

# MLCommons Science Working Group AI Benchmarks Collection

Gregor von Laszewski, Reece Shiraishi, Anjay Krishnan, Nhan Tran,  
Benjamin Hawks, and Geoffrey C. Fox

October 28, 2025

## Abstract

This document provides an overview of various benchmarks, including their descriptions, URLs, domains, focus areas, keywords, task types, AI capabilities measured, metrics, models, and notes. Each benchmark

## Citation

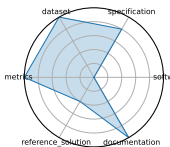
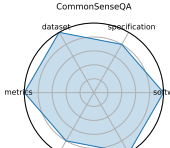
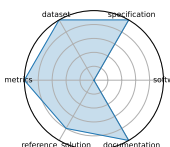
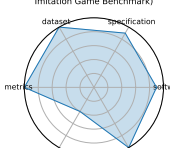
```
@misc{benchmark-collection,  
  title={MLCommons Science Working Group AI Benchmarks Collection}  
  author={Gregor von Laszewski and  
          Reece Shiraishi and  
          Anjay Krishnan and  
          Nhan Tran and  
          Benjamin Hawks and  
          Geoffrey C. Fox}  
  url={https://mlcommons-science.github.io/benchmark/benchmarks.pdf}  
  howpublished={Github},  
  year={2025}  
  month=jul  
}
```

# Contents

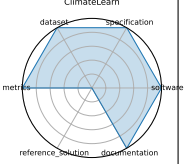
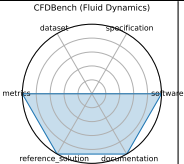
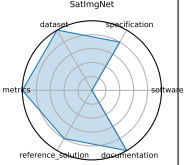
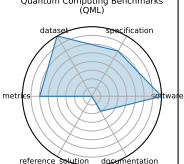
<b>1</b>	<b>Benchmark Overview Table</b>	<b>4</b>
<b>2</b>	<b>Radar Chart Table</b>	<b>30</b>
<b>3</b>	<b>Benchmark Details</b>	<b>33</b>
3.1	MMLU (Massive Multitask Language Understanding)	33
3.2	CommonSenseQA	34
3.3	Winogrande	35
3.4	BIG-Bench (Beyond the Imitation Game Benchmark)	36
3.5	ClimateLearn	37
3.6	CFDBench (Fluid Dynamics)	38
3.7	SatImgNet	39
3.8	Quantum Computing Benchmarks (QML)	40
3.9	OCF (Open Catalyst Project)	41
3.10	MedQA	42
3.11	Materials Project	43
3.12	GPQA Diamond	44
3.13	ARC-Challenge (Advanced Reasoning Challenge)	45
3.14	Humanity’s Last Exam	46
3.15	FrontierMath	47
3.16	SciCode	48
3.17	AIME (American Invitational Mathematics Examination)	49
3.18	MATH-500	50
3.19	CURIE (Scientific Long-Context Understanding, Reasoning and Information Extraction)	51
3.20	FEABench (Finite Element Analysis Benchmark)	52
3.21	SPIQA (Scientific Paper Image Question Answering)	53
3.22	BaisBench (Biological AI Scientist Benchmark)	54
3.23	MOLGEN	55
3.24	Open Graph Benchmark (OGB) - Biology	56
3.25	JARVIS-Leaderboard	57
3.26	Nixtla Neural Forecast TimeGPT	58
3.27	Nixtla Neural Forecast NHITS	59
3.28	Nixtla Neural Forecast TimeLLM	60
3.29	Nixtla NeuralForecast	61
3.30	vLLM Performance Dashboard	62
3.31	vLLM Inference and Serving Engine	63
3.32	SGLang Framework	64
3.33	Jet Classification	65
3.34	Irregular Sensor Data Compression	66
3.35	Beam Control	67
3.36	Ultrafast jet classification at the HL-LHC	68
3.37	Quench detection	69
3.38	Intelligent experiments through real-time AI	70
3.39	Neural Architecture Codesign for Fast Physics Applications	71
3.40	Smart Pixels for LHC	72
3.41	HEDM (BraggNN)	73
3.42	4D-STEM	74
3.43	In-Situ High-Speed Computer Vision	75
3.44	BenchCouncil AIBench	76
3.45	BenchCouncil BigDataBench	77

3.46	MLPerf HPC . . . . .	78
3.47	MLCommons Science . . . . .	79
3.48	LHC New Physics Dataset . . . . .	80
3.49	MLCommons Medical AI . . . . .	81
3.50	CaloChallenge 2022 . . . . .	82
3.51	Papers With Code (SOTA Platform) . . . . .	83
3.52	Codabench . . . . .	84
3.53	Sabath (SBI-FAIR) . . . . .	85
3.54	PDEBench . . . . .	86
3.55	LLM-Inference-Bench . . . . .	87
3.56	Delta Squared-DFT . . . . .	88
3.57	Urban Data Layer (UDL) . . . . .	89
3.58	LLMs for Crop Science . . . . .	90
3.59	DUNE . . . . .	91
3.60	MassSpecGym . . . . .	92
3.61	HDR ML Anomaly Challenge (Gravitational Waves) . . . . .	93
3.62	HDR ML Anomaly Challenge (Butterfly) . . . . .	94
3.63	HDR ML Anomaly Challenge (Sea Level Rise) . . . . .	95
3.64	Single Qubit Readout on QICK System . . . . .	96
3.65	GPQA: A Graduate-Level Google-Proof Question and Answer Benchmark . . . . .	97
3.66	SeafloorAI . . . . .	98
3.67	SuperCon3D . . . . .	99
3.68	GeSS . . . . .	100
3.69	Vocal Call Locator (VCL) . . . . .	101
3.70	SPIQA (LLM) . . . . .	102
3.71	The Well . . . . .	103

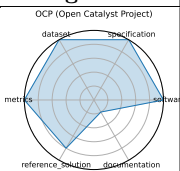
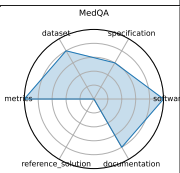
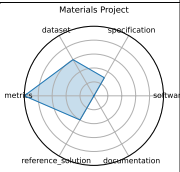
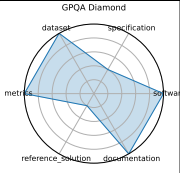
## 1 Benchmark Overview Table

Ratings	Name	Domain	Focus	Keywords	Task Types	AI Capability	Metrics	Models	Citation
	MMLU (Massive Multitask Language Understanding)	Multidomain	Academic knowledge and reasoning across 57 subjects	multitask, multiple-choice, zero-shot, few-shot, knowledge probing	Multiple choice	General reasoning, subject-matter understanding	Accuracy	GPT-4o, Gemini 1.5 Pro, o1, DeepSeek-R1	[1]⇒
	CommonSenseQA	NLP; Commonsense	Commonsense question answering	ConceptNet, multiple-choice, adversarial	Multiple choice	Commonsense reasoning and knowledge integration	Accuracy	BERT-large, RoBERTa, GPT-3	[2]⇒
	Winogrande	NLP; Commonsense	Winograd Schema-style pronoun resolution	adversarial, pronoun resolution	Pronoun resolution	Robust commonsense reasoning	Accuracy, AUC	RoBERTa, BERT, GPT-2	[3]⇒
	BIG-Bench (Beyond the Imitation Game Benchmark)	NLP; AI Evaluation	Diverse reasoning and generalization tasks	few-shot, multi-task, bias analysis	Few-shot evaluation, Multi-task evaluation	Reasoning and generalization across diverse tasks	Accuracy, Task-specific metrics	GPT-3, Dense Transformers, Sparse Transformers	[4]⇒

Continued on next page

Ratings	Name	Domain	Focus	Keywords	Task Types	AI Capability	Metrics	Models	Citation
	ClimateLearn	Climate Science; Forecasting	ML for weather and climate modeling	medium-range forecasting, ERA5, data-driven	Forecasting	Global weather prediction (3-5 days)	RMSE, Anomaly correlation	CNN baselines, ResNet variants	[5]⇒
	CFDBench (Fluid Dynamics)	Fluid Dynamics; Scientific ML	Neural operator surrogate modeling	neural operators, CFD, FNO, DeepONet	Surrogate modeling	Generalization of neural operators for PDEs	L2 error, MAE	FNO, DeepONet, U-Net	[6]⇒
	SatImNet	Remote Sensing	Satellite imagery classification	land-use, zero-shot, multi-task	Image classification	Zero-shot land-use classification	Accuracy	CLIP, BLIP, ALBEF	[7]⇒
	Quantum Computing Benchmarks (QML)	Quantum Computing	Quantum algorithm performance evaluation	quantum circuits, state preparation, error correction	Circuit benchmarking, State classification	Quantum algorithm performance and fidelity	Fidelity, Success probability	IBM Q, IonQ, AQT@LBNL	[8]⇒

Continued on next page

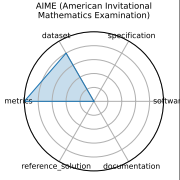
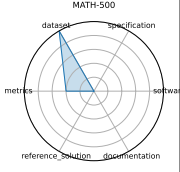
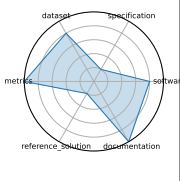
Ratings	Name	Domain	Focus	Keywords	Task Types	AI Capability	Metrics	Models	Citation
	OCP (Open Catalyst Project)	Chemistry; Materials Science	Catalyst adsorption energy prediction	DFT relaxations, adsorption energy, graph neural networks	Energy prediction, Force prediction	Prediction of adsorption energies and forces	MAE (energy), MAE (force)	CGCNN, SchNet, DimeNet++, GemNet-OC	[9]–[12]⇒
	MedQA	Medical Question Answering	Medical board exam QA	USMLE, diagnostic QA, medical knowledge, multilingual	Multiple choice	Medical diagnosis and knowledge retrieval	Accuracy	Neural reader, Retrieval-based QA systems	[13]⇒
	Materials Project	Materials Science	DFT-based property prediction	DFT, materials genome, high-throughput	Property prediction	Prediction of inorganic material properties	MAE, $R^2$	Automatminer[14]⇒ Crystal Graph Neural Networks	
	GPQA Diamond	Science	Graduate-level scientific reasoning	Google-proof, graduate-level, science QA, chemistry, physics	Multiple choice, Multi-step QA	Scientific reasoning, deep knowledge	Accuracy	o1, DeepSeek-R1	[15]⇒

Continued on next page

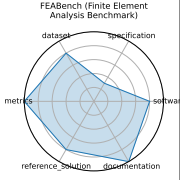
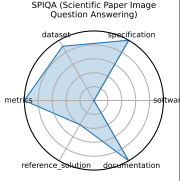
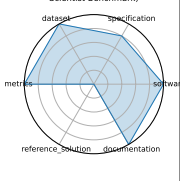
Ratings	Name	Domain	Focus	Keywords	Task Types	AI Capability	Metrics	Models	Citation
	ARC-Challenge (Advanced Reasoning Challenge)	Science	Grade-school science with reasoning emphasis	grade-school, science QA, challenge set, reasoning	Multiple choice	Commonsense and scientific reasoning	Accuracy	GPT-4, Claude	[16]⇒
	Humanity's Last Exam	Multidomain	Broad cross-domain academic reasoning	cross-domain, academic exam, multiple-choice, multi-disciplinary	Multiple choice	Cross-domain academic reasoning	Accuracy	unknown	[17]⇒
	FrontierMath	Mathematics	Challenging advanced mathematical reasoning	symbolic reasoning, number theory, algebraic geometry, category theory	Problem solving	Symbolic and abstract mathematical reasoning	Accuracy	unknown	[18]⇒
	SciCode	Scientific Programming	Scientific code generation and problem solving	code synthesis, scientific computing, programming benchmark	Coding	Program synthesis, scientific computing	Solve rate (%)	Claude3.5-Sonnet	[19]⇒

Continued on next page

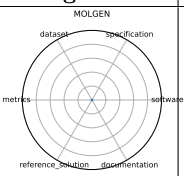
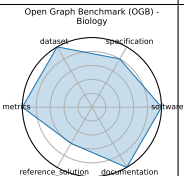
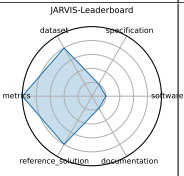


Ratings	Name	Domain	Focus	Keywords	Task Types	AI Capability	Metrics	Models	Citation
	AIME (American Invitational Mathematics Examination)	Mathematics	Pre-college advanced problem solving	algebra, combinatorics, number theory, geometry	Problem solving	Mathematical problem-solving and reasoning	Accuracy	unkown	[20]⇒
	MATH-500	Mathematics	Math reasoning generalization	calculus, algebra, number theory, geometry	Problem solving	Math reasoning and generalization	Accuracy	unkown	[21]⇒
	CURIE (Scientific Long-Context Understanding, Reasoning and Information Extraction)	Multidomain Science	Long-context scientific reasoning	long-context, information extraction, multimodal	Information extraction, Reasoning, Concept tracking, Aggregation, Algebraic manipulation, Multimodal comprehension	Long-context understanding and scientific reasoning	Accuracy	unkown	[22]⇒

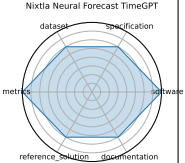
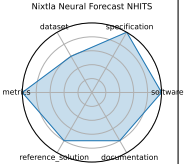
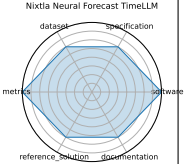
Continued on next page

Ratings	Name	Domain	Focus	Keywords	Task Types	AI Capability	Metrics	Models	Citation
	FEABench (Finite Element Analysis Benchmark)	Computational Engineering	FEA simulation accuracy and performance	finite element, simulation, PDE	Simulation, Performance evaluation	Numerical simulation accuracy and efficiency	Solve time, Error norm	FEniCS, deal.II	[23]⇒
	SPIQA (Scientific Paper Image Question Answering)	Computer Science	Multimodal QA on scientific figures	multimodal QA, figure understanding, table comprehension, chain-of-thought	Question answering, Multimodal QA, Chain-of-Thought evaluation	Visual-textual reasoning in scientific contexts	Accuracy, F1 score	Chain-of-Thought models, Multimodal QA systems	[24]⇒
	BaisBench (Biological AI Scientist Benchmark)	Computational Biology	Omics-driven AI research tasks	single-cell annotation, biological QA, autonomous discovery	Cell type annotation, Multiple choice	Autonomous biological research capabilities	Annotation accuracy, QA accuracy	LLM-based AI scientist agents	[25]⇒

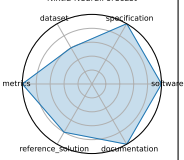
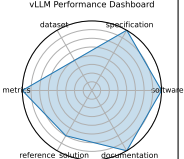
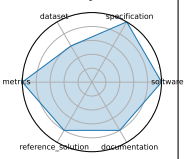
Continued on next page

Ratings	Name	Domain	Focus	Keywords	Task Types	AI Capability	Metrics	Models	Citation
	MOLGEN	Computational Chemistry	Molecular generation and optimization	SELFIES, GAN, property optimization	Distribution learning, Goal-oriented generation	Generation of valid and optimized molecular structures	Validity%, Novelty%, QED, Docking score	MolGen	[26]⇒
	Open Graph Benchmark (OGB) - Biology	Graph ML	Biological graph property prediction	node prediction, link prediction, graph classification	Node property prediction, Link property prediction, Graph property prediction	Scalability and generalization in graph ML for biology	Accuracy, ROC-AUC	GCN, GraphSAGE, GAT	[27]⇒
	JARVIS-Leaderboard	Materials Science; Benchmarking	Comparative evaluation of materials design methods	leaderboards, materials methods, simulation	Method benchmarking, Leaderboard ranking	Performance comparison across diverse materials design methods	MAE, RMSE, Accuracy	unknown	[28]⇒

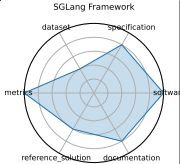
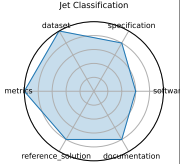
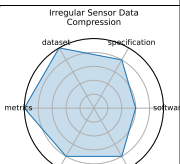
Continued on next page

Ratings	Name	Domain	Focus	Keywords	Task Types	AI Capability	Metrics	Models	Citation
	Nixtla Neural Forecast TimeGPT	Time-series; General ML	Time-series foundation model "TimeGPT" for forecasting and anomaly detection	TimeGPT, foundation model, time-series, generative model	Time-series forecasting, Anomaly detection	Zero-shot forecasting, anomaly detection	RMSE, Anomaly detection metrics	TimeGPT	[29]⇒
	Nixtla Neural Forecast NHITS	Time-series; General ML	Official NHITS implementation for long-horizon time series forecasting	NHITS, long-horizon forecasting, neural interpolation, time-series	Time-series forecasting	Accuracy, compute efficiency for long series	RMSE, MAPE	NHITS	[30]⇒
	Nixtla Neural Forecast TimeLLM	Time-series; General ML	Reprogramming LLMs for time series forecasting	Time-LLM, language model, time-series, reprogramming	Time-series forecasting	Model reuse via LLM, few-shot forecasting	RMSE, MAPE	Time-LLM	[31]⇒

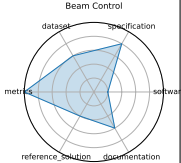
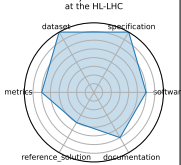
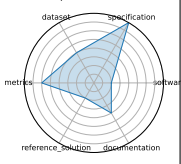
Continued on next page

Ratings	Name	Domain	Focus	Keywords	Task Types	AI Capability	Metrics	Models	Citation
	Nixtla NeuralForecast	Time-series forecasting; General ML	High-performance neural forecasting library with >30 models	time-series, neural forecasting, NBEATS, NHITS, TFT, probabilistic forecasting, usability	Time-series forecasting	Forecast accuracy, interpretability, speed	RMSE, MAPE, CRPS	NBEATS, NHITS, TFT, DeepAR	[32]⇒
	vLLM Performance Dashboard	LLM; HPC/inference	Interactive dashboard showing inference performance of vLLM	Dashboard, Throughput visualization, Latency analysis, Metric tracking	Performance visualization	Throughput, latency, hardware utilization	Tokens/sec, TTFT, Memory usage	LLaMA-2, Mistral, Qwen	[33]⇒
	vLLM Inference and Serving Engine	LLM; HPC/inference	High-throughput, memory-efficient inference and serving engine for LLMs	LLM inference, PageAttention, CUDA graph, streaming API, quantization	Inference Benchmarking	Throughput, latency, memory efficiency	Tokens/sec, Time to First Token (TTFT), Memory footprint	LLaMA, Mistral, FlashAttention-based models	[34]⇒

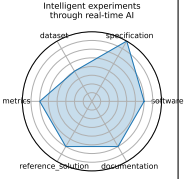
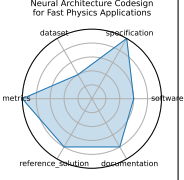
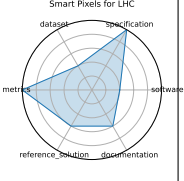
Continued on next page

Ratings	Name	Domain	Focus	Keywords	Task Types	AI Capability	Metrics	Models	Citation
	SGLang Framework	LLM Vi-sion	Fast serv-ing frame-work for LLMs and vision-language models	LLM serv-ing, vision-language, RadixAtten-tion, perfor-mance, JSON decoding	Model serving framework	Serving throughput, JSON/task-specific latency	Tokens/sec, Time-to-first-token, Through-put gain vs baseline	LLaVA, DeepSeek, Llama	[35]⇒
	Jet Classifica-tion	Particle Physics	Real-time classi-fication of particle jets using HL-LHC simulation features	classification, real-time ML, jet tagging, QKeras	Classification	Real-time in-ference, model compression performance	Accuracy, AUC	Keras DNN, QKeras quantized DNN	[36]⇒
	Irregular Sensor Data Compression	Particle Physics	Real-time compres-sion of sparse sensor data with auto-encoders	compression, autoencoder, sparse data, irregular sampling	Compression	Reconstruction quality, com-pression effi-ciency	MSE, Compres-sion ratio	Autoencoder, Quantized autoen-coder	[37]⇒

Continued on next page

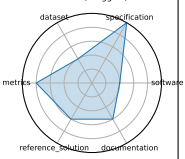
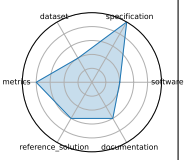
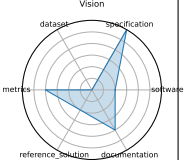
Ratings	Name	Domain	Focus	Keywords	Task Types	AI Capability	Metrics	Models	Citation
	Beam Control	Accelerators and Magnets	Reinforcement learning control of accelerator beam position	RL, beam stabilization, control systems, simulation	Control	Policy performance in simulated accelerator control	Stability, Control loss	DDPG, PPO (planned)	[38], [39]⇒
	Ultrafast jet classification at the HL-LHC	Particle Physics	FPGA-optimized real-time jet origin classification at the HL-LHC	jet classification, FPGA, quantization-aware training, Deep Sets, Interaction Networks	Classification	Real-time inference under constraints	Accuracy, Latency, Resource utilization	MLP, Deep Sets, Interaction Network	[40]⇒
	Quench detection	Accelerators and Magnets	Real-time detection of superconducting magnet quenches using ML	quench detection, autoencoder, anomaly detection, real-time	Anomaly detection, Quench localization	Real-time anomaly detection with multi-modal sensors	ROC-AUC, Detection latency	Autoencoder, RL agents (in development)	[41]⇒

Continued on next page

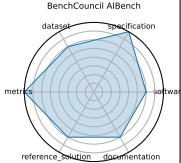
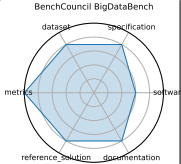
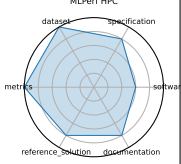
Ratings	Name	Domain	Focus	Keywords	Task Types	AI Capability	Metrics	Models	Citation
	Intelligent experiments through real-time AI	Instrumentation and Detectors; Nuclear Physics; Particle Physics	Real-time FPGA-based triggering and detector control for sPHENIX and future EIC	FPGA, Graph Neural Network, hls4ml, real-time inference, detector control	Trigger classification, Detector control, Real-time inference	Low-latency GNN inference on FPGA	Accuracy (charm and beauty detection), Latency (micros), Resource utilization (LUT/FF/BRAM/DSP)	Bipartite Graph Network with Set Transformers (BGN-ST), GarNet (edge-ADs/DSP)	[42]⇒
	Neural Architecture Codesign for Fast Physics Applications	Physics; Materials Science; Particle Physics	Automated neural architecture search and hardware-efficient model codesign for fast physics applications	neural architecture search, FPGA deployment, quantization, pruning, hls4ml	Classification, Peak finding	Hardware-aware model optimization; low-latency inference	Accuracy, Latency, Resource utilization	NAC-based BraggNN, NAC-optimized Deep Sets (jet)	[43]⇒
	Smart Pixels for LHC	Particle Physics; Instrumentation and Detectors	On-sensor, in-pixel ML filtering for high-rate LHC pixel detectors	smart pixel, on-sensor inference, data reduction, trigger	Image Classification, Data filtering	On-chip, low-power inference; data reduction	Data rejection rate, Power per pixel	2-layer pixel NN	[44]⇒

Continued on next page

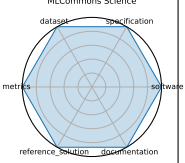
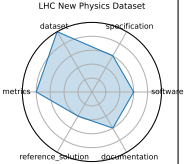


Ratings	Name	Domain	Focus	Keywords	Task Types	AI Capability	Metrics	Models	Citation
	HEDM (BraggNN)	Material Science	Fast Bragg peak analysis using deep learning in diffraction microscopy	BraggNN, diffraction, peak finding, HEDM	Peak detection	High-throughput peak localization	Localization accuracy, Inference time	BraggNN	[45]⇒
	4D-STEM	Material Science	Real-time ML for scanning transmission electron microscopy	4D-STEM, electron microscopy, real-time, image processing	Image Classification, Streamed data inference	Real-time large-scale microscopy inference	Classification accuracy, Throughput	CNN models (prototype)	[46]⇒
	In-Situ High-Speed Computer Vision	Fusion/Plasma	Real-time image classification for in-situ plasma diagnostics	plasma, in-situ vision, real-time ML	Image Classification	Real-time diagnostic inference	Accuracy, FPS	CNN	[47]⇒

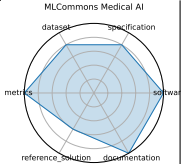
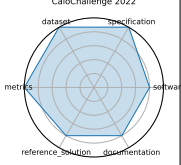
Continued on next page

Ratings	Name	Domain	Focus	Keywords	Task Types	AI Capability	Metrics	Models	Citation
	BenchCouncil AIBench	General	End-to-end AI benchmarking across micro, component, and application levels	benchmarking, AI systems, application-level evaluation	Training, Inference, End-to-end workloads	System-level AI workload performance	Throughput, Latency, Accuracy	ResNet, BERT, GANs, Recommendation systems	[48]⇒
	BenchCouncil Big-DataBench	General	Big data and AI benchmarking across structured, semi-structured, and unstructured data workloads	big data, AI benchmarking, data analytics	Data processing, Inference, End-to-end pipelines	Data processing and AI model inference performance at scale	Data throughput, Latency, Accuracy	CNN, LSTM, SVM, XGBoost	[49]⇒
	MLPerf HPC	Cosmology, Climate, Protein Structure, Catalysis	Scientific ML training and inference on HPC systems	HPC, training, inference, scientific ML	Training, Inference	Scaling efficiency, training time, model accuracy on HPC	Training time, Accuracy, GPU utilization	CosmoFlow, DeepCAM, OpenCatalyst	[50]⇒

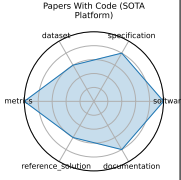
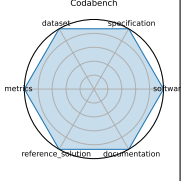
Continued on next page

Ratings	Name	Domain	Focus	Keywords	Task Types	AI Capability	Metrics	Models	Citation
	MLCommons Science	Earthquake, Satellite Image, Drug Discovery, Electron Microscope, CFD	AI benchmarks for scientific applications including time-series, imaging, and simulation	science AI, benchmark, MLCommons, HPC	Time-series analysis, Image classification, Simulation surrogate modeling	Inference accuracy, simulation speed-up, generalization	MAE, Accuracy, Speedup vs simulation	CNN, GNN, Transformer	[51]⇒
	LHC New Physics Dataset	Particle Physics; Real-time Triggering	Real-time LHC event filtering for anomaly detection using proton collision data	anomaly detection, proton collision, real-time inference, event filtering, unsupervised ML	Anomaly detection, Event classification	Unsupervised signal detection under latency and bandwidth constraints	ROC-AUC, Detection efficiency	Autoencoder, Variational autoencoder, Isolation forest	[52]⇒

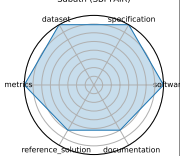
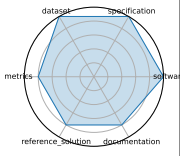
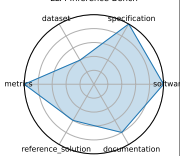
Continued on next page

Ratings	Name	Domain	Focus	Keywords	Task Types	AI Capability	Metrics	Models	Citation
	MLCommons Medical AI	Healthcare; Medical AI	Federated benchmarking and evaluation of medical AI models across diverse real-world clinical data	medical AI, federated evaluation, privacy-preserving, fairness, healthcare benchmarks	Federated evaluation, Model validation	Clinical accuracy, fairness, generalizability, privacy compliance	ROC AUC, Accuracy, Fairness metrics	MedPerf-validated CNNs, GaNDLF workflows	[53]⇒
	CaloChallenge 2022	LHC Calorimeter; Particle Physics	Fast generative-model-based calorimeter shower simulation evaluation	calorimeter simulation, generative models, surrogate modeling, LHC, fast simulation	Surrogate modeling	Simulation fidelity, speed, efficiency	Histogram similarity, Classifier AUC, Generation latency	VAE variants, GAN variants, Normalizing flows, Diffusion models	[54]⇒


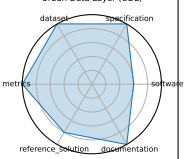
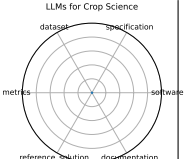
Continued on next page

Ratings	Name	Domain	Focus	Keywords	Task Types	AI Capability	Metrics	Models	Citation
	Papers With Code (SOTA Platform)	General ML; All domains	Open platform tracking state-of-the-art results, benchmarks, and implementations across ML tasks and papers	leaderboard, benchmarking, reproducibility, open-source	Multiple (Classification, Detection, NLP, etc.)	Model performance across tasks (accuracy, F1, BLEU, etc.)	Task-specific (Accuracy, F1, BLEU, etc.)	All published models with code	[55]⇒
	Codabench	General ML; Multiple	Open-source platform for organizing reproducible AI benchmarks and competitions	benchmark platform, code submission, competitions, meta-benchmark	Multiple	Model reproducibility, performance across datasets	Submission count, Leaderboard ranking, Task-specific metrics	Arbitrary code submissions	[56]⇒

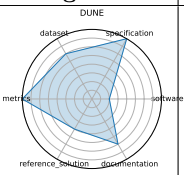
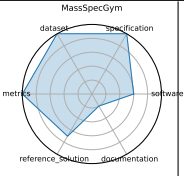
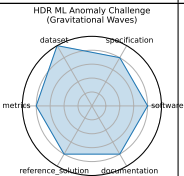
Continued on next page

Ratings	Name	Domain	Focus	Keywords	Task Types	AI Capability	Metrics	Models	Citation
	Sabath (SBI-FAIR)	Systems; Metadata	FAIR metadata framework for ML-driven surrogate workflows in HPC systems	meta-benchmark, metadata, HPC, surrogate modeling	Systems benchmarking	Metadata tracking, reproducible HPC workflows	Metadata completeness, FAIR compliance	NA	[57]⇒
	PDEBench	CFD; Weather Modeling	Benchmark suite for ML-based surrogates solving time-dependent PDEs	PDEs, CFD, scientific ML, surrogate modeling, NeurIPS	Supervised Learning	Time-dependent PDE modeling; physical accuracy	RMSE, boundary RMSE, Fourier RMSE	FNO, U-Net, PINN, Gradient-Based inverse methods	[58]⇒
	LLM-Inference-Bench	LLM; HPC/inference	Hardware performance benchmarking of LLMs on AI accelerators	LLM, inference benchmarking, GPU, accelerator, throughput	Inference Benchmarking	Inference throughput, latency, hardware utilization	Token throughput (tok/s), Latency, Framework-hardware mix performance	LLaMA-2-7B, LLaMA-2-70B, Mistral-7B, Qwen-7B	[59]⇒

Continued on next page

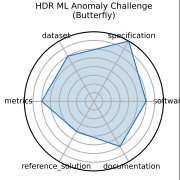
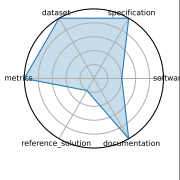
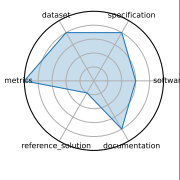
Ratings	Name	Domain	Focus	Keywords	Task Types	AI Capability	Metrics	Models	Citation
	Delta Squared-DFT	Computational Chemistry; Materials Science	Benchmarking machine-learning corrections to DFT using Delta Squared-trained models for reaction energies	density functional theory, Delta Squared-ML correction, reaction energetics, quantum chemistry	Regression	High-accuracy energy prediction, DFT correction	Mean Absolute Error (eV), Energy ranking accuracy	Delta Squared-ML correction networks, Kernel ridge regression	[60]⇒
	Urban Data Layer (UDL)	Urban Computing; Data Engineering	Unified data pipeline for multi-modal urban science research	data pipeline, urban science, multi-modal, benchmark	Prediction, Classification	Multi-modal urban inference, standardization	Task-specific accuracy or RMSE	Baseline regression/classification pipelines	[61]⇒
	LLMs for Crop Science	Agricultural Science; NLP	Evaluating LLMs on crop trait QA and textual inference tasks with domain-specific prompts	crop science, prompt engineering, domain adaptation, question answering	Question Answering, Inference	Scientific knowledge, crop reasoning	Accuracy, F1 score	GPT-4, LLaMA-2-13B, T5-XXL	[62]⇒

Continued on next page

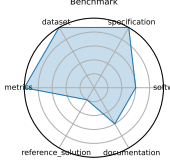
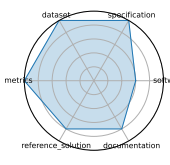
Ratings	Name	Domain	Focus	Keywords	Task Types	AI Capability	Metrics	Models	Citation
	DUNE	Particle Physics	Real-time ML for DUNE DAQ time-series data	DUNE, time-series, real-time, trigger	Trigger selection, Time-series anomaly detection	Low-latency event detection	Detection efficiency, Latency	CNN, LSTM (planned)	[63]⇒
	MassSpecGym	Cheminformatics, Molecular Discovery	Benchmark suite for discovery and identification of molecules via MS/MS	mass spectrometry, molecular structure, de novo generation, retrieval, dataset	De novo generation, Retrieval, Simulation	Molecular identification and generation from spectral data	Structure accuracy, Retrieval precision, Simulation MSE	Graph-based generative models, Retrieval baselines	[64]⇒
	HDR ML Anomaly Challenge (Gravitational Waves)	Astrophysics, Time-series	Detecting anomalous gravitational wave signals from LIGO/Virgo datasets	anomaly detection, gravitational waves, astrophysics, time-series	Anomaly detection	Novel event detection in physical signals	ROC-AUC, Precision/Recall	Deep latent CNNs, Autoencoders	[65]⇒

Continued on next page

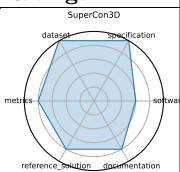
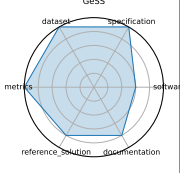


Ratings	Name	Domain	Focus	Keywords	Task Types	AI Capability	Metrics	Models	Citation
	HDR ML Anomaly Challenge (Butterfly)	Genomics; Image/CV	Detecting hybrid butterflies via image anomaly detection in genomic-informed dataset	anomaly detection, computer vision, genomics, butterfly hybrids	Anomaly detection	Hybrid detection in biological systems	Classification accuracy, F1 score	CNN-based detectors	[66]⇒
	HDR ML Anomaly Challenge (Sea Level Rise)	Climate Science; Time-series, Image/CV	Detecting anomalous sea-level rise and flooding events via time-series and satellite imagery	anomaly detection, climate science, sea-level rise, time-series, remote sensing	Anomaly detection	Detection of environmental anomalies	ROC-AUC, Precision/Recall	CNNs, RNNs, Transformers	[67]⇒
	Single Qubit Readout on QICK System	Quantum Computing	Real-time single-qubit state classification using FPGA firmware	qubit readout, hls4ml, FPGA, QICK	Classification	Single-shot fidelity, inference latency	Accuracy, Latency	hls4ml quantized NN	[68]⇒

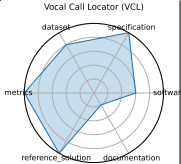
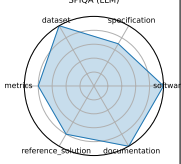
Continued on next page

Ratings	Name	Domain	Focus	Keywords	Task Types	AI Capability	Metrics	Models	Citation
 <p>GPQA: A Graduate-Level Google-Proof Question and Answer Benchmark</p>	GPQA: A Graduate-Level Google-Proof Question and Answer Benchmark	Science (Biology, Physics, Chemistry)	Graduate-level, expert-validated multiple-choice questions hard even with web access	Google-proof, multiple-choice, expert reasoning, science QA	Multiple choice	Scientific reasoning, knowledge probing	Accuracy	GPT-4 baseline	[69]⇒
 <p>SeaFloorAI</p>	SeaFloorAI	Marine Science; Vision-Language	Large-scale vision-language dataset for seafloor mapping and geological classification	sonar imagery, vision-language, seafloor mapping, segmentation, QA	Image segmentation, Vision-language QA	Geospatial understanding, multimodal reasoning	Segmentation pixel accuracy, QA accuracy	SegFormer, ViLT-style multi-modal models	[70]⇒

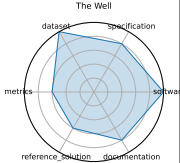
Continued on next page

Ratings	Name	Domain	Focus	Keywords	Task Types	AI Capability	Metrics	Models	Citation
	SuperCon3D	Materials Science; Superconductivity	Dataset and models for predicting and generating high-Tc superconductors using 3D crystal structures	superconductivity, crystal structures, equivariant GNN, generative models	Regression (Tc prediction), Generative modeling	Structure-to-property prediction, structure generation	MAE (Tc), Validity of generated structures	SODNet, DiffCSP-SC	[71]⇒
	GeSS	Scientific ML; Geometric Deep Learning	Benchmark suite evaluating geometric deep learning models under real-world distribution shifts	geometric deep learning, distribution shift, OOD robustness, scientific applications	Classification, Regression	OOD performance in scientific settings	Accuracy, RMSE, OOD robustness delta	GCN, EGNN, DimeNet++	[72]⇒

Continued on next page

Ratings	Name	Domain	Focus	Keywords	Task Types	AI Capability	Metrics	Models	Citation
	Vocal Call Locator (VCL)	Neuroscience; Bioacoustics	Benchmarking source localization of rodent vocalizations from multi-channel audio	source localization, bioacoustics, time-series, SSL	Sound source localization	Source localization accuracy in bioacoustic settings	Localization error (cm), Recall/Precision	CNN-based SSL models	[73]⇒
	SPIQA (LLM)	Multimodal Scientific QA; Computer Vision	Evaluating LLMs on image-based scientific paper figure QA tasks (LLM Adapter performance)	multimodal QA, scientific figures, image+text, chain-of-thought prompting	Multimodal QA	Visual reasoning, scientific figure understanding	Accuracy, F1 score	LLaVA, MiniGPT-4, Owl-LLM adapter variants	[74]⇒

Continued on next page

Ratings	Name	Domain	Focus	Keywords	Task Types	AI Capability	Metrics	Models	Citation
	The Well	biological systems, fluid dynamics, acoustic scattering, astro-physical MHD	Foundation model + surrogate dataset spanning 16 physical simulation domains	surrogate modeling, foundation model, physics simulations, spatiotemporal dynamics	Supervised Learning	Surrogate modeling, physics-based prediction	Dataset size, Domain breadth	FNO baselines, U-Net baselines	[75]⇒

## 2 Radar Chart Table

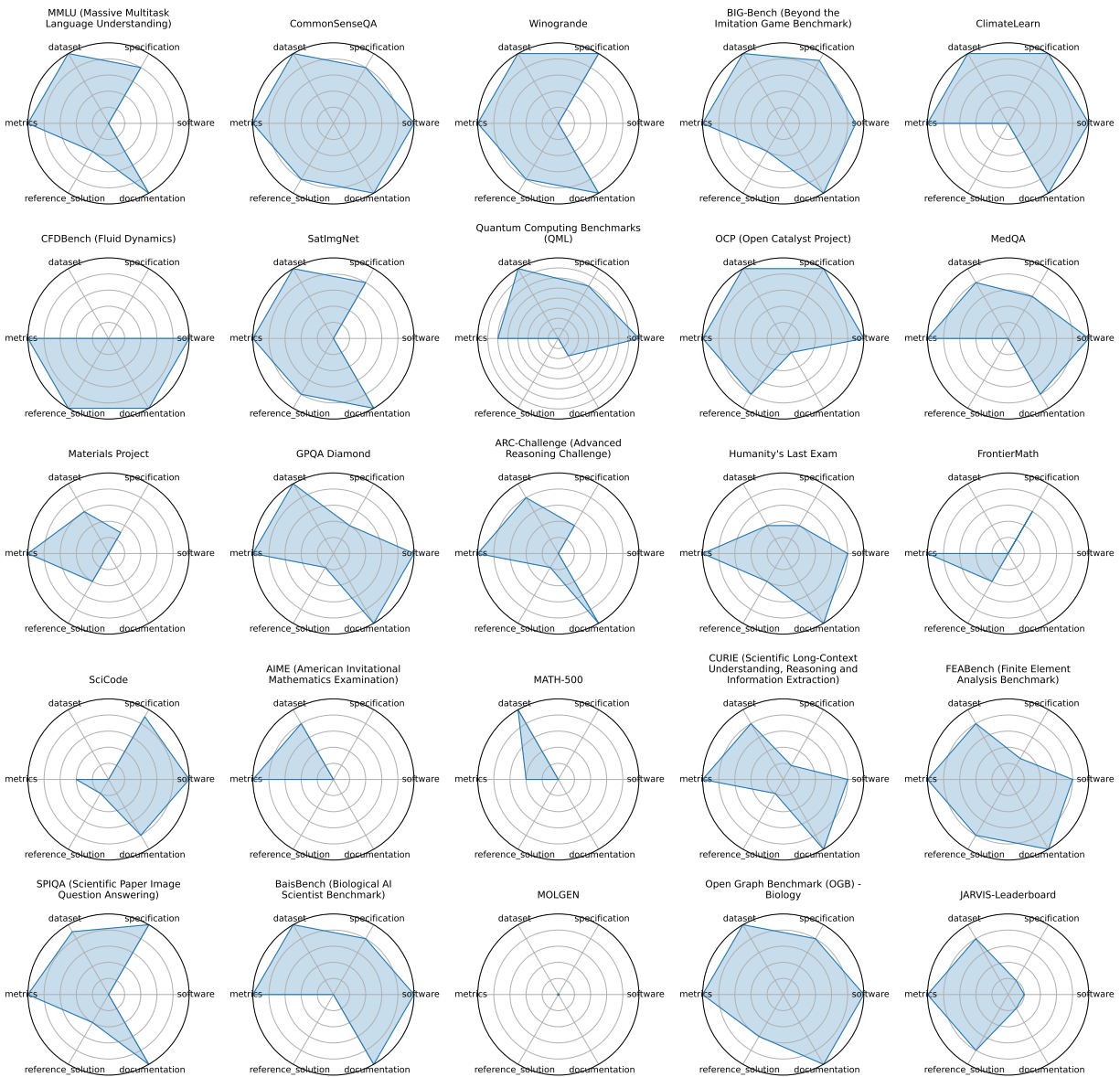


Figure 1: Radar chart overview (page 1)

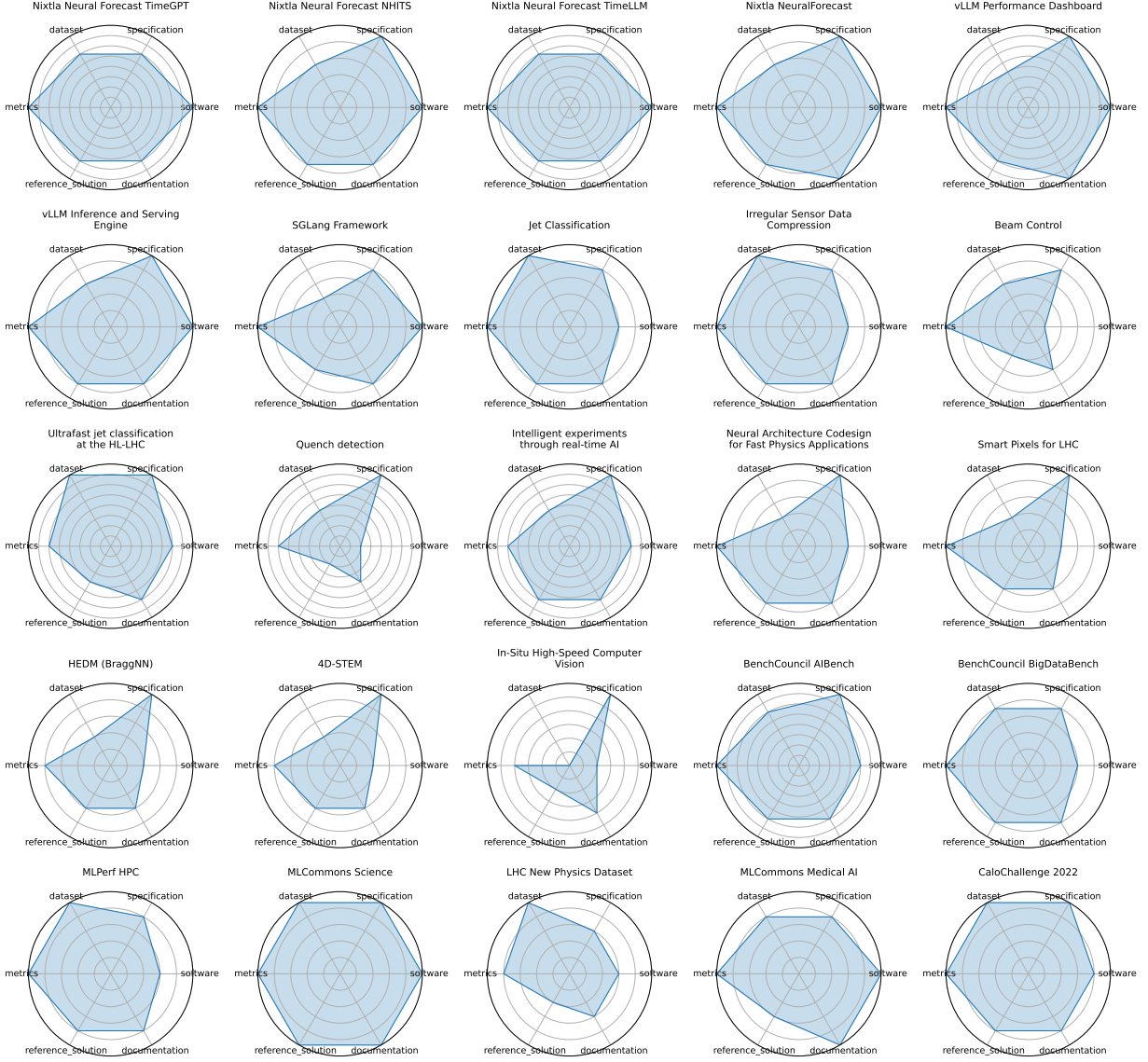


Figure 2: Radar chart overview (page 2)

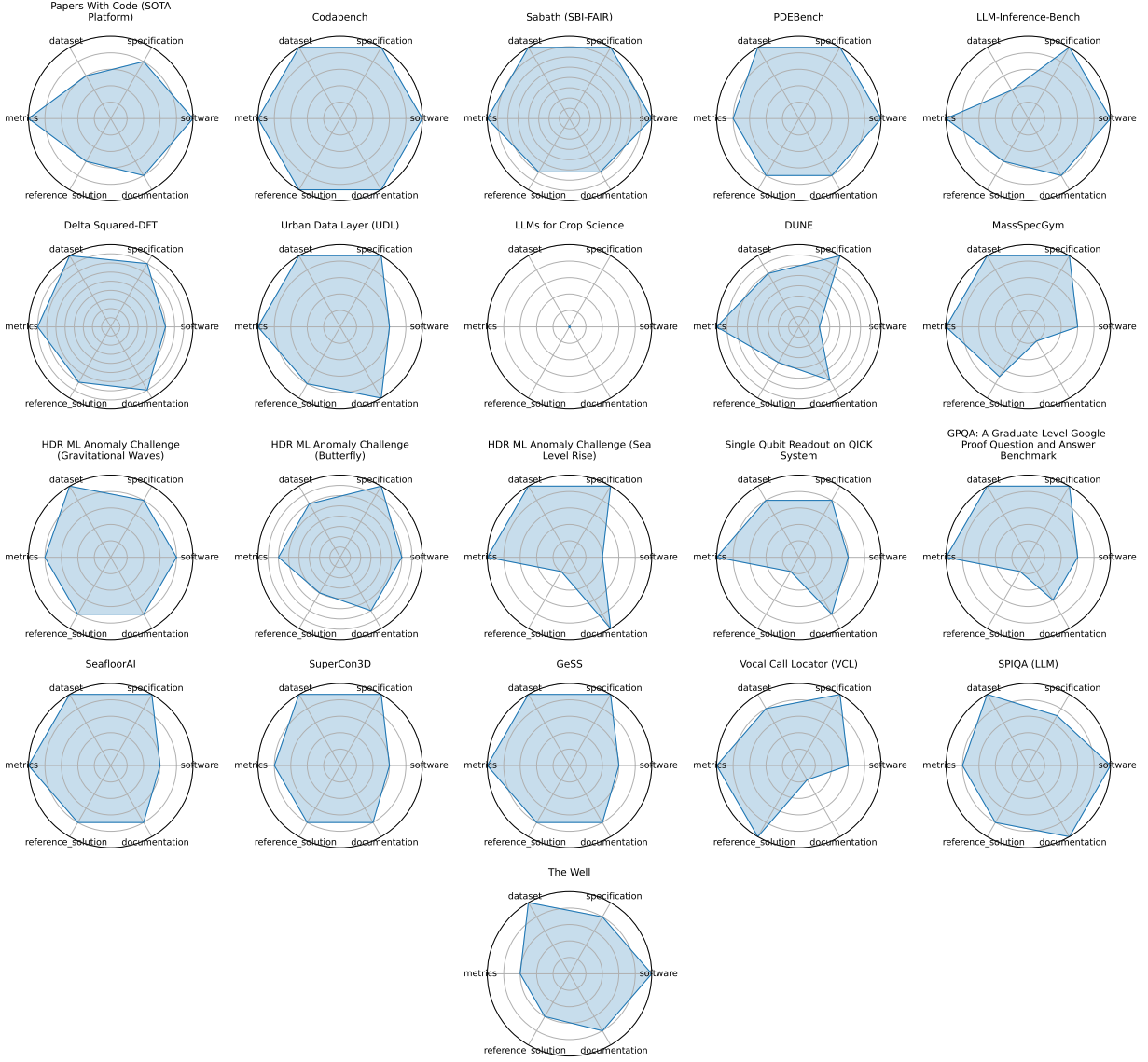


Figure 3: Radar chart overview (page 3)



## 3 Benchmark Details

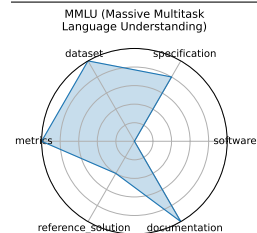
### 3.1 MMLU (Massive Multitask Language Understanding)

Measuring Massive Multitask Language Understanding (MMLU) is a benchmark of 57 multiple-choice tasks covering elementary mathematics, US history, computer science, law, and more, designed to evaluate a model's breadth and depth of knowledge in zero-shot and few-shot settings.

**date:** 2020-09-07  
**version:** 1  
**last\_updated:** 2020-09-07  
**expired:** false  
**valid:** yes  
**valid\_date:** 2025-07-28  
**url:** <https://paperswithcode.com/dataset/mmlu>  
**doi:** 10.48550/arXiv.2009.03300  
**domain:** Multidomain  
**focus:** Academic knowledge and reasoning across 57 subjects  
**keywords:** - multitask - multiple-choice - zero-shot - few-shot - knowledge probing  
**licensing:** MIT License  
**task\_types:** - Multiple choice  
**ai\_capability\_measured:** - General reasoning, subject-matter understanding  
**metrics:** - Accuracy  
**models:** - GPT-4o - Gemini 1.5 Pro - o1 - DeepSeek-R1  
**ml\_motif:** - General knowledge  
**type:** Benchmark  
**ml\_task:** - Supervised Learning  
**solutions:** 1  
**notes:** Good  
**contact.name:** Dan Hendrycks  
**contact.email:** dan (at) safe.ai  
**datasets.links.name:** Papers with Code datasets  
**datasets.links.url:** <https://github.com/paperswithcode/paperswithcode-data>  
**results.links.name:** Chinchilla  
**results.links.url:** <https://arxiv.org/abs/2203.15556>  
**fair.reproducible:** True  
**fair.benchmark\_ready:** True  
**id:** mmlu\_massive\_multitask\_language\_understanding  
**Citations:** [1]

#### Ratings:

Rating	Value	Reason
dataset	5	Meets all FAIR principles and properly versioned.
documentation	5	Well-explained in a provided paper.
metrics	5	Fully defined, represents a solution's performance.
reference_solution	2	Reference models are available (i.e. GPT-3), but are not trainable or publicly documented
software	0	No instructions to download or run data given on the site
specification	4	No system constraints



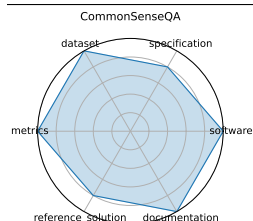
## 3.2 CommonsenseQA

CommonsenseQA is a challenging multiple-choice QA dataset built from ConceptNet, requiring models to apply commonsense knowledge to select the correct answer among five choices.

<b>date:</b>	2019-11-20
<b>version:</b>	1
<b>last_updated:</b>	2019-11-20
<b>expired:</b>	false
<b>valid:</b>	yes
<b>valid_date:</b>	2019-11-20
<b>url:</b>	<a href="https://paperswithcode.com/paper/commonsenseqa-a-question-answering-challenge">https://paperswithcode.com/paper/commonsenseqa-a-question-answering-challenge</a>
<b>doi:</b>	10.48550/arXiv.1811.00937
<b>domain:</b>	NLP; Commonsense
<b>focus:</b>	Commonsense question answering
<b>keywords:</b>	- ConceptNet - multiple-choice - adversarial
<b>licensing:</b>	MIT
<b>task_types:</b>	- Multiple choice
<b>ai_capability_measured:</b>	- Commonsense reasoning and knowledge integration
<b>metrics:</b>	- Accuracy
<b>models:</b>	- BERT-large - RoBERTa - GPT-3
<b>ml_motif:</b>	- Commonsense question answering
<b>type:</b>	Benchmark
<b>ml_task:</b>	- Supervised Learning
<b>solutions:</b>	2
<b>notes:</b>	Baseline 56%, human 89%
<b>contact.name:</b>	Alon Talmor, Jonathan Herzig, Nicholas Lourie, Jonathan Berant
<b>contact.email:</b>	Unknown
<b>datasets.links.name:</b>	CommonsenseQA Dataset (Hugging Face)
<b>datasets.links.url:</b>	<a href="https://huggingface.co/datasets/commonsense_qa">https://huggingface.co/datasets/commonsense_qa</a>
<b>results.links.name:</b>	Papers With Code Leaderboard for CommonsenseQA
<b>results.links.url:</b>	<a href="https://paperswithcode.com/dataset/commonsenseqa">https://paperswithcode.com/dataset/commonsenseqa</a>
<b>fair.reproducible:</b>	True
<b>fair.benchmark_ready:</b>	True
<b>id:</b>	commonsenseqa
<b>Citations:</b>	[2]

### Ratings:

Rating	Value	Reason
dataset	5	Public, versioned, and FAIR-compliant; includes metadata, splits, and licensing; well-integrated with HuggingFace and other ML libraries.
documentation	5	Given in paper.
metrics	5	Accuracy is a simple, reproducible metric aligned with task goals; no ambiguity in evaluation.
reference_solution	4	Several baseline models (e.g., BERT, RoBERTa) are reported with scores; implementations exist in public repos, but not run with hardware constraints
software	5	All code given on Github site
specification	4	Task and format (multiple-choice QA with 5 options) are clearly defined; grounded in ConceptNet with consistent structure, though no hardware/system constraints are specified.



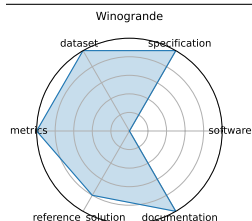
### 3.3 Winogrande

WinoGrande is a large-scale adversarial dataset of 44,000 Winograd Schema-style questions with reduced bias using AFLite, serving as both a benchmark and transfer learning resource.

<b>date:</b>	2019-07-24
<b>version:</b>	1
<b>last_updated:</b>	2019-07-24
<b>expired:</b>	false
<b>valid:</b>	yes
<b>valid_date:</b>	2019-07-24
<b>url:</b>	<a href="https://leaderboard.allenai.org/winogrande/submissions/public">https://leaderboard.allenai.org/winogrande/submissions/public</a>
<b>doi:</b>	10.48550/arXiv.1907.10641
<b>domain:</b>	NLP; Commonsense
<b>focus:</b>	Winograd Schema-style pronoun resolution
<b>keywords:</b>	- adversarial - pronoun resolution
<b>licensing:</b>	CC-BY
<b>task_types:</b>	- Pronoun resolution
<b>ai_capability_measured:</b>	- Robust commonsense reasoning
<b>metrics:</b>	- Accuracy - AUC
<b>models:</b>	- RoBERTa - BERT - GPT-2
<b>ml_motif:</b>	- Commonsense reasoning
<b>type:</b>	Benchmark
<b>ml_task:</b>	- Supervised Learning
<b>solutions:</b>	2
<b>notes:</b>	Human ~94%
<b>contact.name:</b>	Keisuke Sakaguchi
<b>contact.email:</b>	keisukes@allenai.org
<b>datasets.links.name:</b>	Hugging Face / AllenAI
<b>datasets.links.url:</b>	<a href="https://huggingface.co/datasets/allenai/winogrande">https://huggingface.co/datasets/allenai/winogrande</a>
<b>results.links.name:</b>	Papers With Code leaderboard
<b>results.links.url:</b>	<a href="https://paperswithcode.com/dataset/winogrande">https://paperswithcode.com/dataset/winogrande</a>
<b>fair.reproducible:</b>	True
<b>fair.benchmark_ready:</b>	True
<b>id:</b>	winogrande
<b>Citations:</b>	[3]

#### Ratings:

Rating	Value	Reason
dataset	5	Public, versioned, and FAIR-compliant with AFLite-generated splits to reduce annotation artifacts; hosted by AllenAI with good metadata.
documentation	5	Dataset page and paper provide sufficient detail
metrics	5	Accuracy and AUC are quantitative and well-aligned with disambiguation goals; standardized across evaluations.
reference_solution	4	Baseline results available, requiring users to submit their methods along with their submissions. Constraints are not required in submissions.
software	0	No template code provided
specification	5	Task (pronoun/coreference resolution) is clearly defined in Winograd Schema style, with consistent input/output format; no system constraints included.



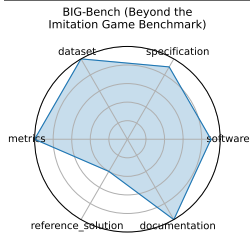
### 3.4 BIG-Bench (Beyond the Imitation Game Benchmark)

BIG-Bench is a collaborative suite of 204 tasks designed to probe LLMs’ reasoning, knowledge, and bias across diverse domains and difficulty levels beyond simple imitation.

**date:** 2022-06-09  
**version:** 1  
**last\_updated:** 2022-06-09  
**expired:** false  
**valid:** yes  
**valid\_date:** 2022-06-09  
**url:** <https://github.com/google/BIG-bench>  
**doi:** 10.48550/arXiv.2206.04615  
**domain:** NLP; AI Evaluation  
**focus:** Diverse reasoning and generalization tasks  
**keywords:** - few-shot - multi-task - bias analysis  
**licensing:** Apache-2.0  
**task\_types:** - Few-shot evaluation - Multi-task evaluation  
**ai\_capability\_measured:** - Reasoning and generalization across diverse tasks  
**metrics:** - Accuracy - Task-specific metrics  
**models:** - GPT-3 - Dense Transformers - Sparse Transformers  
**ml\_motif:** - LLM evaluation  
**type:** Benchmark  
**ml\_task:** - Supervised Learning  
**solutions:** Multiple, including human baselines  
**notes:** Human baselines included  
**contact.name:** Aarohi Srivastava et al.  
**contact.email:** [bigbench@googlegroups.com](mailto:bigbench@googlegroups.com)  
**datasets.links.name:** BIG-Bench GitHub Repository (contains tasks and data)  
**datasets.links.url:** [https://github.com/google/BIG-bench/tree/main/bigbench/benchmark\\_tasks](https://github.com/google/BIG-bench/tree/main/bigbench/benchmark_tasks)  
**results.links.name:** BIG-Bench GitHub Repository (results in papers and code)  
**results.links.url:** <https://github.com/google/BIG-bench>  
**fair.reproducible:** True  
**fair.benchmark\_ready:** True  
**id:** big-bench\_beyond\_the\_imitation\_game\_benchmark  
**Citations:** [4]

#### Ratings:

Rating	Value	Reason
dataset	5	Public, versioned, and well-documented; FAIR overall
documentation	5	Explained in the associated paper.
metrics	5	Many tasks use standard quantitative metrics (accuracy, BLEU, F1). Others involve subjective ratings (e.g., Likert), which reduces cross-task comparability.
reference_solution	2	Human baselines and LLM performance results are included; however, runnable reference solutions are limited and setup is not fully turnkey.
software	4.5	Quick start notebook provided, but instructions on how to run it are lacking.
specification	4.5	Tasks are diverse and clearly described; input/output formats are usually defined but vary widely, and system constraints are not standardized.



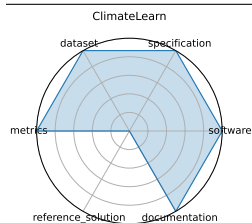
### 3.5 ClimateLearn

ClimateLearn provides standardized datasets and evaluation protocols for machine learning models in medium-range weather and climate forecasting using ERA5 reanalysis.

<b>date:</b>	2023-07-19
<b>version:</b>	1
<b>last_updated:</b>	2023-07-19
<b>expired:</b>	false
<b>valid:</b>	yes
<b>valid_date:</b>	2023-07-19
<b>url:</b>	<a href="https://arxiv.org/abs/2307.01909">https://arxiv.org/abs/2307.01909</a>
<b>doi:</b>	10.48550/arXiv.2307.01909
<b>domain:</b>	Climate Science; Forecasting
<b>focus:</b>	ML for weather and climate modeling
<b>keywords:</b>	- medium-range forecasting - ERA5 - data-driven
<b>licensing:</b>	CC-BY-4.0
<b>task_types:</b>	- Forecasting
<b>ai_capability_measured:</b>	- Global weather prediction (3-5 days)
<b>metrics:</b>	- RMSE - Anomaly correlation
<b>models:</b>	- CNN baselines - ResNet variants
<b>ml_motif:</b>	- Forecasting - Benchmarking
<b>type:</b>	Benchmark
<b>ml_task:</b>	- Supervised Learning
<b>solutions:</b>	Multiple baseline models provided
<b>notes:</b>	Includes physical and ML baselines.
<b>contact.name:</b>	Jason Jewik
<b>contact.email:</b>	<a href="mailto:jason.jewik@ucla.edu">jason.jewik@ucla.edu</a>
<b>datasets.links.name:</b>	ClimateLearn GitHub Repository (data loaders and processing)
<b>datasets.links.url:</b>	<a href="https://github.com/aditya-grover/climate-learn">https://github.com/aditya-grover/climate-learn</a>
<b>results.links.name:</b>	ClimateLearn Paper (results section)
<b>results.links.url:</b>	<a href="https://arxiv.org/abs/2307.01909">https://arxiv.org/abs/2307.01909</a>
<b>fair.reproducible:</b>	True
<b>fair.benchmark_ready:</b>	True
<b>id:</b>	climatelearn
<b>Citations:</b>	[5]

#### Ratings:

Rating	Value	Reason
dataset	5	Provides standardized access to ERA5 and other reanalysis datasets, with ML-ready splits, metadata, and Xarray-compatible formats; versioned and fully FAIR-compliant.
documentation	5	Explained in the benchmark's paper.
metrics	5	ACC and RMSE are standard, quantitative, and appropriate for climate forecasting; well-integrated into the benchmark, though interpretation across domains may vary.
reference_solution	0	The benchmark is geared for CNN architectures, but no specific model was mentioned.
software	5	Quickstart notebook makes for easy usage
specification	5	Task framing (medium-range climate forecasting), input/output formats, and evaluation windows are clearly defined; benchmark supports both physical and learned models with detailed constraints.



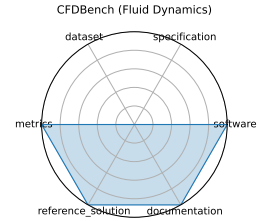
### 3.6 CFDBench (Fluid Dynamics)

CFDBench provides large-scale CFD data for four canonical fluid flow problems, assessing neural operators’ ability to generalize to unseen PDE parameters and domains.

**date:** 2024-10-01  
**version:** 1  
**last\_updated:** 2024-10-01  
**expired:** false  
**valid:** yes  
**valid\_date:** 2024-10-01  
**url:** <https://arxiv.org/abs/2310.05963>  
**doi:** 10.48550/arXiv.2310.05963  
**domain:** Fluid Dynamics; Scientific ML  
**focus:** Neural operator surrogate modeling  
**keywords:** - neural operators - CFD - FNO - DeepONet  
**licensing:** CC-BY-4.0  
**task\_types:** - Surrogate modeling  
**ai\_capability\_measured:** - Generalization of neural operators for PDEs  
**metrics:** - L2 error - MAE  
**models:** - FNO - DeepONet - U-Net  
**ml\_motif:** - Generalization  
**type:** Benchmark  
**ml\_task:** - Supervised Learning  
**solutions:** Numerous, as it’s a benchmark for ML models  
**notes:** 302K frames across 739 cases  
**contact.name:** Yining Luo  
**contact.email:** [yining.luo@mail.utoronto.ca](mailto:yining.luo@mail.utoronto.ca)  
**datasets.links.name:** unknown  
**datasets.links.url:** unknown  
**results.links.name:** unknown  
**results.links.url:** unknown  
**fair.reproducible:** True  
**fair.benchmark\_ready:** True  
**id:** cfdbench\_fluid\_dynamics  
**Citations:** [6]

#### Ratings:

Rating	Value	Reason
dataset	0	Not given
documentation	5	Associated paper gives all necessary information.
metrics	5	Quantitative metrics (L2 error, MAE, relative error) are clearly defined and align with regression task objectives.
reference_solution	5	Baseline models like FNO and DeepONet are implemented, hardware specified.
software	5	The benchmark provides Python scripts for data loading, preprocessing, and model training/evaluation
specification	0	Not listed



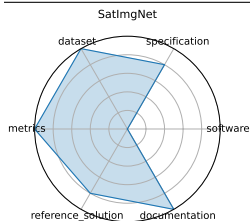
### 3.7 SatImgNet

SATIN (sometimes referred to as SatImgNet) is a multi-task metadataset of 27 satellite imagery classification datasets evaluating zero-shot transfer of vision-language models across diverse remote sensing tasks.

**date:** 2023-04-23  
**version:** 1  
**last\_updated:** 2023-04-23  
**expired:** false  
**valid:** yes  
**valid\_date:** 2023-04-23  
**url:** <https://huggingface.co/datasets/saral-ai/satimagnet>  
**doi:** 10.48550/arXiv.2304.11619  
**domain:** Remote Sensing  
**focus:** Satellite imagery classification  
**keywords:** - land-use - zero-shot - multi-task  
**licensing:** CC-BY-4.0  
**task\_types:** - Image classification  
**ai\_capability\_measured:** - Zero-shot land-use classification  
**metrics:** - Accuracy  
**models:** - CLIP - BLIP - ALBEF  
**ml\_motif:** - Transfer Learning  
**type:** Benchmark  
**ml\_task:** - Supervised Learning  
**solutions:** Numerous, evaluated via leaderboard  
**notes:** Public leaderboard available  
**contact.name:** Jonathan Roberts  
**contact.email:** [j.roberts@cs.ox.ac.uk](mailto:j.roberts@cs.ox.ac.uk)  
**datasets.links.name:** SatImgNet on Hugging Face  
**datasets.links.url:** <https://huggingface.co/datasets/saral-ai/satimagnet>  
**results.links.name:** SatImgNet Leaderboard  
**results.links.url:** <https://huggingface.co/spaces/saral-ai/satin-leaderboard>  
**fair.reproducible:** True  
**fair.benchmark\_ready:** True  
**id:** satimgnet  
**Citations:** [7]

#### Ratings:

Rating	Value	Reason
dataset	5	Hosted on Hugging Face, versioned, FAIR-compliant with rich metadata; covers many well-known remote sensing datasets unified under one metadataset, though documentation depth varies slightly across tasks.
documentation	5	Paper provides all required information
metrics	5	Accuracy of classification is an appropriate metric
reference_solution	4	Baselines like CLIP, BLIP, ALBEF evaluated in the paper; no constraints specified
software	0	No scripts or environment information provided
specification	4	Tasks (image classification across 27 satellite datasets) are clearly defined with multi-task and zero-shot framing; input/output structure is mostly standard but some task-specific nuances require interpretation.



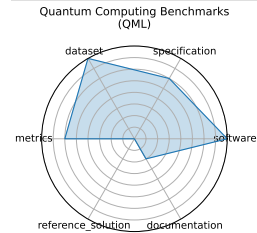
### 3.8 Quantum Computing Benchmarks (QML)

A suite of benchmarks evaluating quantum hardware and algorithms on tasks such as state preparation, circuit optimization, and error correction across multiple platforms.

**date:** 2022-02-22  
**version:** 1  
**last\_updated:** 2022-02-22  
**expired:** false  
**valid:** yes  
**valid\_date:** 2022-02-22  
**url:** <https://github.com/XanaduAI/qml-benchmarks>  
**doi:** 10.48550/arXiv.2307.03901  
**domain:** Quantum Computing  
**focus:** Quantum algorithm performance evaluation  
**keywords:** - quantum circuits - state preparation - error correction  
**licensing:** Apache-2.0  
**task\_types:** - Circuit benchmarking - State classification  
**ai\_capability\_measured:** - Quantum algorithm performance and fidelity  
**metrics:** - Fidelity - Success probability  
**models:** - IBM Q - IonQ - AQT@LBNL  
**ml\_motif:** - Performance Evaluation  
**type:** Benchmark  
**ml\_task:** - Supervised Learning  
**solutions:** Varies per benchmark  
**notes:** Hardware-agnostic, application-level metrics. The citation may not be correct.  
**contact.name:** Xanadu AI  
**contact.email:** [support@xanadu.ai](mailto:support@xanadu.ai)  
**datasets.links.name:** PennyLane QML Benchmarks Datasets  
**datasets.links.url:** <https://pennylane.ai/datasets/collection/qml-benchmarks>  
**results.links.name:** QML Benchmarks GitHub Repository (Results section)  
**results.links.url:** <https://github.com/XanaduAI/qml-benchmarks#results-and-leaderboards>  
**fair.reproducible:** True  
**fair.benchmark\_ready:** True  
**id:** quantum\_computing\_benchmarks\_qml  
**Citations:** [8]

#### Ratings:

Rating	Value	Reason
dataset	4	Datasets are accessible, but not split.
documentation	1	Only the task is defined.
metrics	3	Partially defined, somewhat inferable metrics. Unknown whether a system's performance is captured.
reference_solution	0	Not provided
software	4	Run instructions exist, but are not easy to follow
specification	3	No system constraints. Task clarity and dataset format are not clearly specified.





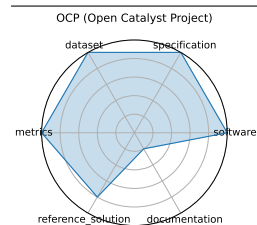
### 3.9 OCP (Open Catalyst Project)

The Open Catalyst Project (OC20 and OC22) provides DFT-calculated catalyst-adsorbate relaxation datasets, challenging ML models to predict energies and forces for renewable energy applications.

**date:** 2020-10-20  
**version:** 1  
**last\_updated:** 2020-10-20  
**expired:** false  
**valid:** yes  
**valid\_date:** 2020-10-20  
**url:** <https://opencatalystproject.org/>  
**doi:** unknown  
**domain:** Chemistry; Materials Science  
**focus:** Catalyst adsorption energy prediction  
**keywords:** - DFT relaxations - adsorption energy - graph neural networks  
**licensing:** OCP Terms of Use  
**task\_types:** - Energy prediction - Force prediction  
**ai\_capability\_measured:** - Prediction of adsorption energies and forces  
**metrics:** - MAE (energy) - MAE (force)  
**models:** - CGCNN - SchNet - DimeNet++ - GemNet-OC  
**ml\_motif:** - Chemistry  
**type:** Benchmark  
**ml\_task:** - Supervised Learning  
**solutions:** 0  
**notes:** Public leaderboards; active community development  
**contact.name:** unknown  
**contact.email:** unknown  
**datasets.links.name:** OCP Dataset  
**datasets.links.url:** <https://fair-chem.github.io/catalysts/datasets/summary>  
**results.links.name:** OCP Pretrained Models  
**results.links.url:** <https://fair-chem.github.io/catalysts/models.html>  
**fair.reproducible:** True  
**fair.benchmark\_ready:** True  
**id:** ocp\_open\_catalyst\_project  
**Citations:** [9], [10], [11], [12]

#### Ratings:

Rating	Value	Reason
dataset	5	Fully FAIR- OC20, per-adsorbate trajectories, and OC22 are versioned; datasets come with standardized splits, metadata, and are downloadable.
documentation	1	Paper exists, but content is behind a paywall.
metrics	5	MAE (energy and force) are standard and reproducible.
reference_solution	4	Multiple baselines (GemNet-OC, DimeNet++, etc.) implemented and evaluated. No hardware listed.
software	5	Data provided in Github links
specification	5	Tasks (energy and force prediction) are clearly defined with explicit I/O specifications, constraints, and physical relevance for renewable energy.



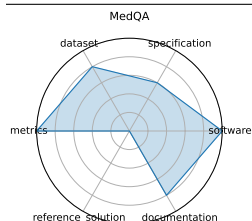
### 3.10 MedQA

MedQA is a large-scale multiple-choice dataset drawn from professional medical board exams (e.g., USMLE), testing AI systems on diagnostic and medical knowledge questions in English and Chinese.

<b>date:</b>	2020-09-28
<b>version:</b>	1
<b>last_updated:</b>	2020-09-28
<b>expired:</b>	false
<b>valid:</b>	yes
<b>valid_date:</b>	2020-09-28
<b>url:</b>	<a href="https://arxiv.org/abs/2009.13081">https://arxiv.org/abs/2009.13081</a>
<b>doi:</b>	10.48550/arXiv.2009.13081
<b>domain:</b>	Medical Question Answering
<b>focus:</b>	Medical board exam QA
<b>keywords:</b>	- USMLE - diagnostic QA - medical knowledge - multilingual
<b>licensing:</b>	Under Association for the Advancement of Artificial Intelligence
<b>task_types:</b>	- Multiple choice
<b>ai_capability_measured:</b>	- Medical diagnosis and knowledge retrieval
<b>metrics:</b>	- Accuracy
<b>models:</b>	- Neural reader - Retrieval-based QA systems
<b>ml_motif:</b>	- Medical diagnosis
<b>type:</b>	Benchmark
<b>ml_task:</b>	- Supervised Learning
<b>solutions:</b>	0
<b>notes:</b>	Multilingual (English, Simplified and Traditional Chinese)
<b>contact.name:</b>	Di Jin
<b>contact.email:</b>	<a href="mailto:jindi15@mit.edu">jindi15@mit.edu</a>
<b>datasets.links.name:</b>	Github
<b>datasets.links.url:</b>	<a href="https://github.com/jind11/MedQA">https://github.com/jind11/MedQA</a>
<b>results.links.name:</b>	unknown
<b>results.links.url:</b>	unknown
<b>fair.reproducible:</b>	True
<b>fair.benchmark_ready:</b>	True
<b>id:</b>	medqa
<b>Citations:</b>	[13]

#### Ratings:

Rating	Value	Reason
dataset	4	Dataset is publicly available (GitHub, paper, Hugging Face), well-structured. However, versioning and metadata could be more standardized to fully meet FAIR criteria.
documentation	4	Paper is available. Evaluation criteria are not mentioned.
metrics	5	Uses clear, quantitative metric (accuracy), standard for multiple-choice benchmarks; easily comparable across models.
reference_solution	0	No reference solution mentioned.
software	5	All code available on the github
specification	3	Task is clearly defined as multiple-choice QA for medical board exams; input and output formats are explicit; task scope is rigorous and structured. System constraints not specified.



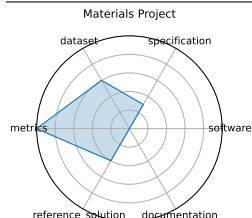
### 3.11 Materials Project

The Materials Project provides an open-access database of computed properties for inorganic materials via high-throughput density functional theory (DFT), accelerating materials discovery.

**date:** 2011-10-01  
**version:** 1  
**last\_updated:** 2011-10-01  
**expired:** false  
**valid:** yes  
**valid\_date:** 2011-10-01  
**url:** <https://materialsproject.org/>  
**doi:** unknown  
**domain:** Materials Science  
**focus:** DFT-based property prediction  
**keywords:** - DFT - materials genome - high-throughput  
**licensing:** <https://next-gen.materialsproject.org/about/terms>  
**task\_types:** - Property prediction  
**ai\_capability\_measured:** - Prediction of inorganic material properties  
**metrics:** - MAE -  $R^2$   
**models:** - Automatminer - Crystal Graph Neural Networks  
**ml\_motif:** - Material properties  
**type:** Benchmark  
**ml\_task:** - Supervised Learning  
**solutions:** 0  
**notes:** Core component of the Materials Genome Initiative  
**contact.name:** unknown  
**contact.email:** unknown  
**datasets.links.name:** Materials Project Catalysis Explorer  
**datasets.links.url:** <https://next-gen.materialsproject.org/catalysis>  
**results.links.name:** unknown  
**results.links.url:** unknown  
**fair.reproducible:** True  
**fair.benchmark\_ready:** True  
**id:** materials\_project  
**Citations:** [14]

#### Ratings:

Rating	Value	Reason
dataset	3	API key required to access data. No predefined splits.
documentation	0	No explanations or paper provided
metrics	5	Uses numerical metrics like MAE and $R^2$
reference_solution	2	Numerous models (e.g., Automatminer, CGCNN) trained on the database, but no constraints or documentation listed.
software	0	No instructions available
specification	1.5	The platform offers a wide range of material property prediction tasks, but task framing and I/O formats vary by API use and are not always standardized across use cases.



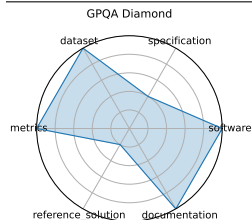
### 3.12 GPQA Diamond

GPQA is a dataset of 448 challenging, multiple-choice questions in biology, physics, and chemistry, written by domain experts. It is Google-proof - experts score 65% (74% after error correction) while skilled non-experts with web access score only 34%. State-of-the-art LLMs like GPT-4 reach around 39% accuracy.

**date:** 2023-11-20  
**version:** 1  
**last\_updated:** 2023-11-20  
**expired:** false  
**valid:** yes  
**valid\_date:** 2023-11-20  
**url:** <https://arxiv.org/abs/2311.12022>  
**doi:** 10.48550/arXiv.2311.12022  
**domain:** Science  
**focus:** Graduate-level scientific reasoning  
**keywords:** - Google-proof - graduate-level - science QA - chemistry - physics  
**licensing:** unknown  
**task\_types:** - Multiple choice - Multi-step QA  
**ai\_capability\_measured:** - Scientific reasoning, deep knowledge  
**metrics:** - Accuracy  
**models:** - o1 - DeepSeek-R1  
**ml\_motif:** - Science and STEM fields  
**type:** Benchmark  
**ml\_task:** - Supervised Learning  
**solutions:** 0  
**notes:** Good  
**contact.name:** Julian Michael  
**contact.email:** julianjm@nyu.edu  
**datasets.links.name:** unknown  
**datasets.links.url:** unknown  
**results.links.name:** unknown  
**results.links.url:** unknown  
**fair.reproducible:** True  
**fair.benchmark\_ready:** True  
**id:** gpqa\_diamond  
**Citations:** [15]

#### Ratings:

Rating	Value	Reason
dataset	5	Easily able to access dataset. Comes with predefined splits as mentioned in the paper
documentation	5	All information is listed in the associated paper
metrics	5	Each question has a correct answer, representing the tested model's performance.
reference_solution	1	Common models such as GPT-3.5 were compared. They are not open and don't provide requirements
software	5	Python version and requirements specified on Github site
specification	2	No system constraints or I/O specified



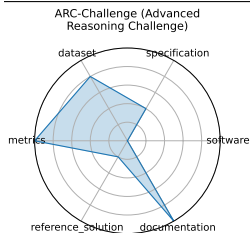
### 3.13 ARC-Challenge (Advanced Reasoning Challenge)

The AI2 Reasoning Challenge (ARC) Challenge set comprises 7,787 natural, grade-school science questions that retrieval-based and word co-occurrence algorithms both fail, requiring advanced reasoning over a 14-million-sentence corpus.

**date:** 2018-03-14  
**version:** 1  
**last\_updated:** 2018-03-14  
**expired:** false  
**valid:** yes  
**valid\_date:** 2018-03-14  
**url:** <https://allenai.org/data/arc>  
**doi:** NA  
**domain:** Science  
**focus:** Grade-school science with reasoning emphasis  
**keywords:** - grade-school - science QA - challenge set - reasoning  
**licensing:** Apache 2.0 License  
**task\_types:** - Multiple choice  
**ai\_capability\_measured:** - Commonsense and scientific reasoning  
**metrics:** - Accuracy  
**models:** - GPT-4 - Claude  
**ml\_motif:** - Elementary science  
**type:** Benchmark  
**ml\_task:** - Supervised Learning  
**solutions:** 0  
**notes:** Good  
**contact.name:** unknown  
**contact.email:** unknown  
**datasets.links.name:** Hugging Face  
**datasets.links.url:** [https://huggingface.co/datasets/allenai/ai2\\_arc](https://huggingface.co/datasets/allenai/ai2_arc)  
**results.links.name:** unknown  
**results.links.url:** unknown  
**fair.reproducible:** True  
**fair.benchmark\_ready:** True  
**id:** arc-challenge\_advanced\_reasoning\_challenge  
**Citations:** [16]

#### Ratings:

Rating	Value	Reason
dataset	4	Data accessible, offers instructions on how to download the data via CLI tools. No splits.
documentation	5	Explains all necessary information inside a paper
metrics	5	(by default) All questions in the dataset are multiple choice, all have a correct answer
reference_solution	1	There are over 300 models listed, but very few, if any, show performance on the dataset or list constraints
software	0	No link to code or documentation
specification	2	Task is clear, but no constraints or format is mentioned



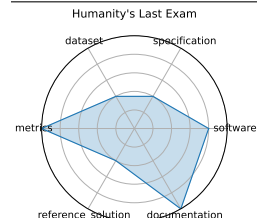
### 3.14 Humanity’s Last Exam

Humanity’s Last Exam is a multi-domain, multiple-choice benchmark containing 2,000 questions across diverse academic disciplines, designed to evaluate LLMs’ ability to reason across domains without external resources.

**date:** 2025-01-24  
**version:** 1  
**last\_updated:** 2025-01-24  
**expired:** false  
**valid:** yes  
**valid\_date:** 2025-01-24  
**url:** <https://arxiv.org/abs/2501.14249>  
**doi:** 10.48550/arXiv.2501.14249  
**domain:** Multidomain  
**focus:** Broad cross-domain academic reasoning  
**keywords:** - cross-domain - academic exam - multiple-choice - multidisciplinary  
**licensing:** MIT License  
**task\_types:** - Multiple choice  
**ai\_capability\_measured:** - Cross-domain academic reasoning  
**metrics:** - Accuracy  
**models:** - unknown  
**ml\_motif:** - Multi-domain  
**type:** Benchmark  
**ml\_task:** - Supervised Learning  
**solutions:** 0  
**notes:** Good  
**contact.name:** HLE team  
**contact.email:** [agibenchmark@safe.ai](mailto:agibenchmark@safe.ai)  
**datasets.links.name:** Hugging Face  
**datasets.links.url:** <https://huggingface.co/datasets/cais/hle>  
**results.links.name:** unknown  
**results.links.url:** unknown  
**fair.reproducible:** True  
**fair.benchmark\_ready:** True  
**id:** humanitys\_last\_exam  
**Citations:** [17]

#### Ratings:

Rating	Value	Reason
dataset	2	Data accessible through Hugging Face, but requires giving contact information to access
documentation	5	Paper available with necessary information
metrics	5	(by default) All questions in the dataset are multiple choice, all have a correct answer
reference_solution	2	Performance for cutting-edge models listed, but does not specify exact version of the models or how to reproduce the result
software	4	Code for testing models posted on the github. Unknown how to run a custom model.
specification	2	Format of inputs (natural language) and outputs (multiple choice or natural language) specified. No HW constraints specified



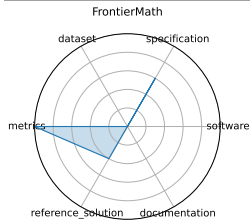
### 3.15 FrontierMath

FrontierMath is a benchmark of hundreds of expert-vetted mathematics problems spanning number theory, real analysis, algebraic geometry, and category theory, measuring LLMs ability to solve problems requiring deep abstract reasoning.

<b>date:</b>	2024-11-07
<b>version:</b>	1
<b>last_updated:</b>	2024-11-07
<b>expired:</b>	false
<b>valid:</b>	yes
<b>valid_date:</b>	2024-11-07
<b>url:</b>	<a href="https://arxiv.org/abs/2411.04872">https://arxiv.org/abs/2411.04872</a>
<b>doi:</b>	10.48550/arXiv.2411.04872
<b>domain:</b>	Mathematics
<b>focus:</b>	Challenging advanced mathematical reasoning
<b>keywords:</b>	- symbolic reasoning - number theory - algebraic geometry - category theory
<b>licensing:</b>	unknown
<b>task_types:</b>	- Problem solving
<b>ai_capability_measured:</b>	- Symbolic and abstract mathematical reasoning
<b>metrics:</b>	- Accuracy
<b>models:</b>	- unknown
<b>ml_motif:</b>	- Math problem solving
<b>type:</b>	Benchmark
<b>ml_task:</b>	- Supervised Learning
<b>solutions:</b>	0
<b>notes:</b>	Good
<b>contact.name:</b>	FrontierMath team
<b>contact.email:</b>	math_evals@epochai.org
<b>datasets.links.name:</b>	unknown
<b>datasets.links.url:</b>	unknown
<b>results.links.name:</b>	unknown
<b>results.links.url:</b>	unknown
<b>fair.reproducible:</b>	True
<b>fair.benchmark_ready:</b>	True
<b>id:</b>	frontiermath
<b>Citations:</b>	[18]

#### Ratings:

Rating	Value	Reason
dataset	0	Paper and website had no link to any dataset. It may still exist somewhere
documentation	0	No specified way to reproduce the reference solution
metrics	5	(by default) All questions in the dataset have a correct answer
reference_solution	2	Displays result of leading models on the benchmark, but none are trainable or list constraints
software	0	No link to code provided
specification	3	Well-specified process for asking questions and receiving answers. No software or hardware constraints



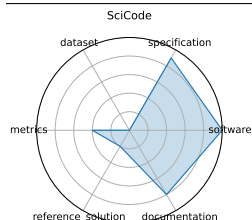
### 3.16 SciCode

SciCode is a scientist-curated coding benchmark with 338 subproblems derived from 80 real research tasks across 16 scientific subfields, evaluating models on knowledge recall, reasoning, and code synthesis for scientific computing tasks.

<b>date:</b>	2024-07-18
<b>version:</b>	1
<b>last_updated:</b>	2024-07-18
<b>expired:</b>	false
<b>valid:</b>	yes
<b>valid_date:</b>	2024-07-18
<b>url:</b>	<a href="https://arxiv.org/abs/2407.13168">https://arxiv.org/abs/2407.13168</a>
<b>doi:</b>	10.48550/arXiv.2407.13168
<b>domain:</b>	Scientific Programming
<b>focus:</b>	Scientific code generation and problem solving
<b>keywords:</b>	- code synthesis - scientific computing - programming benchmark
<b>licensing:</b>	unknown
<b>task_types:</b>	- Coding
<b>ai_capability_measured:</b>	- Program synthesis, scientific computing
<b>metrics:</b>	- Solve rate (%)
<b>models:</b>	- Claude3.5-Sonnet
<b>ml_motif:</b>	- Coding
<b>type:</b>	Benchmark
<b>ml_task:</b>	- Supervised Learning
<b>solutions:</b>	unknown
<b>notes:</b>	Good
<b>contact.name:</b>	Minyang Tian
<b>contact.email:</b>	mtian8@illinois.edu
<b>datasets.links.name:</b>	unknown
<b>datasets.links.url:</b>	unknown
<b>results.links.name:</b>	unknown
<b>results.links.url:</b>	unknown
<b>fair.reproducible:</b>	True
<b>fair.benchmark_ready:</b>	True
<b>id:</b>	scicode
<b>Citations:</b>	[19]

#### Ratings:

Rating	Value	Reason
dataset	0	Paper and website had no link to any dataset. It may still exist somewhere
documentation	4	Paper containing all needed info except for evaluation criteria
metrics	2	Metrics stated, but method of grading is not specified
reference_solution	1	Models presented with scores, but none are open or list constraints
software	5	Code to run exists on github repo
specification	4.5	Expected outputs and broad types of inputs stated. Few details on output grading. No HW constraints.





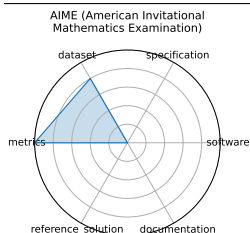
### 3.17 AIME (American Invitational Mathematics Examination)

The AIME is a 15-question, 3-hour exam for high-school students featuring challenging short-answer math problems in algebra, number theory, geometry, and combinatorics, assessing depth of problem-solving ability.

<b>date:</b>	2025-03-13
<b>version:</b>	1
<b>last_updated:</b>	2025-03-13
<b>expired:</b>	false
<b>valid:</b>	yes
<b>valid_date:</b>	2025-03-13
<b>url:</b>	<a href="https://artofproblemsolving.com/wiki/index.php/AIME_Problems_and_Solutions">https://artofproblemsolving.com/wiki/index.php/AIME_Problems_and_Solutions</a>
<b>doi:</b>	NA
<b>domain:</b>	Mathematics
<b>focus:</b>	Pre-college advanced problem solving
<b>keywords:</b>	- algebra - combinatorics - number theory - geometry
<b>licensing:</b>	unknown
<b>task_types:</b>	- Problem solving
<b>ai_capability_measured:</b>	- Mathematical problem-solving and reasoning
<b>metrics:</b>	- Accuracy
<b>models:</b>	- unknown
<b>ml_motif:</b>	- Math problem solving
<b>type:</b>	Benchmark
<b>ml_task:</b>	- Supervised Learning
<b>solutions:</b>	0
<b>notes:</b>	Designed for human test-takers
<b>contact.name:</b>	unknown
<b>contact.email:</b>	unknown
<b>datasets.links.name:</b>	AoPS website
<b>datasets.links.url:</b>	<a href="https://artofproblemsolving.com/wiki/index.php/AIME_Problems_and_Solutions">https://artofproblemsolving.com/wiki/index.php/AIME_Problems_and_Solutions</a>
<b>results.links.name:</b>	unknown
<b>results.links.url:</b>	unknown
<b>fair.reproducible:</b>	True
<b>fair.benchmark_ready:</b>	True
<b>id:</b>	aime_american_invitational_mathematics_examination
<b>Citations:</b>	[20]

#### Ratings:

Rating	Value	Reason
dataset	4	Easily accessible data with problems and solutions, but no splits
documentation	0	Not given
metrics	5	(by default) Answer is correct or it's not
reference_solution	0	Not given. Human performance stats exist, but no mentions of AI performance
software	0	No code available
specification	0	Obvious what the problems are, but not specified how to administer them to AI models. No HW constraints



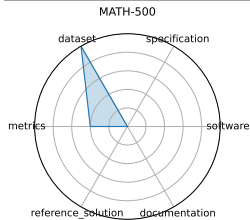
### 3.18 MATH-500

MATH-500 is a curated subset of 500 problems from the OpenAI MATH dataset, spanning high-school to advanced levels, designed to evaluate LLMs mathematical reasoning and generalization.

**date:** 2025-02-15  
**version:** 1  
**last\_updated:** 2025-02-15  
**expired:** false  
**valid:** yes  
**valid\_date:** 2025-02-15  
**url:** <https://huggingface.co/datasets/HuggingFaceH4/MATH-500>  
**doi:** unknown  
**domain:** Mathematics  
**focus:** Math reasoning generalization  
**keywords:** - calculus - algebra - number theory - geometry  
**licensing:** MIT License  
**task\_types:** - Problem solving  
**ai\_capability\_measured:** - Math reasoning and generalization  
**metrics:** - Accuracy  
**models:** - unknown  
**ml\_motif:** - Math problem solving  
**type:** Benchmark  
**ml\_task:** - Supervised Learning  
**solutions:** 0  
**notes:** Dataset hosted on Hugging Face. Data comes from a subset of OpenAI's dataset  
**contact.name:** unknown  
**contact.email:** unknown  
**datasets.links.name:** Hugging Face  
**datasets.links.url:** <https://huggingface.co/datasets/HuggingFaceH4/MATH-500>  
**results.links.name:** unknown  
**results.links.url:** unknown  
**fair.reproducible:** True  
**fair.benchmark\_ready:** True  
**id:** math-  
**Citations:** [21]

#### Ratings:

Rating	Value	Reason
dataset	5	Problems and solutions are easily downloaded. Could not find a way to download the data
documentation	0	Not given. Implicit instructions to download dataset.
metrics	2	Problem spec states that all of the AI reasoning steps are subject to grading, but no specified way to evaluate the steps
reference_solution	0	Not given
software	0	No code provided
specification	0	No method of presentation and evaluation is not stated. No constraints



CURIE is a benchmark of 580 problems across six scientific disciplines-materials science, quantum computing, biology, chemistry, climate science, and astrophysics- designed to evaluate LLMs on long-context understanding, reasoning, and information extraction in realistic scientific workflows.

**Ratings:**

CURIE (Scientific Long-Context Understanding, Reasoning and Information Extraction)

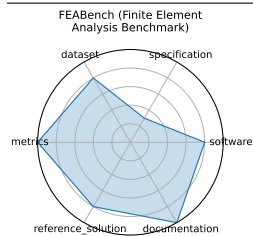
### 3.20 FEABench (Finite Element Analysis Benchmark)

Does not exist

**date:** 2023-01-26  
**version:** 1  
**last\_updated:** 2023-01-26  
**expired:** false  
**valid:** no  
**valid\_date:** 2023-01-26  
**url:** <https://github.com/google/feabench>  
**doi:** unknown  
**domain:** Computational Engineering  
**focus:** FEA simulation accuracy and performance  
**keywords:** - finite element - simulation - PDE  
**licensing:** unknown  
**task\_types:** - Simulation - Performance evaluation  
**ai\_capability\_measured:** - Numerical simulation accuracy and efficiency  
**metrics:** - Solve time - Error norm  
**models:** - FEniCS - deal.II  
**ml\_motif:** - unknown  
**type:** Benchmark  
**ml\_task:** - Supervised Learning  
**solutions:** unknown  
**notes:** OK  
**contact.name:** unknown  
**contact.email:** unknown  
**datasets.links.name:** unknown  
**datasets.links.url:** unknown  
**results.links.name:** unknown  
**results.links.url:** unknown  
**fair.reproducible:** True  
**fair.benchmark\_ready:** True  
**id:** feabench\_finite\_element\_analysis\_benchmark  
**Citations:** [23]

#### Ratings:

Rating	Value	Reason
dataset	4	Available, but not split into sets
documentation	5	In associated paper
metrics	5	Fully defined metrics
reference_solution	4	Three open-source models were used. No system constraints.
software	4	Code is available, but poorly documented
specification	1.5	Output is defined and task clarity is questionable



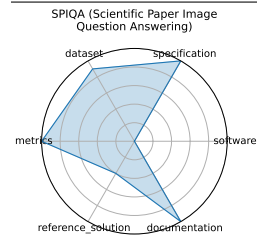
### 3.21 SPIQA (Scientific Paper Image Question Answering)

SPIQA assesses AI models' ability to interpret and answer questions about figures and tables in scientific papers by integrating visual and textual modalities with chain-of-thought reasoning.

<b>date:</b>	2024-07-12
<b>version:</b>	1
<b>last_updated:</b>	2024-07-12
<b>expired:</b>	false
<b>valid:</b>	yes
<b>valid_date:</b>	2024-07-12
<b>url:</b>	<a href="https://arxiv.org/abs/2407.09413">https://arxiv.org/abs/2407.09413</a>
<b>doi:</b>	10.48550/arXiv.2407.09413
<b>domain:</b>	Computer Science
<b>focus:</b>	Multimodal QA on scientific figures
<b>keywords:</b>	- multimodal QA - figure understanding - table comprehension - chain-of-thought
<b>licensing:</b>	Apache 2.0 License
<b>task_types:</b>	- Question answering - Multimodal QA - Chain-of-Thought evaluation
<b>ai_capability_measured:</b>	- Visual-textual reasoning in scientific contexts
<b>metrics:</b>	- Accuracy - F1 score
<b>models:</b>	- Chain-of-Thought models - Multimodal QA systems
<b>ml_motif:</b>	- Scientific paper reading
<b>type:</b>	Benchmark
<b>ml_task:</b>	- Supervised Learning
<b>solutions:</b>	0
<b>notes:</b>	Good
<b>contact.name:</b>	Subhashini Venugopalan
<b>contact.email:</b>	vsubhashini@google.com
<b>datasets.links.name:</b>	Hugging Face
<b>datasets.links.url:</b>	<a href="https://huggingface.co/datasets/google/spiqa">https://huggingface.co/datasets/google/spiqa</a>
<b>results.links.name:</b>	unknown
<b>results.links.url:</b>	unknown
<b>fair.reproducible:</b>	True
<b>fair.benchmark_ready:</b>	True
<b>id:</b>	spiqa_scientific_paper_image_question_answering
<b>Citations:</b>	[24]

#### Ratings:

Rating	Value	Reason
dataset	4.5	Dataset is available (via paper/appendix), includes train/test/valid split. FAIR-compliant with minor gaps in versioning or access standardization.
documentation	5	All information provided in paper
metrics	5	Uses quantitative metrics (Accuracy, F1) aligned with the task
reference_solution	2	Multiple model results (e.g., GPT-4V, Gemini) reported; baselines exist, but full runnable code not confirmed for all.
software	0	Not provided
specification	5	Task administration clearly defined; prompt instructions explicitly given, no ambiguity in format or scope.



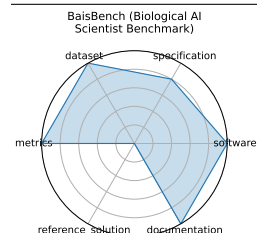
### 3.22 BaisBench (Biological AI Scientist Benchmark)

BaisBench evaluates AI scientists' ability to perform data-driven biological research by annotating cell types in single-cell datasets and answering MCQs derived from biological study insights, measuring autonomous scientific discovery.

**date:** 2025-05-13  
**version:** 1  
**last\_updated:** 2025-05-13  
**expired:** false  
**valid:** yes  
**valid\_date:** 2025-05-13  
**url:** <https://arxiv.org/abs/2505.08341>  
**doi:** 10.48550/arXiv.2505.08341  
**domain:** Computational Biology  
**focus:** Omics-driven AI research tasks  
**keywords:** - single-cell annotation - biological QA - autonomous discovery  
**licensing:** MIT License  
**task\_types:** - Cell type annotation - Multiple choice  
**ai\_capability\_measured:** - Autonomous biological research capabilities  
**metrics:** - Annotation accuracy - QA accuracy  
**models:** - LLM-based AI scientist agents  
**ml\_motif:** - Scientific research  
**type:** Benchmark  
**ml\_task:** - Supervised Learning  
**solutions:** 0  
**notes:** Underperforms human experts; aims to advance AI-driven discovery  
**contact.name:** Xuegong Zhang  
**contact.email:** zhangxg@mail.tsinghua.edu.cn  
**datasets.links.name:** Github  
**datasets.links.url:** <https://github.com/EperLuo/BaisBench>  
**results.links.name:** unknown  
**results.links.url:** unknown  
**fair.reproducible:** True  
**fair.benchmark\_ready:** True  
**id:** baisbench\_biological\_ai\_scientist\_benchmark  
**Citations:** [25]

#### Ratings:

Rating	Value	Reason
dataset	5	Uses public scRNA-seq datasets linked in paper appendix; structured and accessible, though versioning and full metadata not formalized per FAIR standards.
documentation	5	Dataset and paper accessible; IPYNB files for setup are available on the github repo.
metrics	5	Includes precise and interpretable metrics (annotation and QA accuracy); directly aligned with task outputs and benchmarking goals.
reference_solution	0	Model evaluations and LLM agent results discussed; however, no fully packaged, runnable baseline confirmed yet.
software	5	Instructions for environment setup available
specification	4	Task clearly defined-cell type annotation and biological QA; input/output formats are well-described; system constraints are not quantified.



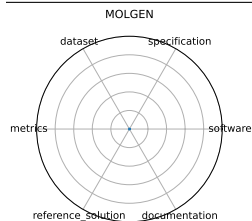
### 3.23 MOLGEN

MolGen is a pre-trained molecular language model that generates chemically valid molecules using SELFIES and reinforcement learning, guided by chemical feedback to optimize properties such as logP, QED, and docking score.

**date:** 2023-01-26  
**version:** 1  
**last\_updated:** 2023-01-26  
**expired:** false  
**valid:** yes  
**valid\_date:** 2023-01-26  
**url:** <https://github.com/zjunlp/MolGen>  
**doi:** 10.48550/arXiv.2301.11259  
**domain:** Computational Chemistry  
**focus:** Molecular generation and optimization  
**keywords:** - SELFIES - GAN - property optimization  
**licensing:** MIT License  
**task\_types:** - Distribution learning - Goal-oriented generation  
**ai\_capability\_measured:** - Generation of valid and optimized molecular structures  
**metrics:** - Validity% - Novelty% - QED - Docking score  
**models:** - MolGen  
**ml\_motif:** - Chemical generation  
**type:** Benchmark  
**ml\_task:** - Supervised Learning  
**solutions:** 0  
**notes:** This is a model, not a benchmark  
**contact.name:** unknown  
**contact.email:** unknown  
**datasets.links.name:** unknown  
**datasets.links.url:** unknown  
**results.links.name:** unknown  
**results.links.url:** unknown  
**fair.reproducible:** True  
**fair.benchmark\_ready:** True  
**id:** molgen  
**Citations:** [26]

#### Ratings:

Rating	Value	Reason
dataset	0	This is a pre-trained model
documentation	0	This is a pre-trained model
metrics	0	This is a pre-trained model
reference_solution	0	This is a pre-trained model
software	0	This is a pre-trained model
specification	0	This is a pre-trained model



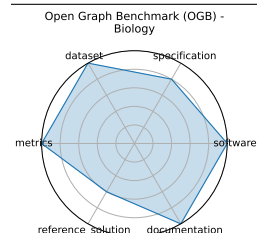
### 3.24 Open Graph Benchmark (OGB) - Biology

OGB-Biology is a suite of large-scale biological network datasets (protein-protein interaction, drug-target, etc.) with standardized splits and evaluation protocols for node, link, and graph property prediction tasks.

**date:** 2020-05-02  
**version:** 1  
**last\_updated:** 2020-05-02  
**expired:** false  
**valid:** yes  
**valid\_date:** 2020-05-02  
**url:** <https://ogb.stanford.edu/docs/home/>  
**doi:** 10.48550/arXiv.2005.00687  
**domain:** Graph ML  
**focus:** Biological graph property prediction  
**keywords:** - node prediction - link prediction - graph classification  
**licensing:** MIT License  
**task\_types:** - Node property prediction - Link property prediction - Graph property prediction  
**ai\_capability\_measured:** - Scalability and generalization in graph ML for biology  
**metrics:** - Accuracy - ROC-AUC  
**models:** - GCN - GraphSAGE - GAT  
**ml\_motif:** - Chemical biology  
**type:** Benchmark  
**ml\_task:** - Supervised Learning  
**solutions:** 0  
**notes:** Community-driven updates  
**contact.name:** OGB Team  
**contact.email:** [ogb@cs.stanford.edu](mailto:ogb@cs.stanford.edu)  
**datasets.links.name:** OGB Webpage  
**datasets.links.url:** [https://ogb.stanford.edu/docs/dataset\\_overview/](https://ogb.stanford.edu/docs/dataset_overview/)  
**results.links.name:** unknown  
**results.links.url:** unknown  
**fair.reproducible:** True  
**fair.benchmark\_ready:** True  
**id:** open\_graph\_benchmark\_ogb\_-\_biology  
**Citations:** [27]

#### Ratings:

Rating	Value	Reason
dataset	5	Fully FAIR- datasets are versioned, split, and accessible via a standardized API; extensive metadata and documentation are included.
documentation	5	All necessary information is included in a paper.
metrics	5	Reproducible, quantitative metrics (e.g., ROC-AUC, accuracy) that are tightly aligned with the tasks.
reference_solution	3	Multiple baselines implemented and documented (GCN, GAT, GraphSAGE). No constraints.
software	5	All necessary information is provided on the Github
specification	4	Tasks (node/link/graph property prediction) are clearly specified with input/output formats and standardized protocols; constraints (e.g., splits) are well-defined. No constraints.





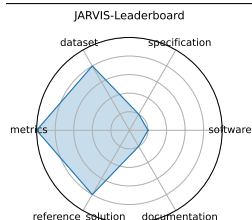
### 3.25 JARVIS-Leaderboard

JARVIS-Leaderboard is a community-driven platform benchmarking AI, electronic structure, force-fields, quantum computing, and experimental methods across hundreds of materials science tasks.

**date:** 2023-06-20  
**version:** 1  
**last\_updated:** 2023-06-20  
**expired:** false  
**valid:** yes  
**valid\_date:** 2023-06-20  
**url:** <https://arxiv.org/abs/2306.11688>  
**doi:** 10.48550/arXiv.2306.11688  
**domain:** Materials Science; Benchmarking  
**focus:** Comparative evaluation of materials design methods  
**keywords:** - leaderboards - materials methods - simulation  
**licensing:** NIST  
**task\_types:** - Method benchmarking - Leaderboard ranking  
**ai\_capability\_measured:** - Performance comparison across diverse materials design methods  
**metrics:** - MAE - RMSE - Accuracy  
**models:** - unknown  
**ml\_motif:** - Material science  
**type:** Benchmark  
**ml\_task:** - Supervised Learning  
**solutions:** 0  
**notes:** 1281 contributions across 274 benchmarks  
**contact.name:** Kamal Choudhary  
**contact.email:** [kamal.choudhary@nist.gov](mailto:kamal.choudhary@nist.gov)  
**datasets.links.name:** AI model specific benchmarks  
**datasets.links.url:** [https://pages.nist.gov/jarvis\\_leaderboard/AI/](https://pages.nist.gov/jarvis_leaderboard/AI/)  
**results.links.name:** unknown  
**results.links.url:** unknown  
**fair.reproducible:** True  
**fair.benchmark\_ready:** True  
**id:** jarvis-leaderboard  
**Citations:** [28]

#### Ratings:

Rating	Value	Reason
dataset	4	Data is public and adheres to FAIR principles across the NIST-hosted infrastructure; however, metadata completeness varies slightly across benchmarks. No splits.
documentation	1	Only the task is specified.
metrics	5	Metrics stated for each benchmark.
reference_solution	4	Many baselines across tasks (CGCNN, ALIGNN, M3GNet, etc.); no constraints specified.
software	1	Setup script provided, but no code provided
specification	1	Only dataset format is defined.



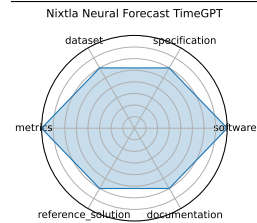
### 3.26 Nixtla Neural Forecast TimeGPT

TimeGPT is a transformer-based generative pretrained model on 100B+ time series data for zero-shot forecasting and anomaly detection via API .

<b>date:</b>	2023-10-05
<b>version:</b>	v3.0.2
<b>last_updated:</b>	2025-06
<b>expired:</b>	unknown
<b>valid:</b>	yes
<b>valid_date:</b>	2023-10-05
<b>url:</b>	<a href="https://github.com/Nixtla/neuralforecast">https://github.com/Nixtla/neuralforecast</a>
<b>doi:</b>	10.48550/arXiv.2310.03589
<b>domain:</b>	Time-series; General ML
<b>focus:</b>	Time-series foundation model "TimeGPT" for forecasting and anomaly detection
<b>keywords:</b>	- TimeGPT - foundation model - time-series - generative model
<b>licensing:</b>	Apache License 2.0
<b>task_types:</b>	- Time-series forecasting - Anomaly detection
<b>ai_capability_measured:</b>	- Zero-shot forecasting - anomaly detection
<b>metrics:</b>	- RMSE - Anomaly detection metrics
<b>models:</b>	- TimeGPT
<b>ml_motif:</b>	- Time-series
<b>type:</b>	Platform
<b>ml_task:</b>	- Forecasting
<b>solutions:</b>	Solution details are described in the referenced paper or repository.
<b>notes:</b>	Offered via Nixtla API and Azure Studio; enterprise-grade support available.
<b>contact.name:</b>	Azul Garza (Nixtla)
<b>contact.email:</b>	unknown
<b>results.links.name:</b>	ChatGPT LLM
<b>fair.reproducible:</b>	Yes
<b>fair.benchmark_ready:</b>	Yes
<b>id:</b>	nixtla_neural_forecast_timegpt
<b>Citations:</b>	[29]

#### Ratings:

Rating	Value	Reason
dataset	3	Evaluated on existing open datasets, but consolidated data release, splits, and FAIR metadata are not provided.
documentation	3	Basic README with installation and usage examples; more detailed API docs and tutorials would improve usability.
metrics	4	Uses standard forecasting metrics such as RMSE, MASE, SMAPE, and anomaly detection metrics consistently across evaluations.
reference_solution	3	TimeGPT implementation is available, but baseline comparisons and additional reference models are limited.
software	4	Fully open-source Apache 2.0 implementation integrated in NeuralForecast, supporting training and evaluation via API. Production-grade deployment available via Nixtla API and Azure.
specification	3	Concept and forecasting goals are described, but formal input/output definitions and task constraints are not rigorously specified.



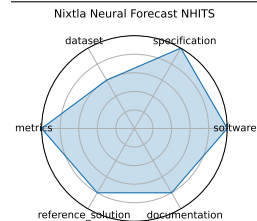
### 3.27 Nixtla Neural Forecast NHITS

NHITS (Neural Hierarchical Interpolation for Time Series) is a state-of-the-art model that improved accuracy by ~25% and reduced compute by 50x compared to Transformer baselines, using hierarchical interpolation and multi-rate sampling .

**date:** 2023-06-01  
**version:** v3.0.2  
**last\_updated:** 2025-06  
**expired:** unknown  
**valid:** yes  
**valid\_date:** 2023-06-01  
**url:** <https://github.com/Nixtla/neuralforecast>  
**doi:** unknown  
**domain:** Time-series; General ML  
**focus:** Official NHITS implementation for long-horizon time series forecasting  
**keywords:** - NHITS - long-horizon forecasting - neural interpolation - time-series  
**licensing:** Apache License 2.0  
**task\_types:** - Time-series forecasting  
**ai\_capability\_measured:** - Accuracy - compute efficiency for long series  
**metrics:** - RMSE - MAPE  
**models:** - NHITS  
**ml\_motif:** - Time-series  
**type:** Platform  
**ml\_task:** - Forecasting  
**solutions:** 0  
**notes:** Official implementation in NeuralForecast, included since its AAAI 2023 release.  
**contact.name:** Kin G. Olivares (Nixtla)  
**contact.email:** unknown  
**datasets.links.name:** Standard forecast datasets, M4  
**results.links.name:** ChatGPT LLM  
**fair.reproducible:** Yes  
**fair.benchmark\_ready:** Yes  
**id:** nixtla\_neural\_forecast\_nhits  
**Citations:** [30]

#### Ratings:

Rating	Value	Reason
dataset	3	Uses standard benchmark datasets like M4, but does not bundle them directly. FAIR compliance depends on external dataset sources and user setup.
documentation	4	Well-documented on GitHub and in AAAI paper, with code examples, training guidance, and usage tutorials. More model-specific docs could improve clarity further.
metrics	5	Evaluated using RMSE, MAPE, and other standard forecasting metrics, integrated into training and evaluation APIs.
reference_solution	4	Official NHITS implementation is fully reproducible with training/eval configs, though pretrained weights are not always provided.
software	5	Implemented within the open-source NeuralForecast library under Apache 2.0. Includes training, evaluation, and hyperparameter tuning pipelines. Actively maintained.
specification	5	The NHITS forecasting task is clearly defined with structured input/output formats. Model design targets long-horizon accuracy and compute efficiency.



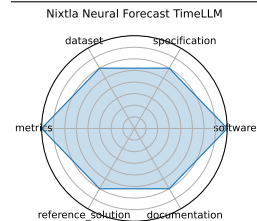
### 3.28 Nixtla Neural Forecast TimeLLM

Time-LLM uses reprogramming layers to adapt frozen LLMs for time series forecasting, treating forecasting as a language task .

**date:** 2023-10-03  
**version:** v3.0.2  
**last\_updated:** 2025-06  
**expired:** unknown  
**valid:** yes  
**valid\_date:** 2023-10-03  
**url:** <https://github.com/Nixtla/neuralforecast>  
**doi:** 10.48550/arXiv.2310.01728  
**domain:** Time-series; General ML  
**focus:** Reprogramming LLMs for time series forecasting  
**keywords:** - Time-LLM - language model - time-series - reprogramming  
**licensing:** Apache License 2.0  
**task\_types:** - Time-series forecasting  
**ai\_capability\_measured:** - Model reuse via LLM - few-shot forecasting  
**metrics:** - RMSE - MAPE  
**models:** - Time-LLM  
**ml\_motif:** - Time-series  
**type:** Platform  
**ml\_task:** - Forecasting  
**solutions:** Solution details are described in the referenced paper or repository.  
**notes:** Fully open-source; transforms forecasting using LLM text reconstruction.  
**contact.name:** Ming Jin (Nixtla)  
**contact.email:** unknown  
**datasets.links.name:** Standard forecast datasets, M4  
**results.links.name:** ChatGPT LLM  
**fair.reproducible:** Yes  
**fair.benchmark\_ready:** Yes  
**id:** nixtla\_neural\_forecast\_timellm  
**Citations:** [31]

#### Ratings:

Rating	Value	Reason
dataset	3	Evaluated on standard datasets like M4 and ETT, but dataset splits and versioning are not bundled or explicitly FAIR-compliant.
documentation	3	GitHub README provides installation and quick usage examples, but lacks detailed API docs, training walkthroughs, or extended tutorials.
metrics	4	Standard forecasting metrics such as RMSE, MAPE, and SMAPE are reported. Evaluation is consistent, though deeper metric justification is limited.
reference_solution	3	Time-LLM implementation is open and reproducible, but limited baselines or comparative implementations are included directly.
software	4	Fully open-source under Apache 2.0, integrated into the NeuralForecast library. Includes Time-LLM implementation with example usage and training scripts.
specification	3	High-level framing of forecasting as language modeling is clear, but detailed input/output specifications, constraints, and task formalization are minimal.



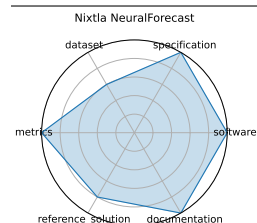
### 3.29 Nixtla NeuralForecast

NeuralForecast offers scalable, user-friendly implementations of over 30 neural forecasting models (NBEATS, NHITS, TFT, DeepAR, etc.), emphasizing quality, usability, interpretability, and performance.

**date:** 2022-04-01  
**version:** v3.0.2  
**last\_updated:** 2025-06  
**expired:** unknown  
**valid:** yes  
**valid\_date:** 2022-04-01  
**url:** <https://github.com/Nixtla/neuralforecast>  
**doi:** unknown  
**domain:** Time-series forecasting; General ML  
**focus:** High-performance neural forecasting library with >30 models  
**keywords:** - time-series - neural forecasting - NBEATS, NHITS, TFT - probabilistic forecasting - usability  
**licensing:** Apache License 2.0  
**task\_types:** - Time-series forecasting  
**ai\_capability\_measured:** - Forecast accuracy - interpretability - speed  
**metrics:** - RMSE - MAPE - CRPS  
**models:** - NBEATS - NHITS - TFT - DeepAR  
**ml\_motif:** - Time-series  
**type:** Platform  
**ml\_task:** - Forecasting  
**solutions:** 0  
**notes:** AutoModel supports hyperparameter tuning and distributed execution via Ray and Optuna. First official NHITS implementation. contentReference oaicite:4 ndex=4  
**contact.name:** Kin G. Olivares (Nixtla)  
**contact.email:** unknown  
**results.links.name:** ChatGPT LLM  
**fair.reproducible:** Yes  
**fair.benchmark\_ready:** Yes  
**id:** nixtla\_neuralforecast  
**Citations:** [32]

#### Ratings:

Rating	Value	Reason
dataset	3	NeuralForecast does not include its own datasets but supports standard datasets (e.g., M4, M5, ETT). FAIR compliance depends on user-supplied data.
documentation	5	Rich documentation with examples, API references, tutorials, notebooks, and CLI support. PyPI, GitHub, and official blog posts offer clear guidance for usage and extension.
metrics	5	RMSE, MAPE, CRPS, and other domain-relevant metrics are well supported and integrated into the evaluation loop.
reference_solution	4	Includes runnable model baselines and training scripts for all supported models. Some models have pretrained weights, but not all are fully benchmarked out-of-the-box.
software	5	Actively maintained open-source library under Apache 2.0. Offers a clean API, extensive model zoo (>30 models), integration with Ray, Optuna, and supports scalable training and inference workflows.
specification	5	Forecasting task is well-defined with clear input/output structures. Framework supports probabilistic and deterministic forecasting, with unified interfaces and support for batch evaluation.



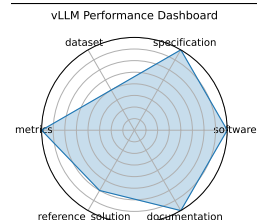
### 3.30 vLLM Performance Dashboard

A live visual dashboard for vLLM showcasing throughput, latency, and other inference metrics across models and hardware configurations.

<b>date:</b>	2022-06-22
<b>version:</b>	v1.0
<b>last_updated:</b>	2025-01
<b>expired:</b>	unknown
<b>valid:</b>	yes
<b>valid_date:</b>	2022-06-22
<b>url:</b>	<a href="https://simon-mo-workspace.observablehq.cloud/vllm-dashboard-v0/">https://simon-mo-workspace.observablehq.cloud/vllm-dashboard-v0/</a>
<b>doi:</b>	unknown
<b>domain:</b>	LLM; HPC/inference
<b>focus:</b>	Interactive dashboard showing inference performance of vLLM
<b>keywords:</b>	- Dashboard - Throughput visualization - Latency analysis - Metric tracking
<b>licensing:</b>	unknown
<b>task_types:</b>	- Performance visualization
<b>ai_capability_measured:</b>	- Throughput - latency - hardware utilization
<b>metrics:</b>	- Tokens/sec - TTFT - Memory usage
<b>models:</b>	- LLaMA-2 - Mistral - Qwen
<b>ml_motif:</b>	- HPC/inference
<b>type:</b>	Framework
<b>ml_task:</b>	- Visualization
<b>solutions:</b>	0
<b>notes:</b>	Built using ObservableHQ; integrates live data from vLLM benchmarks. The URL requires a login to access the content.
<b>contact.name:</b>	Simon Mo
<b>contact.email:</b>	unknown
<b>results.links.name:</b>	ChatGPT LLM
<b>fair.reproducible:</b>	Yes
<b>fair.benchmark_ready:</b>	Yes
<b>id:</b>	vllm_performance_dashboard
<b>Citations:</b>	[33]

#### Ratings:

Rating	Value	Reason
dataset	2	No datasets are bundled; the dashboard visualizes metrics derived from model inference logs or external endpoints, not a formal dataset.
documentation	4	Public dashboard with instructions and tooltips; documentation is clear, though access is restricted (login required) and backend setup is opaque to users.
metrics	4	Tracks tokens/sec, TTFT, memory usage, and platform comparisons. Metrics are clear but focused on visualization rather than statistical robustness.
reference_solution	3	Dashboards include reproducible views of benchmarked models, but do not ship with runnable model code. Relies on external serving infrastructure.
software	4	Interactive dashboard built with ObservableHQ and linked to vLLM benchmarks. Source code is not fully open, but backend integration with vLLM is well-maintained.
specification	4	While primarily a visualization tool, it includes benchmark configurations, metric definitions, and supports comparison across models and hardware.



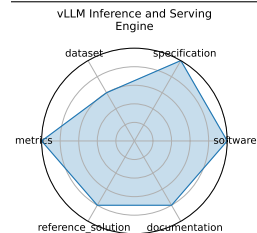
### 3.31 vLLM Inference and Serving Engine

vLLM is a fast, high-throughput, memory-efficient inference and serving engine for large language models, featuring PagedAttention, continuous batching, and support for quantized and pipelined model execution. Benchmarks compare it to TensorRT-LLM, SGLang, and others.

<b>date:</b>	2023-09-12
<b>version:</b>	v0.10.0
<b>last_updated:</b>	2025-06
<b>expired:</b>	unknown
<b>valid:</b>	yes
<b>valid_date:</b>	2023-09-12
<b>url:</b>	<a href="https://github.com/vllm-project/vllm/tree/main/benchmarks">https://github.com/vllm-project/vllm/tree/main/benchmarks</a>
<b>doi:</b>	unknown
<b>domain:</b>	LLM; HPC/inference
<b>focus:</b>	High-throughput, memory-efficient inference and serving engine for LLMs
<b>keywords:</b>	- LLM inference - PagedAttention - CUDA graph - streaming API - quantization
<b>licensing:</b>	Apache License 2.0
<b>task_types:</b>	- Inference Benchmarking
<b>ai_capability_measured:</b>	- Throughput - latency - memory efficiency
<b>metrics:</b>	- Tokens/sec - Time to First Token (TTFT) - Memory footprint
<b>models:</b>	- LLaMA - Mixtral - FlashAttention-based models
<b>ml_motif:</b>	- HPC/inference
<b>type:</b>	Framework
<b>ml_task:</b>	- Inference
<b>solutions:</b>	0
<b>notes:</b>	Incubated by LF AI and Data; achieves up to 24x throughput over HuggingFace Transformers
<b>contact.name:</b>	Woosuk Kwon (vLLM Team)
<b>contact.email:</b>	unknown
<b>results.links.name:</b>	ChatGPT LLM
<b>fair.reproducible:</b>	Yes
<b>fair.benchmark_ready:</b>	Yes
<b>id:</b>	vllm_inference_and_serving_engine
<b>Citations:</b>	[34]

#### Ratings:

Rating	Value	Reason
dataset	3	No traditional dataset is included. Instead, it uses structured configs and logs suitable for inference benchmarking. FAIR principles are only partially applicable.
documentation	4	Well-structured GitHub documentation with setup instructions, config examples, benchmarking comparisons, and performance tuning guides.
metrics	5	Comprehensive performance metrics like tokens/sec, time-to-first-token (TTFT), and memory footprint are consistently applied and benchmarked across frameworks.
reference_solution	4	Provides runnable scripts and configs for several models (LLaMA, Mixtral, etc.) across platforms. Baselines are reproducible, though not all models are fully wrapped or hosted.
software	5	Actively maintained open-source project under Apache 2.0. GitHub repo includes full serving engine, benchmarking scripts, CUDA integration, and deployment examples.
specification	5	Inference benchmarks are well-defined with clear input/output formats and platform-specific constraints. Covers multiple models, hardware backends, and batching configurations.



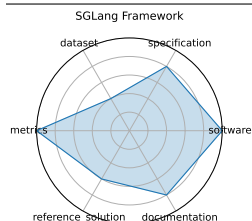
### 3.32 SGLang Framework

A high-performance open-source serving framework combining efficient backend runtime (RadixAttention, batching, quantization) and expressive frontend language, boosting LLM/VLM inference throughput up to  $\sim 3\times$  over alternatives.

<b>date:</b>	2023-12-12
<b>version:</b>	v0.4.9
<b>last_updated:</b>	2025-06
<b>expired:</b>	unknown
<b>valid:</b>	yes
<b>valid_date:</b>	2023-12-12
<b>url:</b>	<a href="https://github.com/sgl-project/sglang/tree/main/benchmark">https://github.com/sgl-project/sglang/tree/main/benchmark</a>
<b>doi:</b>	10.48550/arXiv.2312.07104
<b>domain:</b>	LLM Vision
<b>focus:</b>	Fast serving framework for LLMs and vision-language models
<b>keywords:</b>	- LLM serving - vision-language - RadixAttention - performance - JSON decoding
<b>licensing:</b>	Apache License 2.0
<b>task_types:</b>	- Model serving framework
<b>ai_capability_measured:</b>	- Serving throughput - JSON/task-specific latency
<b>metrics:</b>	- Tokens/sec - Time-to-first-token - Throughput gain vs baseline
<b>models:</b>	- LLaVA - DeepSeek - Llama
<b>ml_motif:</b>	- LLM Vision
<b>type:</b>	Framework
<b>ml_task:</b>	- Model serving
<b>solutions:</b>	Solution details are described in the referenced paper or repository.
<b>notes:</b>	Deployed in production (xAI, NVIDIA, Google Cloud); v0.4.8 release June 2025.
<b>contact.name:</b>	SGLang Team
<b>contact.email:</b>	unknown
<b>datasets.links.name:</b>	Benchmark configs
<b>results.links.name:</b>	ChatGPT LLM
<b>fair.reproducible:</b>	Yes
<b>fair.benchmark_ready:</b>	Yes
<b>id:</b>	sglang_framework
<b>Citations:</b>	[35]

#### Ratings:

Rating	Value	Reason
dataset	2	Does not introduce new datasets; instead, it evaluates performance using existing model benchmarks. Only configuration files are included.
documentation	4	Strong GitHub documentation, install guides, and benchmarks. Some advanced topics (e.g., scaling, hardware tuning) could use deeper walkthroughs.
metrics	5	Serving-related metrics such as tokens/sec, time-to-first-token, and throughput gain vs. baselines are well-defined and consistently applied.
reference_solution	3	Provides benchmark configs and example integrations (e.g., with LLaVA, DeepSeek), but not all models or scripts are runnable out-of-the-box.
software	5	Actively maintained and production-deployed (e.g., xAI, NVIDIA); source code available under Apache 2.0. Includes efficient backends (RadixAttention, quantization, batching) and full serving infrastructure.
specification	4	The framework clearly defines performance targets, serving logic, and model integration. Input/output expectations are consistent, but not all benchmarks are standardized.





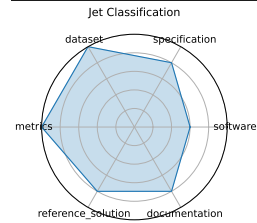
### 3.33 Jet Classification

This benchmark evaluates ML models for real-time classification of particle jets using high-level features derived from simulated LHC data. It includes both full-precision and quantized models optimized for FPGA deployment.

**date:** 2024-05-01  
**version:** v0.2.0  
**last\_updated:** 2024-05  
**expired:** unknown  
**valid:** yes  
**valid\_date:** 2024-05-01  
**url:** <https://github.com/fastmachinelearning/fastml-science/tree/main/jet-classify>  
**doi:** 10.48550/arXiv.2207.07958  
**domain:** Particle Physics  
**focus:** Real-time classification of particle jets using HL-LHC simulation features  
**keywords:** - classification - real-time ML - jet tagging - QKeras  
**licensing:** Apache License 2.0  
**task\_types:** - Classification  
**ai\_capability\_measured:** - Real-time inference - model compression performance  
**metrics:** - Accuracy - AUC  
**models:** - Keras DNN - QKeras quantized DNN  
**ml\_motif:** - Real-time  
**type:** Benchmark  
**ml\_task:** - Supervised Learning  
**solutions:** Solution details are described in the referenced paper or repository.  
**notes:** Includes both float and quantized models using QKeras  
**contact.name:** Jules Muhizi  
**contact.email:** unknown  
**datasets.links.name:** JetClass  
**datasets.links.url:** <https://zenodo.org/record/6619768>  
**results.links.name:** ChatGPT LLM  
**results.links.url:** [https://docs.google.com/document/d/1runrcij-eoH3\\_lgGZ8wm2z1YbL1Qf5cSNbVbHyWFDs4](https://docs.google.com/document/d/1runrcij-eoH3_lgGZ8wm2z1YbL1Qf5cSNbVbHyWFDs4)  
**fair.reproducible:** True  
**fair.benchmark\_ready:** True  
**id:** jet\_classification  
**Citations:** [36]

#### Ratings:

Rating	Value	Reason
dataset	5	None
documentation	4	Full reproducibility requires manual setup
metrics	5	None
reference_solution	4	HW/SW requirements missing; Reference not bundled as official starter kit
software	3	Not containerized; Setup automation/documentation could be improved
specification	4	System constraints missing



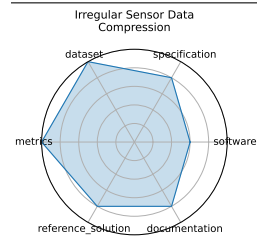
### 3.34 Irregular Sensor Data Compression

This benchmark addresses lossy compression of irregularly sampled sensor data from particle detectors using real-time autoencoder architectures, targeting latency-critical applications in physics experiments.

<b>date:</b>	2024-05-01
<b>version:</b>	v0.2.0
<b>last_updated:</b>	2024-05
<b>expired:</b>	unknown
<b>valid:</b>	yes
<b>valid_date:</b>	2024-05-01
<b>url:</b>	<a href="https://github.com/fastmachinelearning/fastml-science/tree/main/sensor-data-compression">https://github.com/fastmachinelearning/fastml-science/tree/main/sensor-data-compression</a>
<b>doi:</b>	10.48550/arXiv.2207.07958
<b>domain:</b>	Particle Physics
<b>focus:</b>	Real-time compression of sparse sensor data with autoencoders
<b>keywords:</b>	- compression - autoencoder - sparse data - irregular sampling
<b>licensing:</b>	Apache License 2.0
<b>task_types:</b>	- Compression
<b>ai_capability_measured:</b>	- Reconstruction quality - compression efficiency
<b>metrics:</b>	- MSE - Compression ratio
<b>models:</b>	- Autoencoder - Quantized autoencoder
<b>ml_motif:</b>	- Real-time, Image/CV
<b>type:</b>	Benchmark
<b>ml_task:</b>	- Unsupervised Learning
<b>solutions:</b>	Solution details are described in the referenced paper or repository.
<b>notes:</b>	Based on synthetic but realistic physics sensor data
<b>contact.name:</b>	Ben Hawks, Nhan Tran
<b>contact.email:</b>	unknown
<b>datasets.links.name:</b>	Custom synthetic irregular sensor dataset
<b>datasets.links.url:</b>	<a href="https://github.com/fastmachinelearning/fastml-science/tree/main/sensor-data-compression">https://github.com/fastmachinelearning/fastml-science/tree/main/sensor-data-compression</a>
<b>results.links.name:</b>	ChatGPT LLM
<b>fair.reproducible:</b>	True
<b>fair.benchmark_ready:</b>	True
<b>id:</b>	irregular_sensor_data_compression
<b>Citations:</b>	[37]

#### Ratings:

Rating	Value	Reason
dataset	5	All criteria met
documentation	4	Setup for deployment (e.g., FPGA pipeline) requires familiarity with tooling
metrics	5	All criteria met
reference_solution	4	Not fully documented or automated for reproducibility
software	3	Not containerized; Full automation and documentation could be improved
specification	4	Exact latency or resource constraints not numerically specified



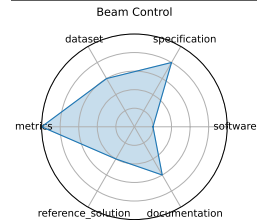
### 3.35 Beam Control

Beam Control explores real-time reinforcement learning strategies for maintaining stable beam trajectories in particle accelerators. The benchmark is based on the BOOSTR environment for accelerator simulation.

<b>date:</b>	2024-05-01
<b>version:</b>	v0.2.0
<b>last_updated:</b>	2024-05
<b>expired:</b>	unknown
<b>valid:</b>	yes
<b>valid_date:</b>	2024-05-01
<b>url:</b>	<a href="https://github.com/fastmachinelearning/fastml-science/tree/main/beam-control">https://github.com/fastmachinelearning/fastml-science/tree/main/beam-control</a>
<b>doi:</b>	10.48550/arXiv.2207.07958
<b>domain:</b>	Accelerators and Magnets
<b>focus:</b>	Reinforcement learning control of accelerator beam position
<b>keywords:</b>	- RL - beam stabilization - control systems - simulation
<b>licensing:</b>	Apache License 2.0
<b>task_types:</b>	- Control
<b>ai_capability_measured:</b>	- Policy performance in simulated accelerator control
<b>metrics:</b>	- Stability - Control loss
<b>models:</b>	- DDPG - PPO (planned)
<b>ml_motif:</b>	- Real-time, RL
<b>type:</b>	Benchmark
<b>ml_task:</b>	- Reinforcement Learning
<b>solutions:</b>	Solution details are described in the referenced paper or repository.
<b>notes:</b>	Environment defined, baseline RL implementation is in progress
<b>contact.name:</b>	Ben Hawks, Nhan Tran
<b>contact.email:</b>	unknown
<b>results.links.name:</b>	ChatGPT LLM
<b>fair.reproducible:</b>	in progress
<b>fair.benchmark_ready:</b>	in progress
<b>id:</b>	beam_control
<b>Citations:</b>	[38], [39]

#### Ratings:

Rating	Value	Reason
dataset	3	Not findable (no DOI/indexing); Not interoperable (format/schema unspecified)
documentation	3	Setup instructions and pretrained model details are missing
metrics	5	All criteria met
reference_solution	2	HW/SW requirements missing; Metrics not evaluated with reference; Baseline not trainable/open
software	1	Code not documented; Incomplete setup and not containerized
specification	4	Latency/resource constraints not fully quantified



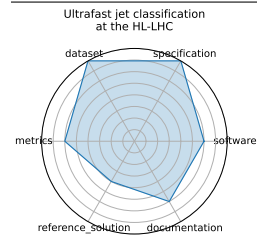
### 3.36 Ultrafast jet classification at the HL-LHC

Demonstrates three ML models (MLP, Deep Sets, Interaction Networks) optimized for FPGA deployment with O(100 ns) inference using quantized models and hls4ml, targeting real-time jet tagging in the L1 trigger environment at the high-luminosity LHC. Data is available on Zenodo DOI:10.5281/zenodo.3602260.

**date:** 2024-07-08  
**version:** v1.0  
**last\_updated:** 2024-07  
**expired:** unknown  
**valid:** yes  
**valid\_date:** 2024-07-08  
**url:** <https://arxiv.org/pdf/2402.01876>  
**doi:** 10.48550/arXiv.2402.01876  
**domain:** Particle Physics  
**focus:** FPGA-optimized real-time jet origin classification at the HL-LHC  
**keywords:** - jet classification - FPGA - quantization-aware training - Deep Sets - Interaction Networks  
**licensing:** CC-BY  
**task\_types:** - Classification  
**ai\_capability\_measured:** - Real-time inference under FPGA constraints  
**metrics:** - Accuracy - Latency - Resource utilization  
**models:** - MLP - Deep Sets - Interaction Network  
**ml\_motif:** - Real-time  
**type:** Model  
**ml\_task:** - Supervised Learning  
**solutions:** Solution details are described in the referenced paper or repository.  
**notes:** Uses quantization-aware training; hardware synthesis evaluated via hls4ml  
**contact.name:** Patrick Odagiu  
**contact.email:** podagiu@ethz.ch  
**datasets.links.name:** Zenodo dataset  
**datasets.links.url:** <https://zenodo.org/records/3602260>  
**results.links.name:** ChatGPT LLM  
**results.links.url:** [https://docs.google.com/document/d/1gDf1CIYtfmfZ9urv1jCRZMYz\\_3WwEETkugUC65OZBdw](https://docs.google.com/document/d/1gDf1CIYtfmfZ9urv1jCRZMYz_3WwEETkugUC65OZBdw)  
**fair.reproducible:** True  
**fair.benchmark\_ready:** False  
**id:** ultrafast\_jet\_classification\_at\_the\_hl-lhc  
**Citations:** [40]

#### Ratings:

Rating	Value	Reason
dataset	4	FAIR metadata limited; no clear mention of dataset format or splits
documentation	3	No linked GitHub repo or setup instructions; paper provides partial guidance only
metrics	3	Metrics exist (accuracy, latency, utilization), but formal definitions and evaluation guidance are limited
reference_solution	2	Reference implementations not fully reproducible; no evaluation pipeline or training setup provided
software	3	Not containerized; Setup and automation incomplete
specification	4	Hardware constraints are referenced but not fully detailed or standardized



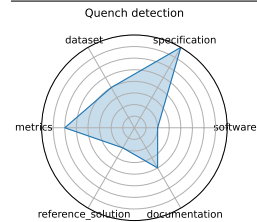
### 3.37 Quench detection

Exploration of real-time quench detection using unsupervised and RL approaches, combining multi-modal sensor data (BPM, power supply, acoustic), operating on kHz-MHz streams with anomaly detection and frequency-domain features.

**date:** 2024-10-15  
**version:** v1.0  
**last\_updated:** 2024-10  
**expired:** no  
**valid:** yes  
**valid\_date:** 2024-10-15  
**url:** [https://indico.cern.ch/event/1387540/contributions/6153618/attachments/2948441/5182077/fast\\_ml\\_magnets\\_](https://indico.cern.ch/event/1387540/contributions/6153618/attachments/2948441/5182077/fast_ml_magnets_)  
**doi:** NA  
**domain:** Accelerators and Magnets  
**focus:** Real-time detection of superconducting magnet quenches using ML  
**keywords:** - quench detection - autoencoder - anomaly detection - real-time  
**licensing:** Via Fermilab  
**task\_types:** - Anomaly detection - Quench localization  
**ai\_capability\_measured:** - Real-time anomaly detection with multi-modal sensors  
**metrics:** - ROC-AUC - Detection latency  
**models:** - Autoencoder - RL agents (in development)  
**ml\_motif:** - Real-time, RL  
**type:** Benchmark  
**ml\_task:** - Reinforcement, Unsupervised Learning  
**solutions:** 0  
**notes:** Precursor detection in progress; multi-modal and dynamic weighting methods  
**contact.name:** Maira Khan  
**contact.email:** unknown  
**datasets.links.name:** BPM and power supply data from BNL  
**results.links.name:** ChatGPT LLM  
**fair.reproducible:** in progress  
**fair.benchmark\_ready:** False  
**id:** quench\_detection  
**Citations:** [41]

#### Ratings:

Rating	Value	Reason
dataset	2	Dataset URL is missing; FAIR principles largely unmet
documentation	2	Only a conference slide deck is available; lacks detailed instructions or repository for reproduction
metrics	3	ROC-AUC and latency are mentioned, but metric definitions and formal evaluation setup are missing
reference_solution	1	No baseline or reproducible model implementation available
software	1	Code not provided; no evidence of documentation or containerization
specification	4	Real-time detection task is clearly described, but exact constraints, inputs/outputs, and evaluation protocol are only partially specified



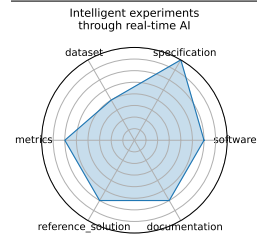
### 3.38 Intelligent experiments through real-time AI

Research and Development demonstrator for real-time processing of high-rate tracking data from the sPHENIX detector (RHIC) and future EIC systems. Uses GNNs with hls4ml for FPGA-based trigger generation to identify rare events (heavy flavor, DIS electrons) within 10 micros latency. Demonstrated improved accuracy and latency on Alveo/FELIX platforms.

**date:** 2025-01-08  
**version:** v1.0  
**last\_updated:** 2025-01  
**expired:** unknown  
**valid:** yes  
**valid\_date:** 2025-01-08  
**url:** <https://arxiv.org/pdf/2501.04845>  
**doi:** 10.48550/arXiv.2501.04845  
**domain:** Instrumentation and Detectors; Nuclear Physics; Particle Physics  
**focus:** Real-time FPGA-based triggering and detector control for sPHENIX and future EIC  
**keywords:** - FPGA - Graph Neural Network - hls4ml - real-time inference - detector control  
**licensing:** CC BY-NC-ND 4.0  
**task\_types:** - Trigger classification - Detector control - Real-time inference  
**ai\_capability\_measured:** - Low-latency GNN inference on FPGA  
**metrics:** - Accuracy (charm and beauty detection) - Latency (micros) - Resource utilization (LUT/FF/BRAM/DSP)  
**models:** - Bipartite Graph Network with Set Transformers (BGN-ST) - GarNet (edge-classifier)  
**ml\_motif:** - Real-time  
**type:** Model  
**ml\_task:** - Supervised Learning  
**solutions:** Solution details are described in the referenced paper or repository.  
**notes:** Achieved ~97.4% accuracy for beauty decay triggers; sub-10 micros latency on Alveo U280; hit-based FPGA design via hls4ml and FlowGNN.  
**contact.name:** Jakub Kvapil  
**contact.email:** Jakub.Kvapil@lanl.gov  
**datasets.links.name:** Internal simulated tracking data (sPHENIX and EIC DIS-electron tagger)  
**results.links.name:** ChatGPT LLM  
**fair.reproducible:** True  
**fair.benchmark\_ready:** False  
**id:** intelligent\_experiments\_through\_real-time\_ai  
**Citations:** [42]

#### Ratings:

Rating	Value	Reason
dataset	2	Dataset is internal and not publicly available or FAIR-compliant
documentation	3	No public GitHub or complete pipeline documentation
metrics	3	Metrics relevant but not supported by evaluation scripts or baselines
reference_solution	3	No public or reproducible implementation released
software	3	No containerized or open-source setup provided
specification	4	Architectural/system specifications are incomplete



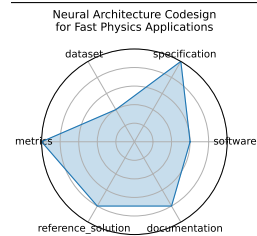
### 3.39 Neural Architecture Codesign for Fast Physics Applications

Introduces a two-stage neural architecture codesign (NAC) pipeline combining global and local search, quantization-aware training, and pruning to design efficient models for fast Bragg peak finding and jet classification, synthesized for FPGA deployment with hls4ml. Achieves  $>30\times$  reduction in BOPs and sub-100 ns inference latency on FPGA.

**date:** 2025-01-09  
**version:** v1.0  
**last\_updated:** 2025-01  
**expired:** unknown  
**valid:** yes  
**valid\_date:** 2025-01-09  
**url:** <https://arxiv.org/abs/2501.05515>  
**doi:** 10.48550/arXiv.2501.05515  
**domain:** Physics; Materials Science; Particle Physics  
**focus:** Automated neural architecture search and hardware-efficient model codesign for fast physics applications  
**keywords:** - neural architecture search - FPGA deployment - quantization - pruning - hls4ml  
**licensing:** Via Fermilab  
**task\_types:** - Classification - Peak finding  
**ai\_capability\_measured:** - Hardware-aware model optimization; low-latency inference  
**metrics:** - Accuracy - Latency - Resource utilization  
**models:** - NAC-based BraggNN - NAC-optimized Deep Sets (jet)  
**ml\_motif:** - Real-time, Image/CV  
**type:** Framework  
**ml\_task:** - Supervised Learning  
**solutions:** Solution details are described in the referenced paper or repository.  
**notes:** Demonstrated two case studies (materials science, HEP); pipeline and code open-sourced.  
**contact.name:** Jason Weitz (UCSD), Nhan Tran (FNAL)  
**contact.email:** unknown  
**results.links.name:** ChatGPT LLM  
**fair.reproducible:** Yes (nac-opt, hls4ml)  
**fair.benchmark\_ready:** False  
**id:** neural\_architecture\_codesign\_for\_fast\_physics\_applications  
**Citations:** [43]

#### Ratings:

Rating	Value	Reason
dataset	2	Simulated datasets referenced but not publicly available or FAIR-compliant
documentation	4	Detailed paper and tools described; open repo planned but not yet complete
metrics	5	Clear, quantitative metrics aligned with task goals and hardware evaluation
reference_solution	4	Models tested on hardware with source code references; full training pipeline not yet released
software	3	Toolchain (hls4ml, nac-opt) described but not yet containerized or fully packaged
specification	5	Fully specified task with constraints and target deployment; includes hardware context



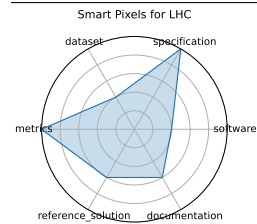
### 3.40 Smart Pixels for LHC

Presents a 256x256-pixel ROIC in 28 nm CMOS with embedded 2-layer NN for cluster filtering at 25 ns, achieving 54-75% data reduction while maintaining noise and latency constraints. Prototype consumes ~300 microW/pixel and operates in combinatorial digital logic.

**date:** 2024-06-24  
**version:** v1.0  
**last\_updated:** 2024-06  
**expired:** unknown  
**valid:** yes  
**valid\_date:** 2024-06-24  
**url:** <https://arxiv.org/abs/2406.14860>  
**doi:** 10.48550/arXiv.2406.14860  
**domain:** Particle Physics; Instrumentation and Detectors  
**focus:** On-sensor, in-pixel ML filtering for high-rate LHC pixel detectors  
**keywords:** - smart pixel - on-sensor inference - data reduction - trigger  
**licensing:** Via Fermilab  
**task\_types:** - Image Classification - Data filtering  
**ai\_capability\_measured:** - On-chip - low-power inference; data reduction  
**metrics:** - Data rejection rate - Power per pixel  
**models:** - 2-layer pixel NN  
**ml\_motif:** - Real-time, Image/CV  
**type:** Benchmark  
**ml\_task:** - Image Classification  
**solutions:** Solution details are described in the referenced paper or repository.  
**notes:** Prototype in CMOS 28 nm; proof-of-concept for Phase III pixel upgrades.  
**contact.name:** Lindsey Gray; Jennet Dickinson  
**contact.email:** unknown  
**results.links.name:** ChatGPT LLM  
**fair.reproducible:** True  
**fair.benchmark\_ready:** Yes (Zenodo:7331128)  
**id:** smart\_pixels\_for\_lhc  
**Citations:** [44]

#### Ratings:

Rating	Value	Reason
dataset	2	No dataset links; not publicly hosted or FAIR-compliant
documentation	3	Paper contains detailed descriptions, but no repo or external guide for reproducing results
metrics	5	None
reference_solution	3	In-pixel 2-layer NN described and evaluated, but reproducibility and source files are not released
software	2	No packaged code or setup scripts available; replication depends on hardware description and paper
specification	5	None





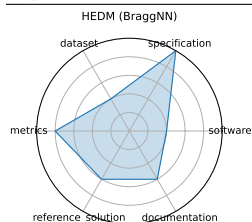
### 3.41 HEDM (BraggNN)

Uses BraggNN, a deep neural network, for rapid Bragg peak localization in high-energy diffraction microscopy, achieving about 13x speedup compared to Voigt-based methods while maintaining sub-pixel accuracy.

**date:** 2023-10-03  
**version:** v1.0  
**last\_updated:** 2023-10  
**expired:** unknown  
**valid:** yes  
**valid\_date:** 2023-10-03  
**url:** <https://arxiv.org/abs/2008.08198>  
**doi:** 10.48550/arXiv.2008.08198  
**domain:** Material Science  
**focus:** Fast Bragg peak analysis using deep learning in diffraction microscopy  
**keywords:** - BraggNN - diffraction - peak finding - HEDM  
**licensing:** DOE Public Access Plan  
**task\_types:** - Peak detection  
**ai\_capability\_measured:** - High-throughput peak localization  
**metrics:** - Localization accuracy - Inference time  
**models:** - BraggNN  
**ml\_motif:** - Real-time, Image/CV  
**type:** Framework  
**ml\_task:** - Peak finding  
**solutions:** Solution details are described in the referenced paper or repository.  
**notes:** Enables real-time HEDM workflows; basis for NAC case study.  
**contact.name:** Jason Weitz (UCSD)  
**contact.email:** unknown  
**results.links.name:** ChatGPT LLM  
**fair.reproducible:** True  
**fair.benchmark\_ready:** True  
**id:** hedm\_braggmn  
**Citations:** [45]

#### Ratings:

Rating	Value	Reason
dataset	2	No dataset links or FAIR metadata; unclear public access
documentation	3	Paper is clear, but lacks a GitHub repo or full reproducibility pipeline
metrics	4	Only localization accuracy and inference time mentioned; not formally benchmarked with scripts
reference_solution	3	BraggNN model is described and evaluated, but no direct implementation or inference scripts available
software	2	No standalone code repository or setup instructions provided
specification	5	None



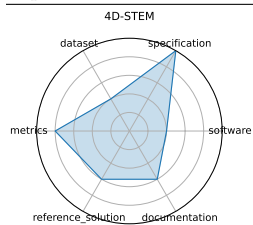
### 3.42 4D-STEM

Proposes ML methods for real-time analysis of 4D scanning transmission electron microscopy datasets; framework details in progress.

**date:** 2023-12-03  
**version:** v1.0  
**last\_updated:** 2023-12  
**expired:** unknown  
**valid:** yes  
**valid\_date:** 2023-12-03  
**url:** <https://openreview.net/pdf?id=7yt3N0o0W9>  
**doi:** unknown  
**domain:** Material Science  
**focus:** Real-time ML for scanning transmission electron microscopy  
**keywords:** - 4D-STEM - electron microscopy - real-time - image processing  
**licensing:** unknown  
**task\_types:** - Image Classification - Streamed data inference  
**ai\_capability\_measured:** - Real-time large-scale microscopy inference  
**metrics:** - Classification accuracy - Throughput  
**models:** - CNN models (prototype)  
**ml\_motif:** - Real-time, Image/CV  
**type:** Model  
**ml\_task:** - Image Classification  
**solutions:** 0  
**notes:** In-progress; model design under development.  
**contact.name:** Shuyu Qin  
**contact.email:** shq219@lehigh.edu  
**results.links.name:** ChatGPT LLM  
**fair.reproducible:** in progress  
**fair.benchmark\_ready:** False  
**id:** d-stem  
**Citations:** [46]

#### Ratings:

Rating	Value	Reason
dataset	2	No dataset links or FAIR metadata; unclear public access
documentation	3	Paper is clear, but lacks a GitHub repo or full reproducibility pipeline
metrics	4	Only localization accuracy and inference time mentioned; not formally benchmarked with scripts
reference_solution	3	BraggNN model is described and evaluated, but no direct implementation or inference scripts available
software	2	No standalone code repository or setup instructions provided
specification	5	None



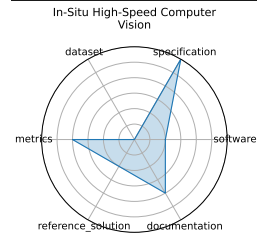
### 3.43 In-Situ High-Speed Computer Vision

Applies low-latency CNN models for image classification of plasma diagnostics streams; supports deployment on embedded platforms.

**date:** 2023-12-05  
**version:** v1.0  
**last\_updated:** 2023-12  
**expired:** unknown  
**valid:** yes  
**valid\_date:** 2023-12-05  
**url:** <https://arxiv.org/abs/2312.00128>  
**doi:** 10.48550/arXiv.2312.00128  
**domain:** Fusion/Plasma  
**focus:** Real-time image classification for in-situ plasma diagnostics  
**keywords:** - plasma - in-situ vision - real-time ML  
**licensing:** Via Fermilab  
**task\_types:** - Image Classification  
**ai\_capability\_measured:** - Real-time diagnostic inference  
**metrics:** - Accuracy - FPS  
**models:** - CNN  
**ml\_motif:** - Real-time, Image/CV  
**type:** Model  
**ml\_task:** - Image Classification  
**solutions:** Solution details are described in the referenced paper or repository.  
**notes:** Embedded/deployment details in progress.  
**contact.name:** unknown  
**contact.email:** unknown  
**results.links.name:** ChatGPT LLM  
**results.links.url:** [https://docs.google.com/document/d/1EqkRHuQs1yQqMvZs\\_L6p9JAy2vKX5OC'TubztFBuRoQ/edit?usp=sha](https://docs.google.com/document/d/1EqkRHuQs1yQqMvZs_L6p9JAy2vKX5OC'TubztFBuRoQ/edit?usp=sha)  
**fair.reproducible:** in progress  
**fair.benchmark\_ready:** False  
**id:** in-situ\_high-speed\_computer\_vision  
**Citations:** [47]

#### Ratings:

Rating	Value	Reason
dataset	0	Dataset not provided or described in any formal way
documentation	2	Some insight via papers, but no working repo, setup, or replication path
metrics	2	Throughput and accuracy mentioned, but not defined or benchmarked
reference_solution	1	Prototype CNNs described; no code, baseline, or training details available
software	1	No public implementation or containerized setup released
specification	3	No standardized I/O, latency constraint, or complete framing



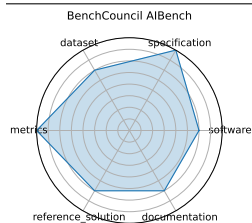
### 3.44 BenchCouncil AIBench

AIBench is a comprehensive benchmark suite that evaluates AI workloads at different levels (micro, component, application) across hardware systems-covering image generation, object detection, translation, recommendation, video prediction, etc.

**date:** 2020-01-01  
**version:** v1.0  
**last\_updated:** 2020-01  
**expired:** unknown  
**valid:** yes  
**valid\_date:** 2020-01-01  
**url:** <https://www.benchcouncil.org/AIBench/>  
**doi:** 10.48550/arXiv.1908.08998  
**domain:** General  
**focus:** End-to-end AI benchmarking across micro, component, and application levels  
**keywords:** - benchmarking - AI systems - application-level evaluation  
**licensing:** Apache License 2.0  
**task\_types:** - Training - Inference - End-to-end AI workloads  
**ai\_capability\_measured:** - System-level AI workload performance  
**metrics:** - Throughput - Latency - Accuracy  
**models:** - ResNet - BERT - GANs - Recommendation systems  
**ml\_motif:** - General  
**type:** Benchmark  
**ml\_task:** - NA  
**solutions:** Solution details are described in the referenced paper or repository.  
**notes:** Covers scenario-distilling, micro, component, and end-to-end benchmarks.  
**contact.name:** Wanling Gao (BenchCouncil)  
**contact.email:** unknown  
**results.links.name:** ChatGPT LLM  
**fair.reproducible:** Yes  
**fair.benchmark\_ready:** Yes  
**id:** benchcouncil\_aibench  
**Citations:** [48]

#### Ratings:

Rating	Value	Reason
dataset	3	Multiple datasets are mentioned, but not consistently FAIR-documented, versioned, or linked
documentation	3	Paper is comprehensive, but minimal user-facing documentation or structured reproduction guide
metrics	4	Metrics are appropriate, but standardization and reproducibility across tasks vary
reference_solution	3	Reference models (e.g., ResNet, BERT) described; no turnkey implementation or results repository for all levels
software	3	No containerized or automated implementation provided for full benchmark suite
specification	4	Task coverage is broad and well-scoped, but system constraints and expected outputs are not uniformly defined



### 3.45 BenchCouncil BigDataBench

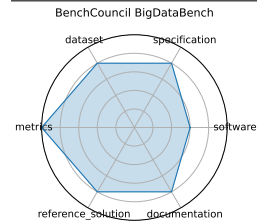
BigDataBench provides benchmarks for evaluating big data and AI workloads with realistic datasets (13 sources) and pipelines across analytics, graph, warehouse, NoSQL, streaming, and AI.

**date:** 2020-01-01  
**version:** v1.0  
**last\_updated:** 2020-01  
**expired:** unknown  
**valid:** yes  
**valid\_date:** 2020-01-01  
**url:** <https://www.benchcouncil.org/BigDataBench/>  
**doi:** 10.48550/arXiv.1802.08254  
**domain:** General  
**focus:** Big data and AI benchmarking across structured, semi-structured, and unstructured data workloads

**keywords:** - big data - AI benchmarking - data analytics  
**licensing:** Apache License 2.0  
**task\_types:** - Data preprocessing - Inference - End-to-end data pipelines  
**ai\_capability\_measured:** - Data processing and AI model inference performance at scale  
**metrics:** - Data throughput - Latency - Accuracy  
**models:** - CNN - LSTM - SVM - XGBoost  
**ml\_motif:** - General  
**type:** Benchmark  
**ml\_task:** - NA  
**solutions:** Solution details are described in the referenced paper or repository.  
**notes:** Built on eight data motifs; provides Hadoop, Spark, Flink, MPI implementations.  
**contact.name:** Jianfeng Zhan (BenchCouncil)  
**contact.email:** unknown  
**results.links.name:** ChatGPT LLM  
**results.links.url:** <https://docs.google.com/document/d/1VFRxhR2G5A83S8PqKBrP99LLVgcCGvX2WW4vTtwxmQ4/edit?usp=s>  
**fair.reproducible:** Yes  
**fair.benchmark\_ready:** Yes  
**id:** benchcouncil\_bigdatabench  
**Citations:** [49]

#### Ratings:

Rating	Value	Reason
dataset	4	Some datasets lack consistent versioning or rich metadata annotations.
documentation	4	Setup requires manual steps; some task-specific instructions lack clarity.
metrics	5	None
reference_solution	4	Not all benchmark components have fully reproducible baselines; deployment across platforms is fragmented.
software	3	No automated setup across all tasks; some components require manual integration.
specification	4	Specific I/O formats and hardware constraints are not uniformly detailed across all tasks.



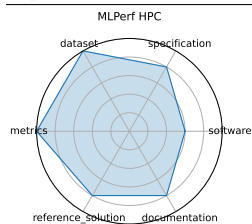
### 3.46 MLPerf HPC

MLPerf HPC introduces scientific model benchmarks (e.g., CosmoFlow, DeepCAM) aimed at large-scale HPC evaluation with >10x performance scaling through system-level optimizations.

**date:** 2021-10-20  
**version:** v1.0  
**last\_updated:** 2021-10  
**expired:** unknown  
**valid:** yes  
**valid\_date:** 2021-10-20  
**url:** <https://github.com/mlcommons/hpc>  
**doi:** 10.48550/arXiv.2110.11466  
**domain:** Cosmology, Climate, Protein Structure, Catalysis  
**focus:** Scientific ML training and inference on HPC systems  
**keywords:** - HPC - training - inference - scientific ML  
**licensing:** Apache License 2.0  
**task\_types:** - Training - Inference  
**ai\_capability\_measured:** - Scaling efficiency - training time - model accuracy on HPC  
**metrics:** - Training time - Accuracy - GPU utilization  
**models:** - CosmoFlow - DeepCAM - OpenCatalyst  
**ml\_motif:** - HPC/inference, HPC/training  
**type:** Framework  
**ml\_task:** - NA  
**solutions:** Solution details are described in the referenced paper or repository.  
**notes:** Shared framework with MLCommons Science; reference implementations included.  
**contact.name:** Steven Farrell (MLCommons)  
**contact.email:** unknown  
**results.links.name:** ChatGPT LLM  
**fair.reproducible:** Yes  
**fair.benchmark\_ready:** Yes  
**id:** mlperf\_hpc  
**Citations:** [50]

#### Ratings:

Rating	Value	Reason
dataset	5	Not all data is independently versioned or comes with standardized FAIR metadata.
documentation	4	Central guidance is available but requires domain-specific effort to replicate results across systems.
metrics	5	None
reference_solution	4	Reproducibility and environment tuning depend on system configuration; baseline models not uniformly bundled.
software	3	Reference implementations exist but containerization and environment setup require manual effort across HPC systems.
specification	4	Hardware constraints and I/O formats are not fully defined for all scenarios.



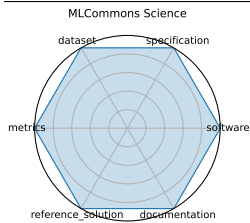
### 3.47 MLCommons Science

MLCommons Science assembles benchmark tasks with datasets, targets, and implementations across earthquake forecasting, satellite imagery, drug screening, electron microscopy, and CFD to drive scientific ML reproducibility.

**date:** 2023-06-01  
**version:** v1.0  
**last\_updated:** 2023-06  
**expired:** unknown  
**valid:** yes  
**valid\_date:** 2023-06-01  
**url:** <https://github.com/mlcommons/science>  
**doi:** unknown  
**domain:** Earthquake, Satellite Image, Drug Discovery, Electron Microscope, CFD  
**focus:** AI benchmarks for scientific applications including time-series, imaging, and simulation  
**keywords:** - science AI - benchmark - MLCommons - HPC  
**licensing:** Apache License 2.0  
**task\_types:** - Time-series analysis - Image classification - Simulation surrogate modeling  
**ai\_capability\_measured:** - Inference accuracy - simulation speed-up - generalization  
**metrics:** - MAE - Accuracy - Speedup vs simulation  
**models:** - CNN - GNN - Transformer  
**ml\_motif:** - Time-series, Image/CV, HPC/inference  
**type:** Framework  
**ml\_task:** - NA  
**solutions:** 0  
**notes:** Joint national-lab effort under Apache-2.0 license.  
**contact.name:** MLCommons Science Working Group  
**contact.email:** unknown  
**results.links.name:** ChatGPT LLM  
**fair.reproducible:** Yes  
**fair.benchmark\_ready:** Yes  
**id:** mlcommons\_science  
**Citations:** [51]

#### Ratings:

Rating	Value	Reason
dataset	5	Public scientific datasets are used with defined splits. At least 4 FAIR principles are followed.
documentation	5	Thorough documentation exists covering the task, background, motivation, evaluation criteria, and includes a supporting paper.
metrics	5	Clearly defined metrics such as accuracy, training time, and GPU utilization are used. These metrics are explained and effectively capture solution performance.
reference_solution	5	A reference implementation is available, well-documented, trainable/open, and includes full metric evaluation and software/hardware details.
software	5	Actively maintained GitHub repository available at <a href="https://github.com/mlcommons/science">https://github.com/mlcommons/science</a> with implementations, scripts, and reproducibility support.
specification	5	All five specification aspects are covered: system constraints, task, dataset format, benchmark inputs, and outputs.



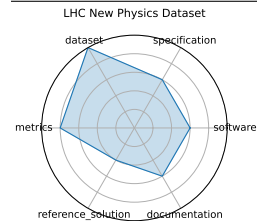
### 3.48 LHC New Physics Dataset

A dataset of proton-proton collision events emulating a 40 MHz real-time data stream from LHC detectors, pre-filtered on electron or muon presence. Designed for unsupervised new-physics detection algorithms under latency/bandwidth constraints.

<b>date:</b>	2021-07-05
<b>version:</b>	v1.0
<b>last_updated:</b>	2021-07
<b>expired:</b>	unknown
<b>valid:</b>	yes
<b>valid_date:</b>	2021-07-05
<b>url:</b>	<a href="https://arxiv.org/pdf/2107.02157">https://arxiv.org/pdf/2107.02157</a>
<b>doi:</b>	unknown
<b>domain:</b>	Particle Physics; Real-time Triggering
<b>focus:</b>	Real-time LHC event filtering for anomaly detection using proton collision data
<b>keywords:</b>	- anomaly detection - proton collision - real-time inference - event filtering - unsupervised ML
<b>licensing:</b>	unknown
<b>task_types:</b>	- Anomaly detection - Event classification
<b>ai_capability_measured:</b>	- Unsupervised signal detection under latency and bandwidth constraints
<b>metrics:</b>	- ROC-AUC - Detection efficiency
<b>models:</b>	- Autoencoder - Variational autoencoder - Isolation forest
<b>ml_motif:</b>	- Multiple
<b>type:</b>	Framework
<b>ml_task:</b>	- NA
<b>solutions:</b>	0
<b>notes:</b>	Includes electron/muon-filtered background and black-box signal benchmarks; 1M events per black box.
<b>contact.name:</b>	Ema Puljak (ema.puljak@cern.ch)
<b>contact.email:</b>	unknown
<b>datasets.links.name:</b>	Zenodo stores, background + 3 black-box signal sets. 1M events each
<b>results.links.name:</b>	ChatGPT LLM
<b>fair.reproducible:</b>	Yes
<b>fair.benchmark_ready:</b>	Yes
<b>id:</b>	lhc_new_physics_dataset
<b>Citations:</b>	[52]

#### Ratings:

Rating	Value	Reason
dataset	5	Large-scale dataset hosted on Zenodo, publicly available, well-documented, with defined train/test structure. Appears to follow at least 4 FAIR principles.
documentation	3	Some description in papers and dataset metadata exists, but lacks a unified guide, README, or training setup in a central location.
metrics	4	Uses reasonable metrics (ROC-AUC, detection efficiency) that capture performance but lacks full explanation and standard evaluation tools.
reference_solution	2	Baselines are described across multiple papers but lack centralized, reproducible implementations and hardware/software setup details.
software	3	While not formally evaluated in the previous version, Zenodo and paper links suggest available code for baseline models (e.g., autoencoders, GANs), though they are scattered and not unified in a single repository.
specification	3	The task and context are clearly described, but system constraints and formal inputs/outputs are not fully specified.





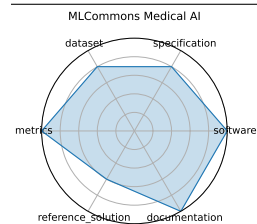
### 3.49 MLCommons Medical AI

The MLCommons Medical AI working group develops benchmarks, best practices, and platforms (MedPerf, GaNDLF, COFE) to accelerate robust, privacy-preserving AI development for healthcare. MedPerf enables federated testing of clinical models on diverse datasets, improving generalizability and equity while keeping data onsite .

<b>date:</b>	2023-07-17
<b>version:</b>	v1.0
<b>last_updated:</b>	2023-07
<b>expired:</b>	unknown
<b>valid:</b>	yes
<b>valid_date:</b>	2023-07-17
<b>url:</b>	<a href="https://github.com/mlcommons/medical">https://github.com/mlcommons/medical</a>
<b>doi:</b>	unknown
<b>domain:</b>	Healthcare; Medical AI
<b>focus:</b>	Federated benchmarking and evaluation of medical AI models across diverse real-world clinical data
<b>keywords:</b>	- medical AI - federated evaluation - privacy-preserving - fairness - healthcare benchmarks
<b>licensing:</b>	Apache License 2.0
<b>task_types:</b>	- Federated evaluation - Model validation
<b>ai_capability_measured:</b>	- Clinical accuracy - fairness - generalizability - privacy compliance
<b>metrics:</b>	- ROC AUC - Accuracy - Fairness metrics
<b>models:</b>	- MedPerf-validated CNNs - GaNDLF workflows
<b>ml_motif:</b>	- Multiple
<b>type:</b>	Platform
<b>ml_task:</b>	- NA
<b>solutions:</b>	0
<b>notes:</b>	Open-source platform under Apache-2.0; used across 20+ institutions and hospitals .
<b>contact.name:</b>	Alex Karargyris (MLCommons Medical AI)
<b>contact.email:</b>	unknown
<b>datasets.links.name:</b>	Multi-institutional clinical datasets, radiology
<b>results.links.name:</b>	ChatGPT LLM
<b>fair.reproducible:</b>	Yes
<b>fair.benchmark_ready:</b>	Yes
<b>id:</b>	mlcommons_medical_ai
<b>Citations:</b>	[53]

#### Ratings:

Rating	Value	Reason
dataset	4	Multi-institutional datasets used in federated settings; real-world data is handled privately onsite, but some FAIR aspects (e.g., accessibility and metadata) are implicit.
documentation	5	Extensive documentation, papers, and community support exist. Clear examples and usage instructions are provided in GitHub and publications.
metrics	5	Metrics such as ROC AUC, accuracy, and fairness are clearly specified and directly support goals like generalizability and equity.
reference_solution	3	GaNDLF workflows and MedPerf-validated CNNs are referenced, but not all baseline models are centrally documented or easily reproducible.
software	5	GitHub repository ( <a href="https://github.com/mlcommons/medical">https://github.com/mlcommons/medical</a> ) provides actively maintained open-source tools like MedPerf and GaNDLF for federated medical AI evaluation.
specification	4	The platform defines federated tasks and model evaluation scenarios. Some clinical and system-level constraints are implied but not uniformly formalized across all use cases.



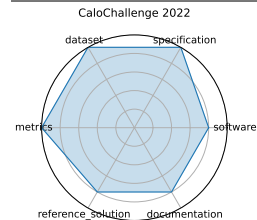
### 3.50 CaloChallenge 2022

The Fast Calorimeter Simulation Challenge 2022 assessed 31 generative-model submissions (VAEs, GANs, Flows, Diffusion) on four calorimeter shower datasets; benchmarking shower quality, generation speed, and model complexity .

**date:** 2024-10-28  
**version:** v1.0  
**last\_updated:** 2024-10  
**expired:** unknown  
**valid:** yes  
**valid\_date:** 2024-10-28  
**url:** <http://arxiv.org/abs/2410.21611>  
**doi:** 10.48550/arXiv.2410.21611  
**domain:** LHC Calorimeter; Particle Physics  
**focus:** Fast generative-model-based calorimeter shower simulation evaluation  
**keywords:** - calorimeter simulation - generative models - surrogate modeling - LHC - fast simulation  
**licensing:** Via Fermilab  
**task\_types:** - Surrogate modeling  
**ai\_capability\_measured:** - Simulation fidelity - speed - efficiency  
**metrics:** - Histogram similarity - Classifier AUC - Generation latency  
**models:** - VAE variants - GAN variants - Normalizing flows - Diffusion models  
**ml\_motif:** - Surrogate  
**type:** Dataset  
**ml\_task:** - Surrogate Modeling  
**solutions:** Solution details are described in the referenced paper or repository.  
**notes:** The most comprehensive survey to date on ML-based calorimeter simulation; 31 submissions over different dataset sizes.  
**contact.name:** Claudius Krause (CaloChallenge Lead)  
**contact.email:** unknown  
**datasets.links.name:** Four LHC calorimeter shower datasets  
**datasets.links.url:** various voxel resolutions  
**results.links.name:** ChatGPT LLM  
**fair.reproducible:** Yes  
**fair.benchmark\_ready:** Yes  
**id:** calochallenge\_  
**Citations:** [54]

#### Ratings:

Rating	Value	Reason
dataset	5	Four well-structured calorimeter datasets are provided, with different voxel resolutions, open access, signal/background separation, and metadata. FAIR principles are well covered.
documentation	4	Accompanied by a detailed paper and dataset description. Reproduction of pipelines may require additional setup or familiarity with the model submissions.
metrics	5	Metrics like histogram similarity, classifier AUC, and generation latency are well defined and relevant for simulation quality, fidelity, and performance.
reference_solution	4	Several baselines (GANs, VAEs, flows, diffusion models) are documented and evaluated. Some are available via community repos, though not all are fully standardized or bundled.
software	4	Community GitHub repos and model implementations are available for the 31 submissions. While not fully unified in one place, the software is accessible and reproducible.
specification	5	The task—evaluating fast generative calorimeter simulations—is clearly defined with benchmarking protocols, constraints like latency and model complexity, and structured evaluation criteria.



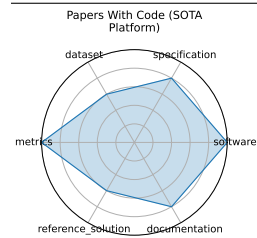
### 3.51 Papers With Code (SOTA Platform)

Papers With Code (PWC) aggregates benchmark suites, tasks, and code across ML research: 12,423 benchmarks, 5,358 unique tasks, and 154,766 papers with code links. It tracks SOTA metrics and fosters reproducibility.

**date:** ongoing  
**version:** v1.0  
**last\_updated:** 2025-06  
**expired:** unknown  
**valid:** yes  
**valid\_date:** ongoing  
**url:** <https://paperswithcode.com/sota>  
**doi:** unknown  
**domain:** General ML; All domains  
**focus:** Open platform tracking state-of-the-art results, benchmarks, and implementations across ML tasks and papers  
**keywords:** - leaderboard - benchmarking - reproducibility - open-source  
**licensing:** Apache License 2.0  
**task\_types:** - Multiple (Classification, Detection, NLP, etc.)  
**ai\_capability\_measured:** - Model performance across tasks (accuracy - F1 - BLEU - etc.)  
**metrics:** - Task-specific (Accuracy, F1, BLEU, etc.)  
**models:** - All published models with code  
**ml\_motif:** - Multiple  
**type:** Platform  
**ml\_task:** - Multiple  
**solutions:** 0  
**notes:** Community-driven open platform; automatic data extraction and versioning.  
**contact.name:** Papers With Code Team  
**contact.email:** unknown  
**results.links.name:** ChatGPT LLM  
**fair.reproducible:** Yes  
**fair.benchmark\_ready:** Yes  
**id:** papers\_with\_code\_sota\_platform  
**Citations:** [55]

#### Ratings:

Rating	Value	Reason
dataset	3	Relies on external datasets submitted by the community. While links are available, FAIR compliance is not guaranteed or systematically enforced across all benchmarks.
documentation	4	Strong front-end documentation and metadata on benchmarks, tasks, and models; however, some benchmark-specific instructions are sparse or dependent on external paper links.
metrics	5	Tracks state-of-the-art using task-specific metrics like Accuracy, F1, BLEU, etc., with consistent aggregation and historical SOTA tracking.
reference_solution	3	Provides links to implementations of many SOTA models, but no single unified reference baseline is required or maintained per benchmark.
software	5	Actively maintained open-source platform ( <a href="https://paperswithcode.com">https://paperswithcode.com</a> ) under Apache 2.0 license; includes automatic integration with GitHub, datasets, and models for reproducibility.
specification	4	Task and benchmark structures are well organized and standardized, but due to its broad coverage, input/output formats vary significantly between tasks and are not always tightly controlled.



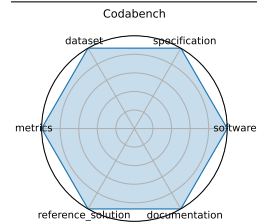
### 3.52 Codabench

Codabench (successor to CodaLab) is a flexible, easy-to-use, reproducible API platform for hosting AI benchmarks and code-submission challenges. It supports custom scoring, inverted benchmarks, and scalable public or private queues .

**date:** 2022-01-01  
**version:** v1.0  
**last\_updated:** 2025-03  
**expired:** unknown  
**valid:** yes  
**valid\_date:** 2022-01-01  
**url:** <https://www.codabench.org/>  
**doi:** <https://doi.org/10.1016/j.patter.2022.100543>  
**domain:** General ML; Multiple  
**focus:** Open-source platform for organizing reproducible AI benchmarks and competitions  
**keywords:** - benchmark platform - code submission - competitions - meta-benchmark  
**licensing:** <https://github.com/codalab/codalab-competitions/wiki/Privacy>  
**task\_types:** - Multiple  
**ai\_capability\_measured:** - Model reproducibility - performance across datasets  
**metrics:** - Submission count - Leaderboard ranking - Task-specific metrics  
**models:** - Arbitrary code submissions  
**ml\_motif:** - Multiple  
**type:** Platform  
**ml\_task:** - Multiple  
**solutions:** Several  
**notes:** Hosts 51 public competitions, ~26 k users, 177 k submissions  
**contact.name:** Isabelle Guyon (Université Paris-Saclay)  
**contact.email:** unknown  
**results.links.name:** ChatGPT LLM  
**fair.reproducible:** Yes  
**fair.benchmark\_ready:** Yes  
**id:** codabench  
**Citations:** [56]

#### Ratings:

Rating	Value	Reason
dataset	1	This is a platform for posting benchmarks, not a benchmark in itself.
documentation	1	This is a platform for posting benchmarks, not a benchmark in itself.
metrics	1	This is a platform for posting benchmarks, not a benchmark in itself.
reference_solution	1	This is a platform for posting benchmarks, not a benchmark in itself.
software	1	This is a platform for posting benchmarks, not a benchmark in itself.
specification	1	This is a platform for posting benchmarks, not a benchmark in itself.



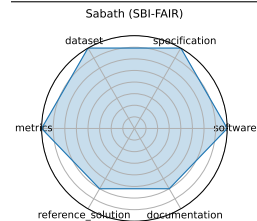
### 3.53 Sabath (SBI-FAIR)

Sabath is a metadata framework from the SBI-FAIR group (UTK, Argonne, Virginia) facilitating FAIR-compliant benchmarking and surrogate execution logging across HPC systems .

**date:** 2021-09-27  
**version:** v1.0  
**last\_updated:** 2023-07  
**expired:** unknown  
**valid:** yes  
**valid\_date:** 2021-09-27  
**url:** <https://sbi-fair.github.io/docs/software/sabath/>  
**doi:** unknown  
**domain:** Systems; Metadata  
**focus:** FAIR metadata framework for ML-driven surrogate workflows in HPC systems  
**keywords:** - meta-benchmark - metadata - HPC - surrogate modeling  
**licensing:** BSD 3-Clause License  
**task\_types:** - Systems benchmarking  
**ai\_capability\_measured:** - Metadata tracking - reproducible HPC workflows  
**metrics:** - Metadata completeness - FAIR compliance  
**models:** - NA  
**ml\_motif:** - Systems  
**type:** Platform  
**ml\_task:** - NA  
**solutions:** 0  
**notes:** Developed by PI Piotr Luszczek at UTK; integrates with MiniWeatherML, AutoPhaseNN, Cosmoflow, etc.  
**contact.name:** Piotr Luszczek  
**contact.email:** luszczek@utk.edu  
**results.links.name:** ChatGPT LLM  
**fair.reproducible:** Yes  
**fair.benchmark\_ready:** N/A  
**id:** sabath\_sbi-fair  
**Citations:** [57]

#### Ratings:

Rating	Value	Reason
dataset	4	Datasets used in surrogate benchmarks are publicly available, well-structured, and FAIR-aligned, but not independently hosted by Sabath itself.
documentation	3	Basic instructions and code are provided on GitHub, but more detailed walkthroughs, use-case examples, or tutorials are limited.
metrics	4	Emphasizes metadata completeness and FAIR compliance. Metrics are clear and well-matched to its metadata-focused benchmarking context.
reference_solution	3	Includes integration with multiple surrogate benchmarks and models, though not all are fully documented or packaged as standardized reference solutions.
software	4	Actively maintained GitHub repository ( <a href="https://github.com/icl-utk-edu/slip/tree/sabath">https://github.com/icl-utk-edu/slip/tree/sabath</a> ) with BSD-licensed tooling for FAIR metadata capture; integrates with existing surrogate modeling benchmarks.
specification	4	FAIR metadata structure and logging goals are clearly described. Input/output definitions are implied through integrations (e.g., MiniWeatherML), though not always formalized.



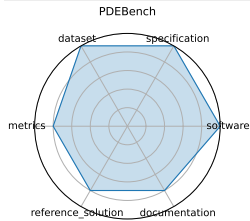
### 3.54 PDEBench

PDEBench offers forward/inverse PDE tasks with large ready-to-use datasets and baselines (FNO, U-Net, PINN), packaged via a unified API. It won the SimTech Best Paper Award 2023 .

<b>date:</b>	2022-10-13
<b>version:</b>	v0.1.0
<b>last_updated:</b>	2025-05
<b>expired:</b>	unknown
<b>valid:</b>	yes
<b>valid_date:</b>	2022-10-13
<b>url:</b>	<a href="https://github.com/pdebench/PDEBench">https://github.com/pdebench/PDEBench</a>
<b>doi:</b>	10.48550/arXiv.2210.07182
<b>domain:</b>	CFD; Weather Modeling
<b>focus:</b>	Benchmark suite for ML-based surrogates solving time-dependent PDEs
<b>keywords:</b>	- PDEs - CFD - scientific ML - surrogate modeling - NeurIPS
<b>licensing:</b>	Other
<b>task_types:</b>	- Supervised Learning
<b>ai_capability_measured:</b>	- Time-dependent PDE modeling; physical accuracy
<b>metrics:</b>	- RMSE - boundary RMSE - Fourier RMSE
<b>models:</b>	- FNO - U-Net - PINN - Gradient-Based inverse methods
<b>ml_motif:</b>	- Multiple
<b>type:</b>	Framework
<b>ml_task:</b>	- Supervised Learning
<b>solutions:</b>	Solution details are described in the referenced paper or repository.
<b>notes:</b>	Datasets hosted on DaRUS (DOI:10.18419/darus-2986); contact maintainers by email
<b>contact.name:</b>	Makoto Takamoto (makoto.takamoto@neclab.eu)
<b>contact.email:</b>	unknown
<b>results.links.name:</b>	ChatGPT LLM
<b>fair.reproducible:</b>	Yes
<b>fair.benchmark_ready:</b>	Yes
<b>id:</b>	pdebench
<b>Citations:</b>	[58]

#### Ratings:

Rating	Value	Reason
dataset	5	Diverse PDE datasets (synthetic and real-world) hosted on DaRUS with DOIs. Datasets are well-documented, structured, and follow FAIR practices.
documentation	4	Strong documentation on GitHub including examples, configs, and usage instructions. Some model-specific details and tutorials could be further expanded.
metrics	4	Includes RMSE, boundary RMSE, and Fourier-domain RMSE. These are well-suited to PDE problems, though rationale behind metric choices could be expanded in some cases.
reference_solution	4	Baselines (FNO, U-Net, PINN, etc.) are available and documented, but not every model includes full training and evaluation reproducibility out-of-the-box.
software	5	GitHub repository ( <a href="https://github.com/pdebench/PDEBench">https://github.com/pdebench/PDEBench</a> ) is actively maintained and includes training pipelines, data loaders, and evaluation scripts. Installation and usage are well-documented.
specification	5	Clearly defined tasks for forward and inverse PDE problems, with structured input/output formats, system constraints, and task specifications.



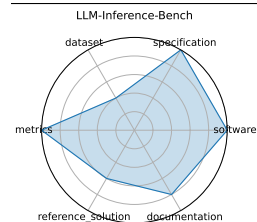
### 3.55 LLM-Inference-Bench

A suite evaluating inference performance of LLMs (LLaMA, Mistral, Qwen) across diverse accelerators (NVIDIA, AMD, Intel, SambaNova) and frameworks (vLLM, DeepSpeed-MII, etc.), with an interactive dashboard and per-platform metrics.

**date:** 2024-10-31  
**version:** v1.0  
**last\_updated:** 2024-11  
**expired:** unknown  
**valid:** yes  
**valid\_date:** 2024-10-31  
**url:** <https://github.com/argonne-lcf/LLM-Inference-Bench>  
**doi:** unknown  
**domain:** LLM; HPC/inference  
**focus:** Hardware performance benchmarking of LLMs on AI accelerators  
**keywords:** - LLM - inference benchmarking - GPU - accelerator - throughput  
**licensing:** BSD 3-Clause New or Revised License  
**task\_types:** - Inference Benchmarking  
**ai\_capability\_measured:** - Inference throughput - latency - hardware utilization  
**metrics:** - Token throughput (tok/s) - Latency - Framework-hardware mix performance  
**models:** - LLaMA-2-7B - LLaMA-2-70B - Mistral-7B - Qwen-7B  
**ml\_motif:** - HPC/inference  
**type:** Dataset  
**ml\_task:** - Inference Benchmarking  
**solutions:** 0  
**notes:** Licensed under BSD-3, maintained by Argonne; supports GPUs and accelerators.  
**contact.name:** Krishna Teja Chitty-Venkata (Argonne LCF)  
**contact.email:** unknown  
**results.links.name:** ChatGPT LLM  
**fair.reproducible:** Yes  
**fair.benchmark\_ready:** Yes  
**id:** llm-inference-bench  
**Citations:** [59]

#### Ratings:

Rating	Value	Reason
dataset	2	No novel dataset is introduced; benchmark relies on pre-trained LLMs and synthetic inference inputs. Dataset structure and FAIR considerations are minimal.
documentation	4	GitHub repo provides clear usage instructions, setup guides, and interactive dashboard tooling. Some areas like benchmarking extensions or advanced tuning are less detailed.
metrics	5	Hardware-specific metrics (token throughput, latency, utilization) are well-defined, consistently measured, and aggregated in dashboards.
reference_solution	3	Inference configurations and baseline performance results are provided, but there are no full reference training pipelines or model implementations.
software	5	Public GitHub repository under BSD-3 license. Includes scripts, configurations, and dashboards for running and visualizing LLM inference benchmarks across multiple accelerator platforms.
specification	5	Benchmark scope, models, accelerator targets, and supported frameworks are clearly specified. Input configurations and output metrics are standardized across hardware types.



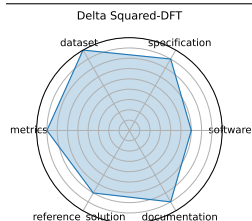
### 3.56 Delta Squared-DFT

Introduces the Delta Squared-ML paradigm-using ML corrections to DFT to predict reaction energies with accuracy comparable to CCSD(T), while training on small CC datasets. Evaluated across 10 reaction datasets covering organic and organometallic transformations.

<b>date:</b>	2024-12-13
<b>version:</b>	v1.0
<b>last_updated:</b>	2024-12
<b>expired:</b>	unknown
<b>valid:</b>	yes
<b>valid_date:</b>	2024-12-13
<b>url:</b>	<a href="https://neurips.cc/virtual/2024/poster/97788">https://neurips.cc/virtual/2024/poster/97788</a>
<b>doi:</b>	10.48550/arXiv.2406.14347
<b>domain:</b>	Computational Chemistry; Materials Science
<b>focus:</b>	Benchmarking machine-learning corrections to DFT using Delta Squared-trained models for reaction energies
<b>keywords:</b>	- density functional theory - Delta Squared-ML correction - reaction energetics - quantum chemistry
<b>licensing:</b>	unknown
<b>task_types:</b>	- Regression
<b>ai_capability_measured:</b>	- High-accuracy energy prediction - DFT correction
<b>metrics:</b>	- Mean Absolute Error (eV) - Energy ranking accuracy
<b>models:</b>	- Delta Squared-ML correction networks - Kernel ridge regression
<b>ml_motif:</b>	- Scientific ML
<b>type:</b>	Dataset + Benchmark
<b>ml_task:</b>	- Regression
<b>solutions:</b>	Solution details are described in the referenced paper or repository.
<b>notes:</b>	Demonstrates CC-level accuracy with ~1% of high-level data. Benchmarks publicly included for reproducibility.
<b>contact.name:</b>	Wei Liu
<b>contact.email:</b>	unknown
<b>results.links.name:</b>	ChatGPT LLM
<b>fair.reproducible:</b>	Yes
<b>fair.benchmark_ready:</b>	Yes
<b>id:</b>	delta_squared-dft
<b>Citations:</b>	[60]

#### Ratings:

Rating	Value	Reason
dataset	4.5	Multi-modal quantum chemistry datasets are standardized and accessible; repository available.
documentation	4	Source code supports pipeline reuse, but formal evaluation splits may vary.
metrics	4	Uses standard regression metrics like MAE and energy ranking accuracy; appropriate for task.
reference_solution	3.5	Includes baseline regression and kernel ridge models; implementations are reproducible.
software	3	Source code and baseline models available for ML correction to DFT; framework maturity is moderate.
specification	4	Benchmark focuses on reaction energy prediction with clear goals, though some task specifics could be formalized further.





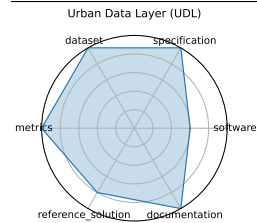
### 3.57 Urban Data Layer (UDL)

UrbanDataLayer standardizes heterogeneous urban data formats and provides pipelines for tasks like air quality prediction and land-use classification, enabling the rapid creation of multi-modal urban benchmarks .

**date:** 2024-12-13  
**version:** v1.0  
**last\_updated:** 2024-12  
**expired:** unknown  
**valid:** yes  
**valid\_date:** 2024-12-13  
**url:** <https://neurips.cc/virtual/2024/poster/97837>  
**doi:** unknown  
**domain:** Urban Computing; Data Engineering  
**focus:** Unified data pipeline for multi-modal urban science research  
**keywords:** - data pipeline - urban science - multi-modal - benchmark  
**licensing:** unknown  
**task\_types:** - Prediction - Classification  
**ai\_capability\_measured:** - Multi-modal urban inference - standardization  
**metrics:** - Task-specific accuracy or RMSE  
**models:** - Baseline regression/classification pipelines  
**ml\_motif:** - Data engineering  
**type:** Framework  
**ml\_task:** - Prediction, classification  
**solutions:** 0  
**notes:** Source code available on GitHub (SJTU-CILAB/udl); promotes reusable urban-science foundation models .  
**contact.name:** Yiheng Wang  
**contact.email:** unknown  
**results.links.name:** ChatGPT LLM  
**fair.reproducible:** Yes  
**fair.benchmark\_ready:** Yes  
**id:** urban\_data\_layer\_udl  
**Citations:** [61]

#### Ratings:

Rating	Value	Reason
dataset	5	Large, multi-modal urban datasets are open-source, well-documented, and support reproducible research.
documentation	5	GitHub repository and conference poster provide comprehensive code and reproducibility instructions.
metrics	5	Uses task-specific accuracy and RMSE metrics appropriate for prediction and classification.
reference_solution	4	Baseline models available but not exhaustive; community adoption and extensions expected.
software	3	Source code is publicly available on GitHub; baseline regression and classification pipelines are included but framework maturity is moderate.
specification	5	Multiple urban science tasks like prediction and classification are well specified with clear input/output and evaluation criteria.



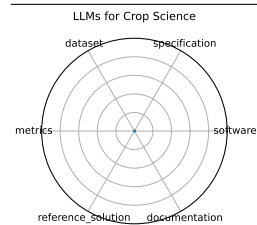
### 3.58 LLMs for Crop Science

Establishes a benchmark of 3,500 expert-annotated prompts and QA pairs covering crop traits, growth stages, and environmental interactions. Tests GPT-style LLMs on accuracy and domain reasoning using in-context, chain-of-thought, and retrieval-augmented prompts.

**date:** 2024-12-13  
**version:** v1.0  
**last\_updated:** 2024-12  
**expired:** unknown  
**valid:** yes  
**valid\_date:** 2024-12-13  
**url:** <https://neurips.cc/virtual/2024/poster/97570>  
**doi:** 10.48550/arXiv.2406.03085  
**domain:** Agricultural Science; NLP  
**focus:** Evaluating LLMs on crop trait QA and textual inference tasks with domain-specific prompts  
**keywords:** - crop science - prompt engineering - domain adaptation - question answering  
**licensing:** unknown  
**task\_types:** - Question Answering - Inference  
**ai\_capability\_measured:** - Scientific knowledge - crop reasoning  
**metrics:** - Accuracy - F1 score  
**models:** - GPT-4 - LLaMA-2-13B - T5-XXL  
**ml\_motif:** - NLP  
**type:** Dataset  
**ml\_task:** - QA, inference  
**solutions:** Solution details are described in the referenced paper or repository.  
**notes:** Includes examples with retrieval-augmented and chain-of-thought prompt templates; supports few-shot adaptation.  
**contact.name:** Deepak Patel  
**contact.email:** unknown  
**results.links.name:** ChatGPT LLM  
**fair.reproducible:** Yes  
**fair.benchmark\_ready:** Yes  
**id:** llms\_for\_crop\_science  
**Citations:** [62]

#### Ratings:

Rating	Value	Reason
dataset	0	This is a model, not a benchmark.
documentation	0	This is a model, not a benchmark.
metrics	0	This is a model, not a benchmark.
reference_solution	0	This is a model, not a benchmark.
software	0	This is a model, not a benchmark.
specification	0	This is a model, not a benchmark.



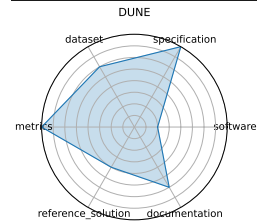
### 3.59 DUNE

Applying real-time ML methods to time-series data from DUNE detectors, exploring trigger-level anomaly detection and event selection with low latency constraints.

**date:** 2024-10-15  
**version:** v1.0  
**last\_updated:** 2024-10  
**expired:** unknown  
**valid:** yes  
**valid\_date:** 2024-10-15  
**url:** [https://indico.fnal.gov/event/66520/contributions/301423/attachments/182439/250508/fast\\_ml\\_dunedaq\\_sonic](https://indico.fnal.gov/event/66520/contributions/301423/attachments/182439/250508/fast_ml_dunedaq_sonic)  
**doi:** 10.48550/arXiv.2103.13910  
**domain:** Particle Physics  
**focus:** Real-time ML for DUNE DAQ time-series data  
**keywords:** - DUNE - time-series - real-time - trigger  
**licensing:** Via Fermilab  
**task\_types:** - Trigger selection - Time-series anomaly detection  
**ai\_capability\_measured:** - Low-latency event detection  
**metrics:** - Detection efficiency - Latency  
**models:** - CNN - LSTM (planned)  
**ml\_motif:** - Real-time, Time-series  
**type:** Benchmark (in progress)  
**ml\_task:** - Supervised Learning  
**solutions:** Solution details are described in the referenced paper or repository.  
**notes:** Prototype models demonstrated on SONIC platform  
**contact.name:** Andrew J. Morgan  
**contact.email:** unknown  
**datasets.links.name:** DUNE SONIC data  
**results.links.name:** ChatGPT LLM  
**fair.reproducible:** in progress  
**fair.benchmark\_ready:** False  
**id:** dune  
**Citations:** [63]

#### Ratings:

Rating	Value	Reason
dataset	3	Dataset lacks a public URL; FAIR metadata and versioning are missing
documentation	3	Documentation exists only in slides/GDocs; no implementation guide or structured release
metrics	4	Metrics are relevant but no benchmark baseline or detailed evaluation guidance is provided
reference_solution	2	Autoencoder prototype exists but is not reproducible; RL model still in development
software	1	Code not available; no containerization or setup provided
specification	4	Constraints like latency thresholds are described qualitatively but not numerically defined



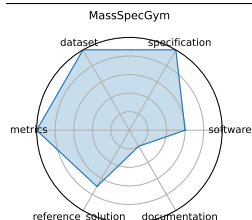
### 3.60 MassSpecGym

MassSpecGym curates the largest public MS/MS dataset with three standardized tasks-de novo structure generation, molecule retrieval, and spectrum simulation-using challenging generalization splits to propel ML-driven molecule discovery.

<b>date:</b>	2024-12-13
<b>version:</b>	v1.0
<b>last_updated:</b>	2024-12
<b>expired:</b>	unknown
<b>valid:</b>	yes
<b>valid_date:</b>	2024-12-13
<b>url:</b>	<a href="https://neurips.cc/virtual/2024/poster/97823">https://neurips.cc/virtual/2024/poster/97823</a>
<b>doi:</b>	unknown
<b>domain:</b>	Cheminformatics; Molecular Discovery
<b>focus:</b>	Benchmark suite for discovery and identification of molecules via MS/MS
<b>keywords:</b>	- mass spectrometry - molecular structure - de novo generation - retrieval - dataset
<b>licensing:</b>	unknown
<b>task_types:</b>	- De novo generation - Retrieval - Simulation
<b>ai_capability_measured:</b>	- Molecular identification and generation from spectral data
<b>metrics:</b>	- Structure accuracy - Retrieval precision - Simulation MSE
<b>models:</b>	- Graph-based generative models - Retrieval baselines
<b>ml_motif:</b>	- Benchmark
<b>type:</b>	Dataset, Benchmark
<b>ml_task:</b>	- Generation, retrieval, simulation
<b>solutions:</b>	0
<b>notes:</b>	Dataset~>1M spectra; open-source GitHub repo; widely cited as a go-to benchmark for MS/MS tasks.
<b>contact.name:</b>	Roman Bushuiev
<b>contact.email:</b>	unknown
<b>results.links.name:</b>	ChatGPT LLM
<b>fair.reproducible:</b>	Yes
<b>fair.benchmark_ready:</b>	Yes
<b>id:</b>	massspecgym
<b>Citations:</b>	[64]

#### Ratings:

Rating	Value	Reason
dataset	5	Largest public MS/MS dataset with extensive annotations; minor point deducted for lack of explicit train/validation/test splits.
documentation	1	Paper and poster describe benchmark goals and design, but documentation and user guides are minimal and repo status uncertain.
metrics	5	Well-defined metrics such as structure accuracy, retrieval precision, and simulation MSE used consistently.
reference_solution	3.5	CNN-based baselines are referenced, but pretrained weights and comprehensive training pipelines are not fully documented.
software	3	Open-source GitHub repository available; baseline models and training code partially provided but overall framework maturity is moderate.
specification	5	Clearly defined tasks including molecule generation, retrieval, and spectrum simulation, scoped for MS/MS molecular identification.



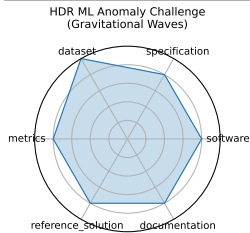
### 3.61 HDR ML Anomaly Challenge (Gravitational Waves)

A benchmark for detecting anomalous transient gravitational-wave signals, including "unknown-unknowns," using preprocessed LIGO time-series at 4096 Hz. Competitors submit inference models on Codabench for continuous 50 ms segments from dual interferometers.

<b>date:</b>	2025-03-03
<b>version:</b>	v1.0
<b>last_updated:</b>	2025-03
<b>expired:</b>	unknown
<b>valid:</b>	yes
<b>valid_date:</b>	2025-03-03
<b>url:</b>	<a href="https://www.codabench.org/competitions/2626/">https://www.codabench.org/competitions/2626/</a>
<b>doi:</b>	10.48550/arXiv.2503.02112
<b>domain:</b>	Astrophysics; Time-series
<b>focus:</b>	Detecting anomalous gravitational-wave signals from LIGO/Virgo datasets
<b>keywords:</b>	- anomaly detection - gravitational waves - astrophysics - time-series
<b>licensing:</b>	NA
<b>task_types:</b>	- Anomaly detection
<b>ai_capability_measured:</b>	- Novel event detection in physical signals
<b>metrics:</b>	- ROC-AUC - Precision/Recall
<b>models:</b>	- Deep latent CNNs - Autoencoders
<b>ml_motif:</b>	- Time-series
<b>type:</b>	Dataset
<b>ml_task:</b>	- Anomaly detection
<b>solutions:</b>	Solution details are described in the referenced paper or repository.
<b>notes:</b>	NSF HDR A3D3 sponsored; prize pool and starter kit provided on Codabench.
<b>contact.name:</b>	HDR A3D3 Team
<b>contact.email:</b>	unknown
<b>results.links.name:</b>	ChatGPT LLM
<b>fair.reproducible:</b>	Yes
<b>fair.benchmark_ready:</b>	Yes
<b>id:</b>	hdr_ml_anomaly_challenge_gravitational_waves
<b>Citations:</b>	[65]

#### Ratings:

Rating	Value	Reason
dataset	5	Uses preprocessed LIGO/Virgo time series data at 4096 Hz, publicly available and standard in astrophysics.
documentation	4	Documentation includes challenge instructions, starter kit details, and baseline descriptions, but could benefit from more thorough tutorials and code walkthroughs.
metrics	4	ROC-AUC, precision, and recall metrics are clearly specified and appropriate for anomaly detection.
reference_solution	4	Baseline deep latent CNNs and autoencoders are provided and reproducible, but not extensively documented.
software	4	Benchmark platform provided on Codabench with starter kits and submission infrastructure. Code and baseline models are publicly accessible but not extensively maintained beyond the challenge.
specification	4	Well-defined anomaly detection task on gravitational-wave time series with clear input/output expectations and challenge constraints.



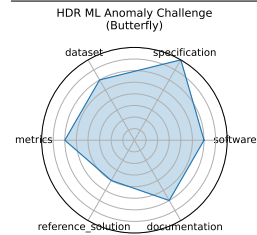
### 3.62 HDR ML Anomaly Challenge (Butterfly)

Image-based challenge for detecting butterfly hybrids in microscopy-driven species data. Participants evaluate models on Codabench using image segmentation/classification.

**date:** 2025-03-03  
**version:** v1.0  
**last\_updated:** 2025-03  
**expired:** unknown  
**valid:** yes  
**valid\_date:** 2025-03-03  
**url:** <https://www.codabench.org/competitions/3764/>  
**doi:** 10.48550/arXiv.2503.02112  
**domain:** Genomics; Image/CV  
**focus:** Detecting hybrid butterflies via image anomaly detection in genomic-informed dataset  
**keywords:** - anomaly detection - computer vision - genomics - butterfly hybrids  
**licensing:** NA  
**task\_types:** - Anomaly detection  
**ai\_capability\_measured:** - Hybrid detection in biological systems  
**metrics:** - Classification accuracy - F1 score  
**models:** - CNN-based detectors  
**ml\_motif:** - Image/CV  
**type:** Dataset  
**ml\_task:** - Anomaly detection  
**solutions:** Solution details are described in the referenced paper or repository.  
**notes:** Hybrid detection benchmarks hosted on Codabench  
**contact.name:** Imageomics/HDR Team  
**contact.email:** unknown  
**results.links.name:** ChatGPT LLM  
**fair.reproducible:** Yes  
**fair.benchmark\_ready:** Yes  
**id:** hdr\_ml\_anomaly\_challenge\_butterfly  
**Citations:** [66]

#### Ratings:

Rating	Value	Reason
dataset	3	Dataset consists of real detector data with synthetic anomaly injections; access is restricted and requires NDA, limiting openness and FAIR compliance.
documentation	3	Challenge website provides basic descriptions and evaluation metrics but lacks comprehensive tutorials or example workflows.
metrics	3	Standard metrics (ROC, F1, precision) are used; evaluation protocols are clear but not deeply elaborated.
reference_solution	2	Baselines are partially described but lack public code or reproducible execution scripts.
software	3	Codabench platform provides submission infrastructure but no fully maintained code repository or reproducible baseline implementations.
specification	4	Task is clearly described with domain-specific anomaly detection objectives and relevant physics motivation.



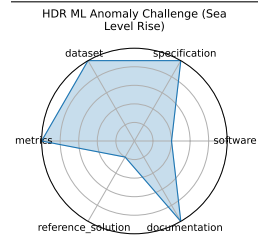
### 3.63 HDR ML Anomaly Challenge (Sea Level Rise)

A challenge combining North Atlantic sea-level time-series and satellite imagery to detect flooding anomalies. Models submitted via Codabench.

<b>date:</b>	2025-03-03
<b>version:</b>	v1.0
<b>last_updated:</b>	2025-03
<b>expired:</b>	unknown
<b>valid:</b>	yes
<b>valid_date:</b>	2025-03-03
<b>url:</b>	<a href="https://www.codabench.org/competitions/3223/">https://www.codabench.org/competitions/3223/</a>
<b>doi:</b>	10.48550/arXiv.2503.02112
<b>domain:</b>	Climate Science; Time-series, Image/CV
<b>focus:</b>	Detecting anomalous sea-level rise and flooding events via time-series and satellite imagery
<b>keywords:</b>	- anomaly detection - climate science - sea-level rise - time-series - remote sensing
<b>licensing:</b>	NA
<b>task_types:</b>	- Anomaly detection
<b>ai_capability_measured:</b>	- Detection of environmental anomalies
<b>metrics:</b>	- ROC-AUC - Precision/Recall
<b>models:</b>	- CNNs, RNNs, Transformers
<b>ml_motif:</b>	- Time-series, Image/CV
<b>type:</b>	Dataset
<b>ml_task:</b>	- Anomaly detection
<b>solutions:</b>	Solution details are described in the referenced paper or repository.
<b>notes:</b>	Sponsored by NSF HDR; integrates sensor and satellite data.
<b>contact.name:</b>	HDR A3D3 Team
<b>contact.email:</b>	unknown
<b>results.links.name:</b>	ChatGPT LLM
<b>fair.reproducible:</b>	Yes
<b>fair.benchmark_ready:</b>	Yes
<b>id:</b>	hdr_ml_anomaly_challenge_sea_level_rise
<b>Citations:</b>	[67]

#### Ratings:

Rating	Value	Reason
dataset	5	Uses preprocessed, public, and well-structured sensor and satellite data for the North Atlantic sea-level rise region.
documentation	5	Challenge page, starter kits, and related papers offer strong guidance for participants.
metrics	5	Standard metrics such as ROC-AUC, precision, and recall are specified and suitable for the anomaly detection tasks.
reference_solution	1	No starter models or baseline implementations linked or provided publicly.
software	2	Benchmark platform exists on Codabench, but no baseline code or maintained repository for reference solutions provided yet.
specification	5	Well-defined anomaly detection task combining satellite imagery and time-series data, with clear physical and domain-specific framing.



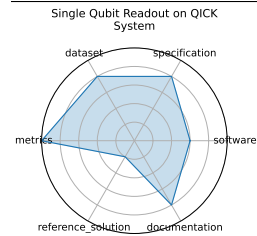
### 3.64 Single Qubit Readout on QICK System

Implements real-time ML models for single-qubit readout on the Quantum Instrumentation Control Kit (QICK), using hls4ml to deploy quantized neural networks on RFSoc FPGAs. Offers high-fidelity, low-latency quantum state discrimination. :contentReference[oaicite:0]{index=0}

**date:** 2025-01-24  
**version:** v1.0  
**last\_updated:** 2025-02  
**expired:** unknown  
**valid:** yes  
**valid\_date:** 2025-01-24  
**url:** <https://github.com/fastmachinelearning/ml-quantum-readout>  
**doi:** 10.48550/arXiv.2501.14663  
**domain:** Quantum Computing  
**focus:** Real-time single-qubit state classification using FPGA firmware  
**keywords:** - qubit readout - hls4ml - FPGA - QICK  
**licensing:** NA  
**task\_types:** - Classification  
**ai\_capability\_measured:** - Single-shot fidelity - inference latency  
**metrics:** - Accuracy - Latency  
**models:** - hls4ml quantized NN  
**ml\_motif:** - Real-time  
**type:** Benchmark  
**ml\_task:** - Supervised Learning  
**solutions:** Solution details are described in the referenced paper or repository.  
**notes:** Achieves ~96% fidelity with ~32 ns latency and low FPGA resource utilization.  
**contact.name:** Javier Campos, Giuseppe Di Guglielmo  
**contact.email:** unknown  
**datasets.links.name:** Zenodo: ml-quantum-readout dataset  
**datasets.links.url:** [zenodo.org/records/14427490](https://zenodo.org/records/14427490)  
**results.links.name:** ChatGPT LLM  
**fair.reproducible:** Yes  
**fair.benchmark\_ready:** Yes  
**id:** single\_qubit\_readout\_on\_qick\_system  
**Citations:** [68]

#### Ratings:

Rating	Value	Reason
dataset	4	Dataset hosted on Zenodo with structured data; however, detailed documentation on image acquisition and labeling pipeline is limited.
documentation	4	Codabench task page and GitHub repo provide descriptions and usage instructions, but detailed API or deployment tutorials are limited.
metrics	5	Standard classification metrics (accuracy, latency) are used and directly relevant to the quantum readout task.
reference_solution	1	No baseline or starter models with runnable code are linked publicly.
software	3	Code and FPGA firmware available on GitHub; integration with hls4ml demonstrated. Some deployment details and examples are provided but overall software maturity is moderate.
specification	4	Task clearly defined: real-time single-qubit state classification with latency and fidelity constraints. Labeling and ground truth definitions could be more explicit.





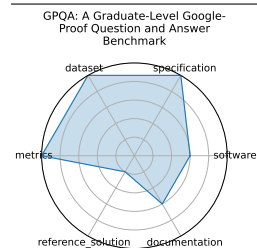
### 3.65 GPQA: A Graduate-Level Google-Proof Question and Answer Benchmark

Contains 448 challenging questions written by domain experts, with expert accuracy at 65% (74% discounting clear errors) and non-experts reaching just 34%. GPT-4 baseline scores ~39%-designed for scalable oversight evaluation.

<b>date:</b>	2023-11-20
<b>version:</b>	v1.0
<b>last_updated:</b>	2023-11
<b>expired:</b>	unknown
<b>valid:</b>	yes
<b>valid_date:</b>	2023-11-20
<b>url:</b>	<a href="https://arxiv.org/abs/2311.12022">https://arxiv.org/abs/2311.12022</a>
<b>doi:</b>	10.48550/arXiv.2311.12022
<b>domain:</b>	Science (Biology, Physics, Chemistry)
<b>focus:</b>	Graduate-level, expert-validated multiple-choice questions hard even with web access
<b>keywords:</b>	- Google-proof - multiple-choice - expert reasoning - science QA
<b>licensing:</b>	NA
<b>task_types:</b>	- Multiple choice
<b>ai_capability_measured:</b>	- Scientific reasoning - knowledge probing
<b>metrics:</b>	- Accuracy
<b>models:</b>	- GPT-4 baseline
<b>ml_motif:</b>	- Multiple choice
<b>type:</b>	Benchmark
<b>ml_task:</b>	- Multiple choice
<b>solutions:</b>	Solution details are described in the referenced paper or repository.
<b>notes:</b>	Google-proof, supports oversight research.
<b>contact.name:</b>	David Rein (NYU)
<b>contact.email:</b>	unknown
<b>datasets.links.name:</b>	GPQA dataset
<b>datasets.links.url:</b>	<a href="https://huggingface.co/datasets/AskPapers/GPQA">zip/HuggingFace</a>
<b>results.links.name:</b>	ChatGPT LLM
<b>fair.reproducible:</b>	Yes
<b>fair.benchmark_ready:</b>	Yes
<b>id:</b>	gpqa_a_graduate-level_google-proof_question_and_answer_benchmark
<b>Citations:</b>	[69]

#### Ratings:

Rating	Value	Reason
dataset	5	The GPQA dataset is publicly released, well curated, with metadata and clearly documented splits.
documentation	3	Documentation includes dataset description and benchmark instructions, but lacks detailed usage tutorials or pipelines.
metrics	5	Accuracy is the primary metric and is clearly defined and appropriate for multiple-choice QA.
reference_solution	1	No baseline implementations or starter code are linked or provided for reproduction.
software	3	Dataset and benchmark materials are publicly available via HuggingFace and GitHub, but no integrated runnable code or software framework is provided.
specification	5	Task is clearly defined as a multiple-choice benchmark requiring expert-level scientific reasoning. Input/output formats and evaluation criteria are well described.



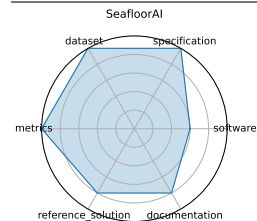
### 3.66 SeafloorAI

A first-of-its-kind dataset covering 17,300 sq.km of seafloor with 696K sonar images, 827K segmentation masks, and 696K natural-language descriptions plus ~7M QA pairs-designed for both vision and language-based ML models in marine science

**date:** 2024-12-13  
**version:** v1.0  
**last\_updated:** 2024-12  
**expired:** unknown  
**valid:** yes  
**valid\_date:** 2024-12-13  
**url:** <https://neurips.cc/virtual/2024/poster/97432>  
**doi:** 10.48550/arXiv.2411.00172  
**domain:** Marine Science; Vision-Language  
**focus:** Large-scale vision-language dataset for seafloor mapping and geological classification  
**keywords:** - sonar imagery - vision-language - seafloor mapping - segmentation - QA  
**licensing:** unknown  
**task\_types:** - Image segmentation - Vision-language QA  
**ai\_capability\_measured:** - Geospatial understanding - multimodal reasoning  
**metrics:** - Segmentation pixel accuracy - QA accuracy  
**models:** - SegFormer - ViLT-style multimodal models  
**ml\_motif:** - Vision-Language  
**type:** Dataset  
**ml\_task:** - Segmentation, QA  
**solutions:** Solution details are described in the referenced paper or repository.  
**notes:** Data processing code publicly available, covering five geological layers; curated with marine scientists  
**contact.name:** Kien X. Nguyen  
**contact.email:** unknown  
**datasets.links.name:** Sonar imagery + annotations  
**datasets.links.url:** unknown  
**results.links.name:** ChatGPT LLM  
**results.links.url:** unknown  
**fair.reproducible:** Yes  
**fair.benchmark\_ready:** Yes  
**id:** seafloorai  
**Citations:** [70]

#### Ratings:

Rating	Value	Reason
dataset	5	Large-scale, well-annotated sonar imagery dataset with segmentation masks and natural language descriptions; curated with domain experts.
documentation	4	Dataset description and data processing instructions are provided, but tutorials and benchmark usage guides are limited.
metrics	5	Standard segmentation pixel accuracy and QA accuracy metrics are clearly specified and appropriate for the tasks.
reference_solution	4	Some baseline models (e.g., SegFormer, ViLT-style) are mentioned, but reproducible code or pretrained weights are not fully available yet.
software	3	Data processing code is publicly available, but no full benchmark framework or runnable model implementations are provided yet.
specification	5	Tasks (image segmentation and vision-language QA) are clearly defined with geospatial and multimodal objectives well specified.



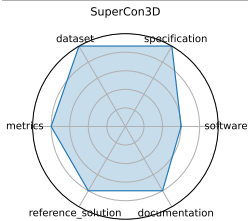
### 3.67 SuperCon3D

SuperCon3D introduces 3D crystal structures with associated critical temperatures (Tc) and two deep-learning models: SODNet (equivariant graph model) and DiffCSP-SC (diffusion generator) designed to screen and synthesize high-Tc candidates .

date:	2024-12-13
version:	v1.0
last_updated:	2024-12
expired:	unknown
valid:	yes
valid_date:	2024-12-13
url:	https://neurips.cc/virtual/2024/poster/97553
doi:	unknown
domain:	Materials Science; Superconductivity
focus:	Dataset and models for predicting and generating high-Tc superconductors using 3D crystal structures
keywords:	- superconductivity - crystal structures - equivariant GNN - generative models
licensing:	unknown
task_types:	- Regression (Tc prediction) - Generative modeling
ai_capability_measured:	- Structure-to-property prediction - structure generation
metrics:	- MAE (Tc) - Validity of generated structures
models:	- SODNet - DiffCSP-SC
ml_motif:	- Materials Modeling
type:	Dataset + Models
ml_task:	- Regression, Generation
solutions:	0
notes:	Demonstrates advantage of combining ordered and disordered structural data in model design
contact.name:	Zhong Zuo
contact.email:	unknown
results.links.name:	ChatGPT LLM
fair.reproducible:	Yes
fair.benchmark_ready:	Yes
id:	supercond
Citations:	[71]

**Ratings:**

Rating	Value	Reason
dataset	5	Dataset contains 3D crystal structures and associated properties; well-curated but not fully released publicly at this time.
documentation	4	Paper and GitHub provide good metadata and data processing descriptions; tutorials and user guides could be expanded.
metrics	4	Metrics such as MAE for Tc prediction and validity checks for generated structures are appropriate and clearly described.
reference_solution	4	Paper provides model architecture details and some training insights, but no complete open-source reference implementations yet.
software	3	Baseline models (SODNet, DiffCSP-SC) are described in the paper; however, fully reproducible code and pretrained models are not publicly available yet.
specification	5	Tasks for regression (Tc prediction) and generative modeling with clear input/output structures and domain constraints are well defined.



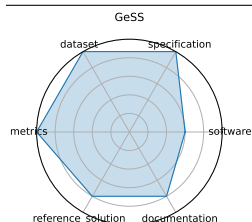
### 3.68 GeSS

GeSS provides 30 benchmark scenarios across particle physics, materials science, and biochemistry, evaluating 3 GDL backbones and 11 algorithms under covariate, concept, and conditional shifts, with varied OOD access .

**date:** 2024-12-13  
**version:** v1.0  
**last\_updated:** 2024-12  
**expired:** unknown  
**valid:** yes  
**valid\_date:** 2024-12-13  
**url:** <https://neurips.cc/virtual/2024/poster/97816>  
**doi:** unknown  
**domain:** Scientific ML; Geometric Deep Learning  
**focus:** Benchmark suite evaluating geometric deep learning models under real-world distribution shifts  
**keywords:** - geometric deep learning - distribution shift - OOD robustness - scientific applications  
**licensing:** unknown  
**task\_types:** - Classification - Regression  
**ai\_capability\_measured:** - OOD performance in scientific settings  
**metrics:** - Accuracy - RMSE - OOD robustness delta  
**models:** - GCN - EGNN - DimeNet++  
**ml\_motif:** - Geometric DL  
**type:** Benchmark  
**ml\_task:** - Classification, Regression  
**solutions:** 0  
**notes:** Includes no-OOD, unlabeled-OOD, and few-label scenarios .  
**contact.name:** Deyu Zou  
**contact.email:** unknown  
**results.links.name:** ChatGPT LLM  
**fair.reproducible:** Yes  
**fair.benchmark\_ready:** Yes  
**id:** gess  
**Citations:** [72]

#### Ratings:

Rating	Value	Reason
dataset	5	Curated datasets of 3D crystal structures and material properties are included and publicly available for reproducible research.
documentation	4	Paper and poster provide solid explanation of benchmarks and scientific motivation; more extensive user documentation forthcoming.
metrics	5	Uses well-established metrics such as MAE and structural validity for materials modeling, plus accuracy and OOD robustness deltas.
reference_solution	4	Two reference models (SODNet, DiffCSP-SC) are reported with results, code expected to be released soon.
software	3	Reference code expected post-conference; current public software availability limited. Benchmark infrastructure partially described but not fully released yet.
specification	5	Benchmark clearly defines OOD robustness scenarios with classification and regression tasks in scientific domains, though no explicit hardware constraints are given.



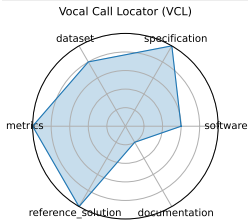
### 3.69 Vocal Call Locator (VCL)

The first large-scale benchmark (767K sounds across 9 conditions) for localizing rodent vocal calls using synchronized audio and video in standard lab environments, enabling systematic evaluation of sound-source localization algorithms in bioacoustics .

date:	2024-12-13
version:	v1.0
last_updated:	2024-12
expired:	unknown
valid:	yes
valid_date:	2024-12-13
url:	<a href="https://neurips.cc/virtual/2024/poster/97470">https://neurips.cc/virtual/2024/poster/97470</a>
doi:	unknown
domain:	Neuroscience; Bioacoustics
focus:	Benchmarking sound-source localization of rodent vocalizations from multi-channel audio
keywords:	- source localization - bioacoustics - time-series - SSL
licensing:	unknown
task_types:	- Sound source localization
ai_capability_measured:	- Source localization accuracy in bioacoustic settings
metrics:	- Localization error (cm) - Recall/Precision
models:	- CNN-based SSL models
ml_motif:	- Real-time
type:	Dataset
ml_task:	- Anomaly detection / localization
solutions:	0
notes:	Dataset spans real, simulated, and mixed audio; supports benchmarking across data types .
contact.name:	Ralph Peterson
contact.email:	unknown
results.links.name:	ChatGPT LLM
fair.reproducible:	Yes
fair.benchmark_ready:	Yes
id:	vocal_call_locator_vcl
Citations:	[73]

**Ratings:**

Rating	Value	Reason
dataset	4	Large-scale audio dataset covering real and simulated data with standardized splits, though exact data formats are not fully detailed.
documentation	1	Methodology and paper are thorough, but setup instructions and runnable code are not publicly provided, limiting user onboarding.
metrics	5	Includes localization error, precision, recall, and other relevant metrics for robust evaluation.
reference_solution	5	Multiple baselines evaluated over diverse models and architectures, supporting reproducibility of benchmark comparisons.
software	3	Some baseline CNN models for sound source localization are reported, but no publicly available or fully integrated runnable codebase yet.
specification	5	Well-defined localization tasks with multiple scenarios and real-world environment conditions; input/output formats clearly described.



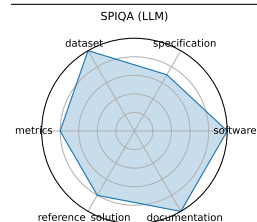
### 3.70 SPIQA (LLM)

A workshop version of SPIQA comparing 10 LLM adapter methods on the SPIQA benchmark with scientific diagram/questions. Highlights performance differences between chain-of-thought and end-to-end adapter models.

**date:** 2024-12-13  
**version:** v1.0  
**last\_updated:** 2024-12  
**expired:** unknown  
**valid:** yes  
**valid\_date:** 2024-12-13  
**url:** <https://neurips.cc/virtual/2024/poster/97575>  
**doi:** 10.48550/arXiv.2407.09413  
**domain:** Multimodal Scientific QA; Computer Vision  
**focus:** Evaluating LLMs on image-based scientific paper figure QA tasks (LLM Adapter performance)  
**keywords:** - multimodal QA - scientific figures - image+text - chain-of-thought prompting  
**licensing:** unknown  
**task\_types:** - Multimodal QA  
**ai\_capability\_measured:** - Visual reasoning - scientific figure understanding  
**metrics:** - Accuracy - F1 score  
**models:** - LLaVA - MiniGPT-4 - Owl-LLM adapter variants  
**ml\_motif:** - Multimodal QA  
**type:** Benchmark  
**ml\_task:** - Multimodal QA  
**solutions:** Solution details are described in the referenced paper or repository.  
**notes:** Companion to SPIQA main benchmark; compares adapter strategies using same images and QA pairs.  
**contact.name:** Xiaoyan Zhong  
**contact.email:** unknown  
**results.links.name:** ChatGPT LLM  
**fair.reproducible:** Yes  
**fair.benchmark\_ready:** Yes  
**id:** spiqqa\_llm  
**Citations:** [74]

#### Ratings:

Rating	Value	Reason
dataset	5	Full dataset available on Hugging Face with train/test/valid splits.
documentation	5	Full paper available
metrics	4	Reports accuracy and F1; fair but no visual reasoning-specific metric.
reference_solution	4	10 LLM adapter baselines; results included without constraints.
software	5	Well-documented codebase available on Github
specification	3.5	Task of QA over scientific figures is sufficient but not fully formalized in input/output terms. No hardware constraints.

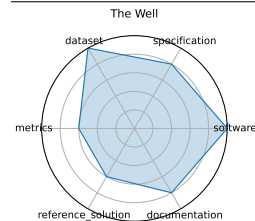


### 3.71 The Well

A 15 TB collection of ML-ready physics simulation datasets (HDF5), covering 16 domains-from biology to astrophysical magnetohydrodynamic simulations-with unified API and metadata. Ideal for training surrogate and foundation models on scientific data.

<b>date:</b>	2024-12-03
<b>version:</b>	v1.0
<b>last_updated:</b>	2025-06
<b>expired:</b>	unknown
<b>valid:</b>	yes
<b>valid_date:</b>	2024-12-03
<b>url:</b>	<a href="https://polymathic-ai.org/the_well/">https://polymathic-ai.org/the_well/</a>
<b>doi:</b>	unknown
<b>domain:</b>	biological systems, fluid dynamics, acoustic scattering, astrophysical MHD
<b>focus:</b>	Foundation model + surrogate dataset spanning 16 physical simulation domains
<b>keywords:</b>	- surrogate modeling - foundation model - physics simulations - spatiotemporal dynamics
<b>licensing:</b>	BSD 3-Clause License
<b>task_types:</b>	- Supervised Learning
<b>ai_capability_measured:</b>	- Surrogate modeling - physics-based prediction
<b>metrics:</b>	- Dataset size - Domain breadth
<b>models:</b>	- FNO baselines - U-Net baselines
<b>ml_motif:</b>	- Foundation model, Surrogate
<b>type:</b>	Dataset
<b>ml_task:</b>	- Supervised Learning
<b>solutions:</b>	1
<b>notes:</b>	Includes unified API and dataset metadata; see 2025 NeurIPS paper for full benchmark details. Size: 15 TB.
<b>contact.name:</b>	Ruben Ohana
<b>contact.email:</b>	<a href="mailto:rohana@flatironinstitute.org">rohana@flatironinstitute.org</a>
<b>datasets.links.name:</b>	16 simulation datasets
<b>datasets.links.url:</b>	HDF5) via PyPI/GitHub
<b>results.links.name:</b>	ChatGPT LLM
<b>fair.reproducible:</b>	Yes
<b>fair.benchmark_ready:</b>	Yes
<b>id:</b>	the_well
<b>Citations:</b>	[75]
<b>Ratings:</b>	

Rating	Value	Reason
dataset	5	15 TB of ML-ready HDF5 datasets across 16 physics domains. Public, well-structured, richly annotated, and designed with FAIR principles in mind.
documentation	4	The GitHub repo and NeurIPS paper provide detailed guidance on dataset use, structure, and training setup. Tutorials and walkthroughs could be expanded further.
metrics	3	Domain breadth and dataset size are emphasized. Standardized quantitative metrics for model evaluation (e.g., RMSE, accuracy) are not uniformly applied across all domains.
reference_solution	3	Includes FNO and U-Net baselines, but does not yet provide fully trained, reproducible models or scripts across all datasets.
software	5	BSD-licensed software and unified API are available via GitHub and PyPI. Supports loading and manipulating large HDF5 datasets across 16 domains.
specification	4	The benchmark includes clearly defined surrogate modeling tasks, data structure, and metadata. However, constraints and formal task specs vary slightly across domains.



## References

- [1] D. Hendrycks, C. Burns, and S. Kadavath, *Measuring massive multitask language understanding*, 2021. [Online]. Available: <https://arxiv.org/abs/2009.03300>.
- [2] A. Talmor, J. Herzig, N. Lourie, and J. Berant, *Commonsenseqa: A question answering challenge targeting commonsense knowledge*, 2019. arXiv: 1811.00937 [cs.CL]. [Online]. Available: <https://arxiv.org/abs/1811.00937>.
- [3] K. Sakaguchi, R. L. Bras, C. Bhagavatula, and Y. Choi, *Winogrande: An adversarial winograd schema challenge at scale*, 2019. arXiv: 1907.10641 [cs.CL]. [Online]. Available: <https://arxiv.org/abs/1907.10641>.
- [4] A. Srivastava, A. Rastogi, A. Rao, *et al.*, *Beyond the imitation game: Quantifying and extrapolating the capabilities of language models*, 2023. arXiv: 2206.04615 [cs.CL]. [Online]. Available: <https://arxiv.org/abs/2206.04615>.
- [5] T. Nguyen, J. Jewik, H. Bansal, P. Sharma, and A. Grover, *Climatelearn: Benchmarking machine learning for weather and climate modeling*, 2023. arXiv: 2307.01909 [cs.LG]. [Online]. Available: <https://arxiv.org/abs/2307.01909>.
- [6] 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>.
- [7] J. Roberts, K. Han, and S. Albanie, *Satin: A multi-task metadataset for classifying satellite imagery using vision-language models*, 2023. [Online]. Available: <https://huggingface.co/datasets/saral-ai/satimaget>.
- [8] F. J. Kiwit, M. Marso, P. Ross, C. A. Riofrío, J. Klepsch, and A. Luckow, “Application-oriented benchmarking of quantum generative learning using quark,” in *2023 IEEE International Conference on Quantum Computing and Engineering (QCE)*, IEEE, Sep. 2023, pp. 475–484. DOI: 10.1109/qce57702.2023.00061. [Online]. Available: <http://dx.doi.org/10.1109/QCE57702.2023.00061>.
- [9] L. Chanussot, A. Das, S. Goyal, *et al.*, “The open catalyst 2020 (oc20) dataset and community challenges,” *ACS Catalysis*, vol. 11, no. 10, pp. 6059–6072, 2021. DOI: 10.1021/acscatal.0c04525. [Online]. Available: <https://pubs.acs.org/doi/10.1021/acscatal.0c04525>.
- [10] R. Tran, J. Lan, M. Shuaibi, *et al.*, “The open catalyst 2022 (oc22) dataset and challenges for oxide electrocatalysts,” *ACS Catalysis*, vol. 13, no. 5, pp. 3066–3084, 2023. DOI: 10.1021/acscatal.2c05426. [Online]. Available: <https://pubs.acs.org/doi/10.1021/acscatal.2c05426>.
- [11] L. Chanussot, A. Das, S. Goyal, *et al.*, “Open catalyst 2020 (oc20) dataset and community challenges,” *ACS Catalysis*, vol. 11, no. 10, pp. 6059–6072, 2021. DOI: 10.1021/acscatal.0c04525. eprint: <https://doi.org/10.1021/acscatal.0c04525>. [Online]. Available: <https://doi.org/10.1021/acscatal.0c04525>.
- [12] R. Tran, J. Lan, M. Shuaibi, *et al.*, “The open catalyst 2022 (oc22) dataset and challenges for oxide electrocatalysts,” *ACS Catalysis*, vol. 13, no. 5, pp. 3066–3084, Feb. 2023, ISSN: 2155-5435. DOI: 10.1021/acscatal.2c05426. [Online]. Available: <http://dx.doi.org/10.1021/acscatal.2c05426>.
- [13] 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*, 2020. arXiv: 2009.13081 [cs.CL]. [Online]. Available: <https://arxiv.org/abs/2009.13081>.
- [14] A. Jain, S. P. Ong, G. Hautier, *et al.*, “The materials project: A materials genome approach,” *APL Materials*, vol. 1, no. 1, 2013. DOI: 10.1063/1.4812323. [Online]. Available: <https://materialsproject.org/>.
- [15] 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>.



- [16] P. Clark, I. Cowhey, and O. Etzioni, “Think you have solved question answering? try arc, the ai2 reasoning challenge,” in *EMNLP 2018*, 2018, pp. 237–248. [Online]. Available: <https://allenai.org/data/arc>.
- [17] L. Phan, A. Gatti, Z. Han, *et al.*, *Humanity’s last exam*, 2025. arXiv: 2501.14249 [cs.LG]. [Online]. Available: <https://arxiv.org/abs/2501.14249>.
- [18] E. Glazer, E. Erdil, T. Besiroglu, *et al.*, *Frontiermath: A benchmark for evaluating advanced mathematical reasoning in ai*, 2024. arXiv: 2411.04872 [cs.AI]. [Online]. Available: <https://arxiv.org/abs/2411.04872>.
- [19] M. Tian, L. Gao, S. D. Zhang, *et al.*, *Scicode: A research coding benchmark curated by scientists*, 2024. arXiv: 2407.13168 [cs.AI]. [Online]. Available: <https://arxiv.org/abs/2407.13168>.
- [20] TBD, *Aime*, [Online accessed 2025-06-24], Mar. 2025. [Online]. Available: <https://www.vals.ai/benchmarks/aime-2025-03-13>.
- [21] HuggingFaceH4, *Math-500*, 2025. [Online]. Available: <https://huggingface.co/datasets/HuggingFaceH4/MATH-500>.
- [22] H. Cui, Z. Shamsi, G. Cheon, *et al.*, *Curie: Evaluating llms on multitask scientific long context understanding and reasoning*, 2025. arXiv: 2503.13517 [cs.CL]. [Online]. Available: <https://arxiv.org/abs/2503.13517>.
- [23] N. Mudur, H. Cui, S. Venugopalan, P. Raccuglia, M. P. Brenner, and P. Norgaard, *Feabench: Evaluating language models on multiphysics reasoning ability*, 2025. arXiv: 2504.06260 [cs.AI]. [Online]. Available: <https://arxiv.org/abs/2504.06260>.
- [24] X. Zhong, Y. Gao, and S. Gururangan, *Spiga: Scientific paper image question answering*, 2024. [Online]. Available: <https://arxiv.org/abs/2407.09413>.
- [25] E. Luo, J. Jia, Y. Xiong, *et al.*, *Benchmarking ai scientists in omics data-driven biological research*, 2025. arXiv: 2505.08341 [cs.AI]. [Online]. Available: <https://arxiv.org/abs/2505.08341>.
- [26] Y. Fang, N. Zhang, Z. Chen, L. Guo, X. Fan, and H. Chen, *Domain-agnostic molecular generation with chemical feedback*, 2024. arXiv: 2301.11259 [cs.LG]. [Online]. Available: <https://arxiv.org/abs/2301.11259>.
- [27] W. Hu, M. Fey, M. Zitnik, *et al.*, *Open graph benchmark: Datasets for machine learning on graphs*, 2021. arXiv: 2005.00687 [cs.LG]. [Online]. Available: <https://arxiv.org/abs/2005.00687>.
- [28] K. Choudhary, D. Wines, K. Li, *et al.*, “JARVIS-Leaderboard: A large scale benchmark of materials design methods,” *npj Computational Materials*, vol. 10, no. 1, p. 93, 2024. DOI: 10.1038/s41524-024-01259-w. [Online]. Available: <https://doi.org/10.1038/s41524-024-01259-w>.
- [29] A. Garza, C. Challu, and M. Mergenthaler-Canseco, *Timegpt-1*, 2024. arXiv: 2310.03589 [cs.LG]. [Online]. Available: <https://arxiv.org/abs/2310.03589>.
- [30] C. Challu, K. G. Olivares, B. N. Oreshkin, F. G. Ramirez, M. M. Canseco, and A. Dubrawski, “Nhits: Neural hierarchical interpolation for time series forecasting,” in *Proceedings of the AAAI conference on artificial intelligence*, vol. 37, 2023, pp. 6989–6997.
- [31] M. Jin, S. Wang, L. Ma, *et al.*, *Time-llm: Time series forecasting by reprogramming large language models*, 2024. arXiv: 2310.01728 [cs.LG]. [Online]. Available: <https://arxiv.org/abs/2310.01728>.
- [32] K. G. Olivares, C. Challu, F. Garza, M. M. Canseco, and A. Dubrawski, *Neuralforecast: User friendly state-of-the-art neural forecasting models*. PyCon Salt Lake City, Utah, US 2022, 2022. [Online]. Available: <https://github.com/Nixtla/neuralforecast>.
- [33] S. Mo, *Vllm performance dashboard*, 2024. [Online]. Available: <https://simon-mo-workspace.observablehq.cloud/vllm-dashboard-v0/>.

- [34] W. Kwon, Z. Li, S. Zhuang, *et al.*, “Efficient memory management for large language model serving with pagedattention,” in *Proceedings of the 29th Symposium on Operating Systems Principles*, ser. SOSP ’23, Koblenz, Germany: Association for Computing Machinery, 2023, pp. 611–626. DOI: 10.1145/3600006.3613165. [Online]. Available: <https://doi.org/10.1145/3600006.3613165>.
- [35] L. Zheng, L. Yin, Z. Xie, *et al.*, *Sglang: Efficient execution of structured language model programs*, 2024. arXiv: 2312.07104 [cs.AI]. [Online]. Available: <https://arxiv.org/abs/2312.07104>.
- [36] J. Duarte, N. Tran, B. Hawks, *et al.*, *Fastml science benchmarks: Accelerating real-time scientific edge machine learning*, 2022. arXiv: 2207.07958 [cs.LG]. [Online]. Available: <https://arxiv.org/abs/2207.07958>.
- [37] J. Duarte, N. Tran, B. Hawks, *et al.*, *Fastml science benchmarks: Accelerating real-time scientific edge machine learning*, 2022. arXiv: 2207.07958 [cs.LG]. [Online]. Available: <https://arxiv.org/abs/2207.07958>.
- [38] J. Duarte, N. Tran, B. Hawks, *et al.*, *Fastml science benchmarks: Accelerating real-time scientific edge machine learning*, 2022. arXiv: 2207.07958 [cs.LG]. [Online]. Available: <https://arxiv.org/abs/2207.07958>.
- [39] D. Kafkes and J. S. John, *Boostr: A dataset for accelerator control systems*, 2021. arXiv: 2101.08359 [physics.acc-ph]. [Online]. Available: <https://arxiv.org/abs/2101.08359>.
- [40] P. Odagiu, Z. Que, J. Duarte, *et al.*, *Ultrafast jet classification on fpgas for the hl-lhc*, 2024. DOI: <https://doi.org/10.1088/2632-2153/ad5f10>. arXiv: 2402.01876 [hep-ex]. [Online]. Available: <https://arxiv.org/abs/2402.01876>.
- [41] 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](https://indico.cern.ch/event/1387540/contributions/6153618/attachments/2948441/5182077/fast_ml_magnets_2024_final.pdf).
- [42] J. Kvapil, G. Borca-Tasciuc, H. Bossi, *et al.*, *Intelligent experiments through real-time ai: Fast data processing and autonomous detector control for sphenix and future eic detectors*, 2025. arXiv: 2501.04845 [physics.ins-det]. [Online]. Available: <https://arxiv.org/abs/2501.04845>.
- [43] J. Weitz, D. Demler, L. McDermott, N. Tran, and J. Duarte, *Neural architecture codesign for fast physics applications*, 2025. arXiv: 2501.05515 [cs.LG]. [Online]. Available: <https://arxiv.org/abs/2501.05515>.
- [44] B. Parpillon, C. Syal, J. Yoo, *et al.*, *Smart pixels: In-pixel ai for on-sensor data filtering*, 2024. arXiv: 2406.14860 [physics.ins-det]. [Online]. Available: <https://arxiv.org/abs/2406.14860>.
- [45] Z. Liu, H. Sharma, J.-S. Park, *et al.*, *Braggmn: Fast x-ray bragg peak analysis using deep learning*, 2021. arXiv: 2008.08198 [eess.IV]. [Online]. Available: <https://arxiv.org/abs/2008.08198>.
- [46] 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>.
- [47] Y. Wei, R. F. Forelli, C. Hansen, *et al.*, *Low latency optical-based mode tracking with machine learning deployed on fpgas on a tokamak*, 2024. DOI: <https://doi.org/10.1063/5.0190354>. arXiv: 2312.00128 [physics.plasm-ph]. [Online]. Available: <https://arxiv.org/abs/2312.00128>.
- [48] W. Gao, F. Tang, L. Wang, *et al.*, *Aibench: An industry standard internet service ai benchmark suite*, 2019. arXiv: 1908.08998 [cs.CV]. [Online]. Available: <https://arxiv.org/abs/1908.08998>.
- [49] W. Gao, J. Zhan, L. Wang, *et al.*, *Bigdatabench: A scalable and unified big data and ai benchmark suite*, 2018. arXiv: 1802.08254 [cs.DC]. [Online]. Available: <https://arxiv.org/abs/1802.08254>.

- [50] S. Farrell, M. Emani, J. Balma, *et al.*, *Mlperf hpc: A holistic benchmark suite for scientific machine learning on hpc systems*, 2021. arXiv: 2110.11466 [cs.LG]. [Online]. Available: <https://arxiv.org/abs/2110.11466>.
- [51] J. Thiyyagalingam, G. von Laszewski, J. Yin, *et al.*, “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, ISBN: 978-3-031-23220-6.
- [52] T. Aarrestad, E. Govorkova, J. Ngadiuba, E. Puljak, M. Pierini, and K. A. Wozniak, *Unsupervised new physics detection at 40 mhz: Training dataset*, 2021. DOI: 10.5281/ZENODO.5046389. [Online]. Available: <https://zenodo.org/record/5046389>.
- [53] A. Karargyris, R. Umeton, M. J. Sheller, *et al.*, “Federated benchmarking of medical artificial intelligence with medperf,” *Nature Machine Intelligence*, vol. 5, no. 7, pp. 799–810, Jul. 2023. DOI: 10.1038/s42256-023-00652-2. [Online]. Available: <https://doi.org/10.1038/s42256-023-00652-2>.
- [54] C. Krause, M. F. Giannelli, G. Kasieczka, *et al.*, *Calochallenge 2022: A community challenge for fast calorimeter simulation*, 2024. arXiv: 2410.21611 [physics.ins-det]. [Online]. Available: <https://arxiv.org/abs/2410.21611>.
- [55] A. Blum and M. Hardt, “The ladder: A reliable leaderboard for machine learning competitions,” in *Proceedings of the 32nd International Conference on Machine Learning*, F. Bach and D. Blei, Eds., ser. Proceedings of Machine Learning Research, vol. 37, Lille, France: PMLR, Jul. 2015, pp. 1006–1014. [Online]. Available: <https://proceedings.mlr.press/v37/blum15.html>.
- [56] Z. Xu, S. Escalera, A. Pavão, *et al.*, “Codabench: Flexible, easy-to-use, and reproducible meta-benchmark platform,” *Patterns*, vol. 3, no. 7, p. 100543, Jul. 2022, ISSN: 2666-3899. DOI: 10.1016/j.patter.2022.100543. [Online]. Available: <http://dx.doi.org/10.1016/j.patter.2022.100543>.
- [57] P. Luszczek, “Sabath: Fair metadata technology for surrogate benchmarks,” University of Tennessee, Tech. Rep., 2021. [Online]. Available: <https://github.com/icl-utk-edu/slip/tree/sabath>.
- [58] M. Takamoto, T. Praditia, R. Leiteritz, *et al.*, *Pdebench: An extensive benchmark for scientific machine learning*, 2024. arXiv: 2210.07182 [cs.LG]. [Online]. Available: <https://arxiv.org/abs/2210.07182>.
- [59] K. T. Chitty-Venkata, S. Raskar, B. Kale, *et al.*, “Llm-inference-bench: Inference benchmarking of large language models on ai accelerators,” in *SC24-W: Workshops of the International Conference for High Performance Computing, Networking, Storage and Analysis*, 2024, pp. 1362–1379. DOI: 10.1109/SCW63240.2024.00178.
- [60] K. Khrabrov, A. Ber, A. Tsypin, *et al.*, *Delta-squared dft: A universal quantum chemistry dataset of drug-like molecules and a benchmark for neural network potentials*, 2024. arXiv: 2406.14347 [physics.chem-ph]. [Online]. Available: <https://arxiv.org/abs/2406.14347>.
- [61] Y. Wang, T. Wang, Y. Zhang, *et al.*, “Urbandatalayer: A unified data pipeline for urban science,” in *Advances in Neural Information Processing Systems*, A. Globerson, L. Mackey, D. Belgrave, *et al.*, 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](https://proceedings.neurips.cc/paper_files/paper/2024/file/0db7f135f6991e8cec5e516ecc66bfba-Paper-Datasets_and_Benchmarks_Track.pdf).
- [62] T. Shen, H. Wang, J. Zhang, *et al.*, *Exploring user retrieval integration towards large language models for cross-domain sequential recommendation*, 2024. arXiv: 2406.03085 [cs.LG]. [Online]. Available: <https://arxiv.org/abs/2406.03085>.
- [63] A. A. Abud, B. Abi, R. Acciarri, *et al.*, *Deep underground neutrino experiment (dune) near detector conceptual design report*, 2021. arXiv: 2103.13910 [physics.ins-det]. [Online]. Available: <https://arxiv.org/abs/2103.13910>.

- [64] R. Bushuiev, A. Bushuiev, N. F. de Jonge, *et al.*, “Massspecgym: A benchmark for the discovery and identification of molecules,” in *Advances in Neural Information Processing Systems*, A. Globerson, L. Mackey, D. Belgrave, *et al.*, 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](https://proceedings.neurips.cc/paper_files/paper/2024/file/c6c31413d5c53b7d1c343c1498734b0f-Paper-Datasets_and_Benchmarks_Track.pdf).
- [65] E. G. Campolongo, Y.-T. Chou, E. Govorkova, *et al.*, *Building machine learning challenges for anomaly detection in science*, 2025. arXiv: 2503.02112 [cs.LG]. [Online]. Available: <https://arxiv.org/abs/2503.02112>.
- [66] E. G. Campolongo, Y.-T. Chou, E. Govorkova, *et al.*, *Building machine learning challenges for anomaly detection in science*, 2025. arXiv: 2503.02112 [cs.LG]. [Online]. Available: <https://arxiv.org/abs/2503.02112>.
- [67] E. G. Campolongo, Y.-T. Chou, E. Govorkova, *et al.*, *Building machine learning challenges for anomaly detection in science*, 2025. arXiv: 2503.02112 [cs.LG]. [Online]. Available: <https://arxiv.org/abs/2503.02112>.
- [68] G. D. Guglielmo, B. Du, J. Campos, *et al.*, *End-to-end workflow for machine learning-based qubit readout with qick and hls4ml*, 2025. arXiv: 2501.14663 [quant-ph]. [Online]. Available: <https://arxiv.org/abs/2501.14663>.
- [69] D. Rein, B. L. Hou, A. C. Stickland, *et al.*, *Gpqa: A graduate-level google-proof q and a benchmark*, 2023. arXiv: 2311.12022 [cs.AI]. [Online]. Available: <https://arxiv.org/abs/2311.12022>.
- [70] K. X. Nguyen, F. Qiao, A. Trembanis, and X. Peng, *Seafloorai: A large-scale vision-language dataset for seafloor geological survey*, 2024. arXiv: 2411.00172 [cs.CV]. [Online]. Available: <https://arxiv.org/abs/2411.00172>.
- [71] P. Chen, L. Peng, R. Jiao, *et al.*, “Learning superconductivity from ordered and disordered material structures,” in *Advances in Neural Information Processing Systems*, A. Globerson, L. Mackey, D. Belgrave, *et al.*, 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](https://proceedings.neurips.cc/paper_files/paper/2024/file/c4e3b55ed4ac9ba52d7df11f8bddbbf4-Paper-Datasets_and_Benchmarks_Track.pdf).
- [72] 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, *et al.*, 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](https://proceedings.neurips.cc/paper_files/paper/2024/file/a8063075b00168dc39bc81683619f1a8-Paper-Datasets_and_Benchmarks_Track.pdf).
- [73] R. E. Peterson, A. Tanelus, C. Ick, *et al.*, “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, *et al.*, 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](https://proceedings.neurips.cc/paper_files/paper/2024/file/c00d37d6b04d73b870b963a4d70051c1-Paper-Datasets_and_Benchmarks_Track.pdf).
- [74] S. Pramanick, R. Chellappa, and S. Venugopalan, *Spiga: A dataset for multimodal question answering on scientific papers*, 2025. arXiv: 2407.09413 [cs.CL]. [Online]. Available: <https://arxiv.org/abs/2407.09413>.
- [75] R. Ohana, M. McCabe, L. Meyer, *et al.*, “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, *et al.*, 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](https://proceedings.neurips.cc/paper_files/paper/2024/file/4f9a5acd91ac76569f2fe291b1f4772b-Paper-Datasets_and_Benchmarks_Track.pdf).