgado, A. Dell'Acqua, P. D. Lurgio, J. R. T. de Mello Neto, D. M. DeMuth, S. Dennis, C. Densham, G. W. Deptuch, A. D. Roeck, V. D. Romeri, G. D. Souza, R. Dharmapalan, F. Diaz, J. S. Díaz, S. D. Domizio, L. D. Giulio, P. Ding, L. D. Noto, C. Distefano, R. Diurba, M. Diwan, Z. Djurcic, N. Dokania, S. Dolan, M. J. Dolinski, L. Domine, D. Douglas, D. Douillet, G. Drake, F. Drielsma, D. Duchesneau, K. Duffy, P. Dunne, T. Durkin, H. Duyang, O. Dvornikov, D. A. Dwyer, A. S. Dyshkant, M. Eads, A. Earle, D. Edmunds, J. Eisch, L. Emberger, S. Emery, A. Ereditato, C. O. Escobar, G. Eurin, J. J. Evans, E. Ewart, A. C. Ezeribe, K. Fahey, A. Falcone, C. Farnese, Y. Farzan, J. Felix, M. F. C. da Silva, E. Fernandez-Martinez, P. F. Menendez, F. Ferraro, L. Fields, F. Filthaut, A. Fiorentini, R. S. Fitzpatrick, W. Flanagan, B. Fleming, R. Flight, D. V. Forero, J. Fowler, W. Fox, J. Franc, K. Francis, D. Franco, J. Freeman, J. Freestone, J. Fried, A. Friedland, S. Fuess, I. Furic, A. P. Furmanski, A. Gago, H. Gallagher, A. Gallas, A. Gallego-Ros, N. Gallice, V. Galymov, E. Gamberini, T. Gamble, R. Gandhi, R. Gandrajula, F. Gao, S. Gao, D. Garcia-Gamez, M. A. García-Peris, S. Gardiner, D. Gastler, G. Ge, B. Gelli, A. Gendotti, S. Gent, Z. Ghorbani-Moghaddam, D. Gibin, I. Gil-Botella, S. Gilligan, C. Girerd, A. K. Giri, D. Gnani, O. Gogota, M. Gold, S. Gollapinni, K. Gollwitzer, R. A. Gomes, L. V. G. Bermeo, L. S. G. Fajardo, F. Gonnella, J. A. Gonzalez-Cuevas, D. Gonzalez-Diaz, M. Gonzalez-Lopez, M. C. Goodman, O. Goodwin, S. Goswami, C. Gotti, E. Goudzovski, C. Grace, M. Graham, R. Gran, E. Granados, P. Granger, A. Grant, C. Grant, D. Gratieri,

P. Green, L. Greenler, J. Greer, W. C. Griffith, M. Groh, J. Grudzinski, K. Grzelak, W. Gu, V. Guarino, R. Guenette, E. Guerard, A. Guglielmi, B. Guo, K. K. Guthikonda, R. Gutierrez, P. Guzowski, M. M. Guzzo, S. Gwon, A. Habig, H. Hadavand, R. Haenni, A. Hahn, J. Haiston, P. Hamacher-Baumann, T. Hamernik, P. Hamilton, J. Han, D. A. Harris, J. Hartnell, J. Harton, T. Hasegawa, C. Hasnip, R. Hatcher, K. W. Hatfield, A. Hatzikoutelis, C. Hayes, E. Hazen, A. Heavey, K. M. Heeger, J. Heise, K. Hennessy, S. Henry, M. A. H. Morquecho, K. Herner, L. Hertel, V. Hewes, A. Higuera, T. Hill, S. J. Hillier, A. Himmel, J. Hoff, C. Hohl, A. Holin, E. Hoppe, G. A. Horton-Smith, M. Hostert, A. Hourlier, B. Howard, R. Howell, J. Huang, J. Huang, J. Hugon, G. Iles, N. Ilic, A. M. Iliescu, R. Illingworth, A. Ioannisian, L. Isenhower, R. Itay, A. Izmaylov, S. Jackson, V. Jain, E. James, B. Jargowsky, F. Jediny, D. Jena, Y. S. Jeong, C. Jesús-Valls, X. Ji, L. Jiang, S. Jiménez, A. Jipa, R. Johnson, B. Jones, S. B. Jones, M. Judah, C. K. Jung, T. Junk, Y. Jwa, M. Kabirnezhad, A. Kaboth, I. Kadenko, I. Kakorin, F. Kamiya, N. Kaneshige, G. Karagiorgi, G. Karaman, A. Karcher, M. Karolak, Y. Karyotakis, S. Kasai, S. P. Kasetti, L. Kashur, N. Kazaryan, E. Kearns, P. Keener, K. J. Kelly, E. Kemp, O. Kemularia, W. Ketchum, S. H. Kettell, M. Khabibullin, A. Khotjantsev, A. Khvedelidze, D. Kim, B. King, B. Kirby, M. Kirby, J. Klein, K. Koehler, L. W. Koerner, S. Kohn, P. P. Koller, L. Kolupaeva, M. Kordosky, T. Kosc, U. Kose, V. A. Kostelecký, K. Kothekar, F. Krennrich, I. Kreslo, Y. Kudenko, V. A. Kudryavtsev, S. Kulagin, J. Kumar, P. Kumar, P. Kunze, N. Kurita, C. Kuruppu, V. Kus, T. Kutter, A. Lambert, B. Land, K. Lande, C. E. Lane, K. Lang, T. Langford, J. Larkin, P. Lasorak, D. Last, C. Lastoria, A. Laundrie, A. Lawrence, I. Lazanu, R. LaZur, T. Le, S. Leardini, J. Learned, P. LeBrun, T. LeCompte, G. L. Miotto, R. Lehnert, M. A. L. de Oliveira, M. Leitner, L. Li, S. W. Li, T. Li, Y. Li, H. Liao, C. S. Lin, Q. Lin, S. Lin, A. Lister, B. R. Littlejohn, J. Liu, S. Lockwitz, T. Loew, M. Lokajicek, I. Lomidze, K. Long, K. Loo, D. Lorca, T. Lord, J. M. LoSecco, W. C. Louis, X. G. Lu, K. B. Luk, X. Luo, N. Lurkin, T. Lux, V. P. Luzio, D. MacFarlane, A. A. Machado, P. Machado, C. T. Macias, J. R. Macier, A. Maddalena, A. Madera, P. Madigan, S. Magill, K. Mahn, A. Maio, A. Major, J. A. Maloney, G. Mandrioli, R. C. Mandujano, J. Maneira, L. Manenti, S. Manly, A. Mann, K. Manolopoulos, M. M. Plata, V. N. Manyam, L. Manzanillas, M. Marchan, A. Marchionni, W. Marciano, D. Marfatia, C. Mariani, J. Maricic, R. Marie, F. Marinho, A. D. Marino, D. Marsden, M. Marshak, C. M. Marshall, J. Marshall, J. Marteau, J. Martin-Albo, N. Martinez, D. A. M. Caicedo, S. Martynenko, K. Mason, A. Mastbaum, M. Masud, S. Matsuno, J. Matthews, C. Mauger, N. Mauri, K. Mavrokoridis, I. Mawby, R. Mazza, A. Mazzacane, E. Mazzucato, T. McAskill, E. McCluskey, N. McConkey, K. S. McFarland, C. McGrew, A. McNab, A. Mefodiev, P. Mehta, P. Melas, O. Mena, S. Menary, H. Mendez, D. P. Méndez, A. Menegolli, G. Meng, M. D. Messier, W. Metcalf, T. Mettler, M. Mewes, H. Meyer, T. Miao, G. Michna, T. Miedema, J. Migenda, V. Mikola, R. Milincic, W. Miller, J. Mills, C. Milne, O. Mineev, O. G. Miranda, S. Miryala, C. S. Mishra, S. R. Mishra, A. Mislivec, D. Mladenov, I. Mocioiu, K. Moffat, N. Moggi, R. Mohanta, T. A. Mohayai, N. Mokhov, J. Molina, L. M. Bueno, A. Montanari, C. Montanari, D. Montanari, L. M. M. Zetina, J. Moon, M. Mooney, A. F. Moor, D. Moreno, C. Morris, C. Mossey, E. Motuk, C. A. Moura, J. Mousseau, W. Mu, L. Mualem, J. Mueller, M. Muether, S. Mufson, F. Muheim, A. Muir, M. Mulhearn, D. Munford, H. Muramatsu, S. Murphy, J. Musser, J. Nachtman, S. Nagu, M. Nalbandyan, R. Nandakumar, D. Naples, S. Narita, D. Navas-Nicolás, A. Navrer-Agasson, N. Nayak, M. Nebot-Guinot, K. Negishi, J. K. Nelson, J. Nesbit, M. Nessi, D. Newbold, M. Newcomer, D. Newhart, H. Newton, R. Nichol, F. Nicolas-Arnaldos, E. Niner,

K. Nishimura, A. Norman, A. Norrick, R. Northrop, P. Novella, J. A. Nowak, M. Oberling, J. P. Ochoa-Ricoux, A. O. D. Campo, A. Olivier, A. Olshevskiy, Y. Onel, Y. Onishchuk, J. Ott, L. Pagani, S. Pakvasa, G. Palacio, O. Palamara, S. Palestini, J. M. Paley, M. Pallavicini, C. Palomares, J. L. Palomino-Gallo, E. Pantic, V. Paolone, V. Papadimitriou, R. Papaleo, A. Papanestis, S. Paramesvaran, S. Parke, Z. Parsa, M. Parvu, S. Pascoli, L. Pasqualini, J. Pasternak, J. Pater, C. Patrick, L. Patrizii, R. B. Patterson, S. J. Patton, T. Patzak, A. Paudel, B. Paulos, L. Paulucci, Z. Pavlovic, G. Pawloski, D. Payne, V. Pec, S. J. M. Peeters, E. Pennacchio, A. Penzo, O. L. G. Peres, J. Perry, D. Pershey, G. Pessina, G. Petrillo, C. Petta, R. Petti, F. Piastra, L. Pickering, F. Pietropaolo, R. Plunkett, R. Poling, X. Pons, N. Poonthottathil, S. Pordes, J. Porter, M. Potekhin, R. Potenza, B. V. K. S. Potukuchi, J. Pozimski, M. Pozzato, S. Prakash, T. Prakash, S. Prince, D. Pugnere, X. Qian, M. C. Q. Bazetto, J. L. Raaf, V. Radeka, J. Rademacker, B. Radics, A. Rafique, E. Raguzin, M. Rai, M. Rajaoalisoa, I. Rakhno, A. Rakotonandrasana, L. Rakotondravohitra, Y. A. Ramachers, R. Rameika, M. A. R. Delgado, B. Ramson, A. Rappoldi, G. Raselli, P. Ratoff, S. Raut, R. F. Razakamiandra, J. S. Real, B. Rebel, M. Reggiani-Guzzo, T. Rehak, J. Reichenbacher, S. D. Reitzner, H. R. Sfar, A. Renshaw, S. Rescia, F. Resnati, A. Reynolds, C. Riccio, G. Riccobene, L. C. J. Rice, J. Ricol, A. Rigamonti, Y. Rigaut, D. Rivera, L. Rochester, M. Roda, P. Rodrigues, M. J. R. Alonso, E. R. Bonilla, J. R. Rondon, S. Rosauro-Alcaraz, M. Rosenberg, P. Rosier, B. Roskovec, M. Rossella, J. Rout, P. Roy, S. Roy, A. Rubbia, C. Rubbia, F. C. Rubio, B. Russell, D. Ruterbories, R. Saakyan, S. Sacerdoti, T. Safford, R. Sahay, N. Sahu, P. Sala, N. Samios, O. Samoylov, M. C. Sanchez, D. A. Sanders, D. Sankey, S. Santana, M. Santos-Maldonado, N. Saoulidou, P. Sapienza, C. Sarasty, I. Sarcevic, G. Savage, V. Savinov, A. Scaramelli, A. Scarff, A. Scarpelli, T. Schaffer, H. Schellman, P. Schlabach, D. Schmitz, K. Scholberg, A. Schukraft, E. Segreto, J. Sensenig, I. Seong, A. Sergi, D. Sgalaberna, M. H. Shaevitz, S. Shafaq, M. Shamma, R. Sharankova, H. R. Sharma, R. Sharma, R. Kumar, T. Shaw, C. Shepherd-Themistocleous, S. Shin, D. Shooltz, R. Shrock, L. Simard, F. Simon, N. Simos, J. Sinclair, G. Sinev, J. Singh, J. Singh, V. Singh, R. Sipos, F. W. Sippach, G. Sirri, A. Sitraka, K. Siyeon, K. S. VIII, A. Smith, E. Smith, P. Smith, J. Smolik, M. Smy, E. L. Snider, P. Snopok, M. S. Nunes, H. Sobel, M. Soderberg, C. J. S. Salinas, S. Söldner-Rembold, N. Solomey, V. Solovov, W. E. Sondheim, M. Sorel, J. Soto-Oton, A. Sousa, K. Soustruznik, F. Spagliardi, M. Spanu, J. Spitz, N. J. C. Spooner, K. Spurgeon, R. Staley, M. Stancari, L. Stanco, R. Stanley, R. Stein, H. M. Steiner, J. Stewart, B. Stillwell, J. Stock, F. Stocker, T. Stokes, M. Strait, T. Strauss, S. Striganov, A. Stuart, J. G. Suarez, H. Sullivan, D. Summers, A. Surdo, V. Susic, L. Suter, C. M. Sutera, R. Svoboda, B. Szczerbinska, A. M. Szelc, R. Talaga, H. A. Tanaka, B. T. Oregui, A. Tapper, S. Tariq, E. Tatar, R. Tayloe, A. M. Teklu, M. Tenti, K. Terao, C. A. Ternes, F. Terranova, G. Testera, A. Thea, J. L. Thompson, C. Thorn, S. C. Timm, J. Todd, A. Tonazzo, D. Torbunov, M. Torti, M. Tortola, F. Tortorici, D. Totani, M. Toups, C. Touramanis, J. Trevor, S. Trilov, W. H. Trzaska, Y. T. Tsai, Z. Tsamalaidze, K. V. Tsang, N. Tsverava, S. Tufanli, C. Tull, E. Tyley, M. Tzanov, M. A. Uchida, J. Urheim, T. Usher, S. Uzunyan, M. R. Vagins, P. Vahle, G. A. Valdiviesso, E. Valencia, Z. Vallari, J. W. F. Valle, S. Vallecorsa, R. V. Berg, R. G. V. de Water, F. Varanini, D. Vargas, G. Varner, J. Vasel, S. Vasina, G. Vasseur, N. Vaughan, K. Vaziri, S. Ventura, A. Verdugo, S. Vergani, M. A. Vermeulen, M. Verzocchi, M. Vicenzi, H. V. de Souza, C. Vignoli, C. Vilela, B. Viren, T. Vrba, T. Wachala, A. V. Waldron, M. Wallbank, H. Wang, J. Wang, M. H. L. S. Wang, Y. Wang, Y. Wang, K. Warburton, D. Warner, M. Wascko, D. Waters, A. Watson, P. Weatherly,

A. Weber, M. Weber, H. Wei, A. Weinstein, D. Wenman, M. Wetstein, A. White, L. H. Whitehead, D. Whittington, M. J. Wilking, C. Wilkinson, Z. Williams, F. Wilson, R. J. Wilson, J. Wolcott, T. Wongjirad, A. Wood, K. Wood, E. Worcester, M. Worcester, C. Wret, W. Wu, W. Wu, Y. Xiao, E. Yandel, G. Yang, K. Yang, S. Yang, T. Yang, A. Yankelevich, N. Yershov, K. Yonehara, T. Young, B. Yu, H. Yu, J. Yu, W. Yuan, R. Zaki, J. Zalesak, L. Zambelli, B. Zamorano, A. Zani, L. Zazueta, G. Zeit, G. P. Zeller, J. Zennamo, K. Zeug, C. Zhang, M. Zhao, E. Zhivun, G. Zhu, P. Zilberman, E. D. Zimmerman, M. Zito, S. Zucchelli, J. Zuklin, V. Zutshi, and R. Zwaska, "Deep Underground Neutrino Experiment (DUNE) Near Detector Conceptual Design Report," arXiv, 2021. [Online]. Available: https://arxiv.org/abs/2103.13910

48.E. Glazer, E. Erdil, T. Besiroglu, D. Chicharro, E. Chen, A. Gunning, C. F. Olsson, J.-S. Denain, A. Ho, E. de Oliveira Santos, O. Järviniemi, M. Barnett, R. Sandler, M. Vrzala, J. Sevilla, Q. Ren, E. Pratt, L. Levine, G. Barkley, N. Stewart, B. Grechuk, T. Grechuk, S. V. Enugandla, and M. Wildon, "FrontierMath: A Benchmark for Evaluating Advanced Mathematical Reasoning in AI," arXiv, 2024. [Online]. Available: https://arxiv.org/abs/2411.04872

49.vals.ai, "Public Enterprise LLM Benchmarks: AIME," Mar. 2025, [Online accessed 2025-06-24]. [Online]. Available: https://www.vals.ai/benchmarks/aime

50.M. Khan, S. Krave, V. Marinozzi, J. Ngadiuba, S. Stoynev, and N. Tran, "Benchmarking and interpreting real time quench detection algorithms," in *Fast Machine Learning for Science Conference 2024*. Purdue University, IN: indico.cern.ch, Oct. 2024. [Online]. Available: https://indico.cern.ch/event/1387540/contributions/6153618/attachments/2948441/5182077/fast_ml_magnets_2024_final.pdf

51.A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, and K. A. Persson, "The Materials Project: A Materials Genome Approach," *APL Materials*, vol. 1, no. 1, 2013. [Online]. Available: https://materialsproject.org/

52.Y. Wei, R. F. Forelli, C. Hansen, J. P. Levesque, N. Tran, J. C. Agar, G. D. Guglielmo, M. E. Mauel, and G. A. Navratil, "Low latency optical-based mode tracking with machine learning deployed on fpgas on a tokamak," arXiv, 2024. [Online]. Available: https://arxiv.org/abs/2312.00128

53.M. Abadi *et al.*, "TensorFlow: A System for Large-Scale Machine Learning," *OSDI*, vol. 16, pp. 265–283, 2016, tensorBoard Profiler is a component of the TensorFlow ecosystem.

54.A. Paszke *et al.*, "PyTorch: An Imperative Style, High-Performance Deep Learning Library," *Advances in Neural Information Processing Systems*, vol. 32, 2019, pyTorch Profiler is part of the PyTorch library.

55.Google, *JAX Profiler*, Manual, 2025, accessed: 2025-10-22. [Online]. Available: https://jax.readthedocs.io/en/latest/profiling.html

56.NVIDIA, *NVIDIA Deep Learning Profiler (DLProf)*, Manual, 2025, accessed: 2025-10-22. [Online]. Available: https://developer.nvidia.com/deep-learning-profiler

57.NVIDIA Corporation, "NVIDIA Nsight Systems Documentation," https://developer.nvidia.com/nsight-systems, 2024, accessed: [Current Date].

58.——, "NVIDIA Nsight Compute Documentation," https://docs.nvidia.com/nsight-compute/NsightCompute/index.html, 2024, accessed: [Current Date].

59.——, *NVIDIA CUDA Profiler User's Guide*, current edition number, e.g., 12.0 ed., NVIDIA Corporation, Current Year, e.g., 2024. [Online]. Available: https://docs.nvidia.com/cuda/profiler-users-guide/index.html

60.Intel Corporation, "Intel VTune Profiler Documentation," https://www.intel.com/content/www/us/en/develop/documentation/vtune-help/, 2024, accessed: [Current Date].

61.AMD, "ROCm Documentation," https://rocm.docs.amd.com/en/latest/, 2024, accessed: [Current Date].

62.J. Mellor-Crummey *et al.*, "HPCToolkit: Tools for Performance Analysis of Optimized Parallel Programs," *Concurrency and Computation: Practice and Experience*, vol. 24, no. 6, pp. 680–708, 2012.

63.S. Shende and A. D. Malony, "The TAU Parallel Performance System," in *International Conference on Parallel and Distributed Computing and Systems*, 1998, pp. 489–493.

64.G. Poulakos *et al.*, "Perfetto: A System-Wide Tracing Tool for Modern Applications," in *Proceedings of the 35th International Conference on Software Engineering (ICSE 2023)*, 2023, based on the Perfetto project's scope for system tracing.

65.S. Browne, J. J. Dongarra, N. Garner, K. S. London, and P. Mucci, "A Scalable Cross-Platform Infrastructure for Application Performance Tuning Using Hardware Counters," in *Proceedings Supercomputing 2000, November 4-10, 2000, Dallas, Texas, USA. IEEE Computer Society, CD-ROM*, J. Donnelley, Ed. IEEE Computer Society, 2000, p. 42. [Online]. Available: https://doi.org/10.1109/SC.2000.10029

66.D. Wang *et al.*, "XLA: Optimizing Compiler for TensorFlow," in *Proceedings of the First Workshop on Systems for ML*, 2017, xLA Profiler is a feature of the XLA compiler.

67.M. (Facebook), *TorchDynamo & TorchInductor Debug Tools*, Manual, 2025, accessed: 2025-10-22. [Online]. Available: https://pytorch.org/blog/torchdynamo/

68.OpenAI, "Triton: An Intermediate Language for Tiled Tensor Computations," https://openai.com/research/triton, 2019, the Triton profiler is part of the Triton compiler and language.

69.NVIDIA Corporation, "NVIDIA Collective Communications Library (NCCL) Documentation," https://docs.nvidia.com/deeplearning/nccl/user-guide/docs/index.html, 2024, accessed: [Current Date].

70.A. W. Services, *AWS SageMaker Debugger*, 2025, accessed: 2025-10-22. [Online]. Available: https://docs.aws.amazon.com/sagemaker/latest/dg/debugger.html

71.Microsoft, *Azure Machine Learning Profiler*, 2025, accessed: 2025-10-22. [Online]. Available: https://learn.microsoft.com/en-us/azure/machine-learning/how-to-monitor-profiling

72.W. . Biases, *Weights & Biases*, 2025, accessed: 2025-10-22. [Online]. Available: https://wandb.ai/site

73.C. ML, *Comet ML*, 2025, accessed: 2025-10-22. [Online]. Available: https://www.comet.com/site/

74.Databricks, *MLflow*, 2025, accessed: 2025-10-22. [Online]. Available: https://mlflow.org/

75.M. . Google, *Torch/TensorFlow Memory Tools*, 2025, accessed: 2025-10-22. [Online]. Available: https://pytorch.org/docs/stable/notes/memory.html

76.P. S. Foundation, *Python Profilers (cProfile, py-spy)*, 2025, accessed: 2025-10-22. [Online]. Available: https://docs.python.org/3/library/profile.html