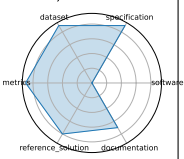
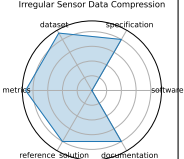
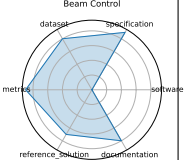
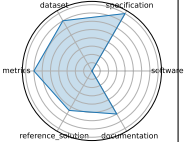
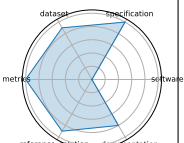
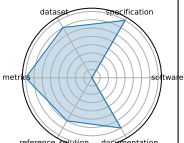



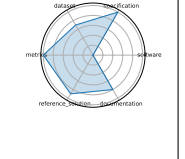
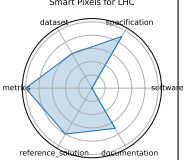
## 1 Benchmark Overview Table

Ratings	Name	Domain	Focus	Keywords	Task Types	AI Capability	Metrics	Models	Citation
	Jet Classification	Particle Physics	Real-time classification of particle jets using HL-LHC simulation features	classification, real-time ML, jet tagging, QKeras	Classification	Real-time inference, model compression performance	Accuracy, AUC	Keras DNN, QKeras quantized DNN	[1]⇒
	Irregular Sensor Data Compression	Particle Physics	Real-time compression of sparse sensor data with autoencoders	compression, autoencoder, sparse data, irregular sampling	Compression	Reconstruction quality, compression efficiency	MSE, Compression ratio	Autoencoder, Quantized autoencoder	[2]⇒
	Beam Control	Accelerators and Magnets	Reinforcement RL, beam stabilization, control systems, simulation	Reinforcement RL, beam stabilization, control systems, simulation	Control	Policy performance in simulated accelerator control	Stability, Control loss	DDPG, PPO (planned)	[2], [3]⇒

Continued on next page

Ratings	Name	Domain	Focus	Keywords	Task Types	AI Capability	Metrics	Models	Citation
	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 FPGA constraints	Accuracy, Latency, Resource utilization	MLP, Deep Sets, Interaction Network	[4]⇒
	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)	
	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)	[5]⇒

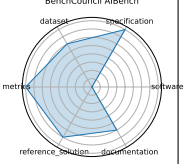
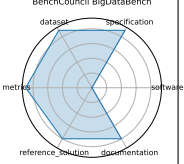
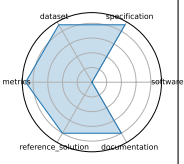
Continued on next page

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 ( $\mu$ s), Resource utilization (LUT/FF/BRAM/DSP)	Bipartite Graph Network with Set Transformers (BGN-ST), GarNet (edge-AM/DSP)	[6]⇒
	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)	[7]⇒
	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	[8]⇒

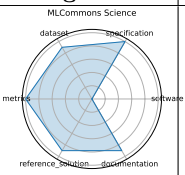
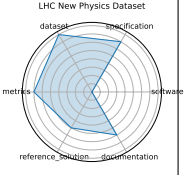
Continued on next page

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	[9]⇒
	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)	[10]⇒
	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	[11]⇒

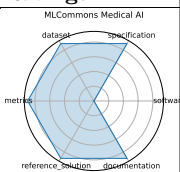
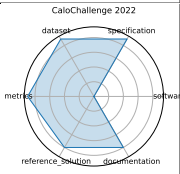
Continued on next page

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	[12]⇒
	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	[13]⇒
	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	[14]⇒

Continued on next page

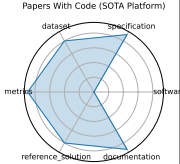
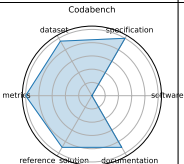
Ratings	Name	Domain	Focus	Keywords	Task Types	AI Capability	Metrics	Models	Citation
	MLCommons Science	Earthquake, Satellite Image, Drug Discovery, Electron Microscope, CFD	AI benchmarks for scientific applications including time-series, imaging, and simulation	science AI, benchmark, MLCommons, HPC	Time-series analysis, Image classification, Simulation surrogate modeling	Inference accuracy, simulation speed-up, generalization	MAE, Accuracy, Speedup vs simulation	CNN, GNN, Transformer	[15]⇒
	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	[16]⇒

Continued on next page

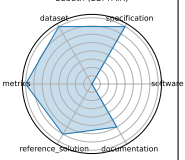
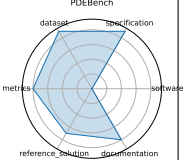
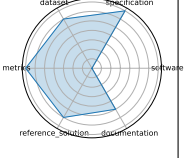
Ratings	Name	Domain	Focus	Keywords	Task Types	AI Capability	Metrics	Models	Citation
	MLCommons Medical AI	Healthcare; Medical AI	Federated benchmarking and evaluation of medical AI models across diverse real-world clinical data	medical AI, federated evaluation, privacy-preserving, fairness, healthcare benchmarks	Federated evaluation, Model validation	Clinical accuracy, fairness, generalizability, privacy compliance	ROC AUC, Accuracy, Fairness metrics	MedPerf-validated CNNs, GaNDLF workflows	[17]⇒
	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	[18]⇒

Continued on next page

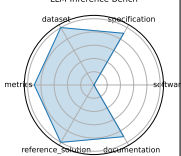
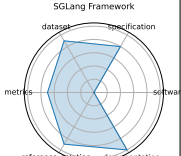
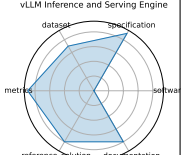


Ratings	Name	Domain	Focus	Keywords	Task Types	AI Capability	Metrics	Models	Citation
 <p>Papers With Code (SOTA Platform)</p>	Papers With Code (SOTA Platform)	General ML; All domains	Open platform tracking state-of-the-art results, benchmarks, and implementations across ML tasks and papers	leaderboard, 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	[19]⇒
 <p>Codabench</p>	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	[20]⇒

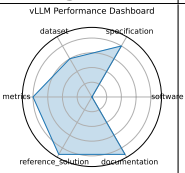
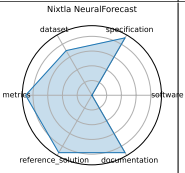
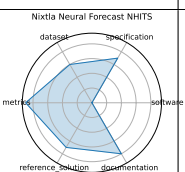
Continued on next page

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	N/A	[21]⇒
	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	[22]⇒
	The Well	biological systems, fluid dynamics, acoustic scattering, astrophysical 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	[23]⇒

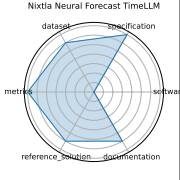
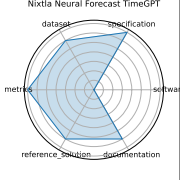
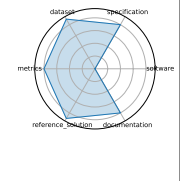
Continued on next page

Ratings	Name	Domain	Focus	Keywords	Task Types	AI Capability	Metrics	Models	Citation
	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	[24]⇒
	SGLang Framework	LLM Vision	Fast serving framework for LLMs and vision-language models	LLM serving, vision-language, RadixAttention, performance, JSON decoding	Model serving framework	Serving throughput, JSON/task-specific latency	Tokens/sec, Time-to-first-token, Throughput gain vs baseline	LLaVA, DeepSeek, Llama	[25]⇒
	vLLM Inference and Serving Engine	LLM; HPC/inference	High-throughput, memory-efficient inference and serving engine for LLMs	LLM inference, PagedAttention, CUDA graph, streaming API, quantization	Inference Benchmarking	Throughput, latency, memory efficiency	Tokens/sec, Time to First Token (TTFT), Memory footprint	LLaMA, Mixtral, FlashAttention-based models	[26]⇒

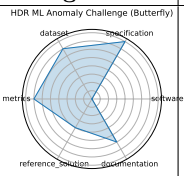

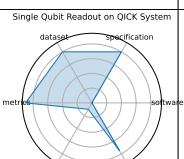
Continued on next page

Ratings	Name	Domain	Focus	Keywords	Task Types	AI Capability	Metrics	Models	Citation
	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	[27]⇒
	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	[28]⇒
	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	[29]⇒


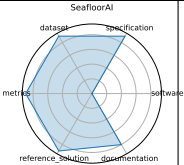
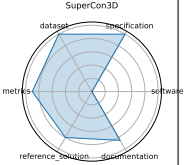
Continued on next page

Ratings	Name	Domain	Focus	Keywords	Task Types	AI Capability	Metrics	Models	Citation
	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	[30]⇒
	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	[31]⇒
	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	[32]⇒

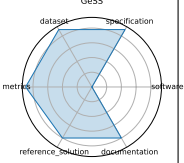
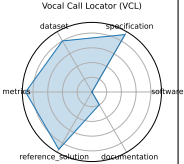
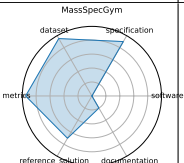
Continued on next page

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	[32]⇒
	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	[32]⇒
	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	[33]⇒

Continued on next page

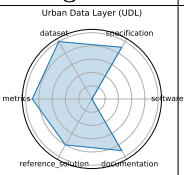
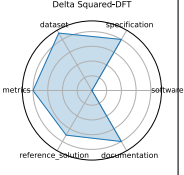
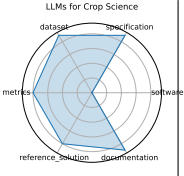
Ratings	Name	Domain	Focus	Keywords	Task Types	AI Capability	Metrics	Models	Citation
	GPQA: A Graduate-Level Google-Proof Question and Answer Benchmark	Science (Biology, Physics, Chemistry)	Graduate-level, expert-validated multiple-choice questions hard even with web access	Google-proof, multiple-choice, expert reasoning, science QA	Multiple choice	Scientific reasoning, knowledge probing	Accuracy	GPT-4 baseline	[34]⇒
	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	[35]⇒
	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, DiffC-SP-SC	[36]⇒

Continued on next page

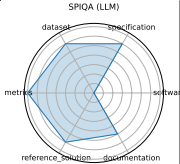
Ratings	Name	Domain	Focus	Keywords	Task Types	AI Capability	Metrics	Models	Citation
	GeSS	Scientific ML; Geometric Deep Learning	Benchmark suite evaluating geometric deep learning models under real-world distribution shifts	geometric deep learning, distribution shift, OOD robustness, scientific applications	Classification, Regression	OOD performance in scientific settings	Accuracy, RMSE, OOD robustness delta	GCN, EGNN, DimeNet++	[37]⇒
	Vocal Call Locator (VCL)	Neuroscience, Bioacoustics	Benchmarking sound-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	[38]⇒
	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	[39]⇒

Continued on next page



Ratings	Name	Domain	Focus	Keywords	Task Types	AI Capability	Metrics	Models	Citation
	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	[40]⇒
	Delta Squared-DFT	Computational Chemistry; Materials Science	Benchmarking density functional theory, Delta Squared-ML correction, reaction energetics, quantum chemistry	machine-learning corrections to DFT using Delta Squared-trained models for reaction energies	Regression	High-accuracy energy prediction, DFT correction	Mean Absolute Error (eV), Energy ranking accuracy	Delta Squared-ML correction networks, Kernel ridge regression	[41]⇒
	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	[42]⇒

Continued on next page

Ratings	Name	Domain	Focus	Keywords	Task Types	AI Capability	Metrics	Models	Citation
	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	[43]⇒

## 2 Radar Chart Table

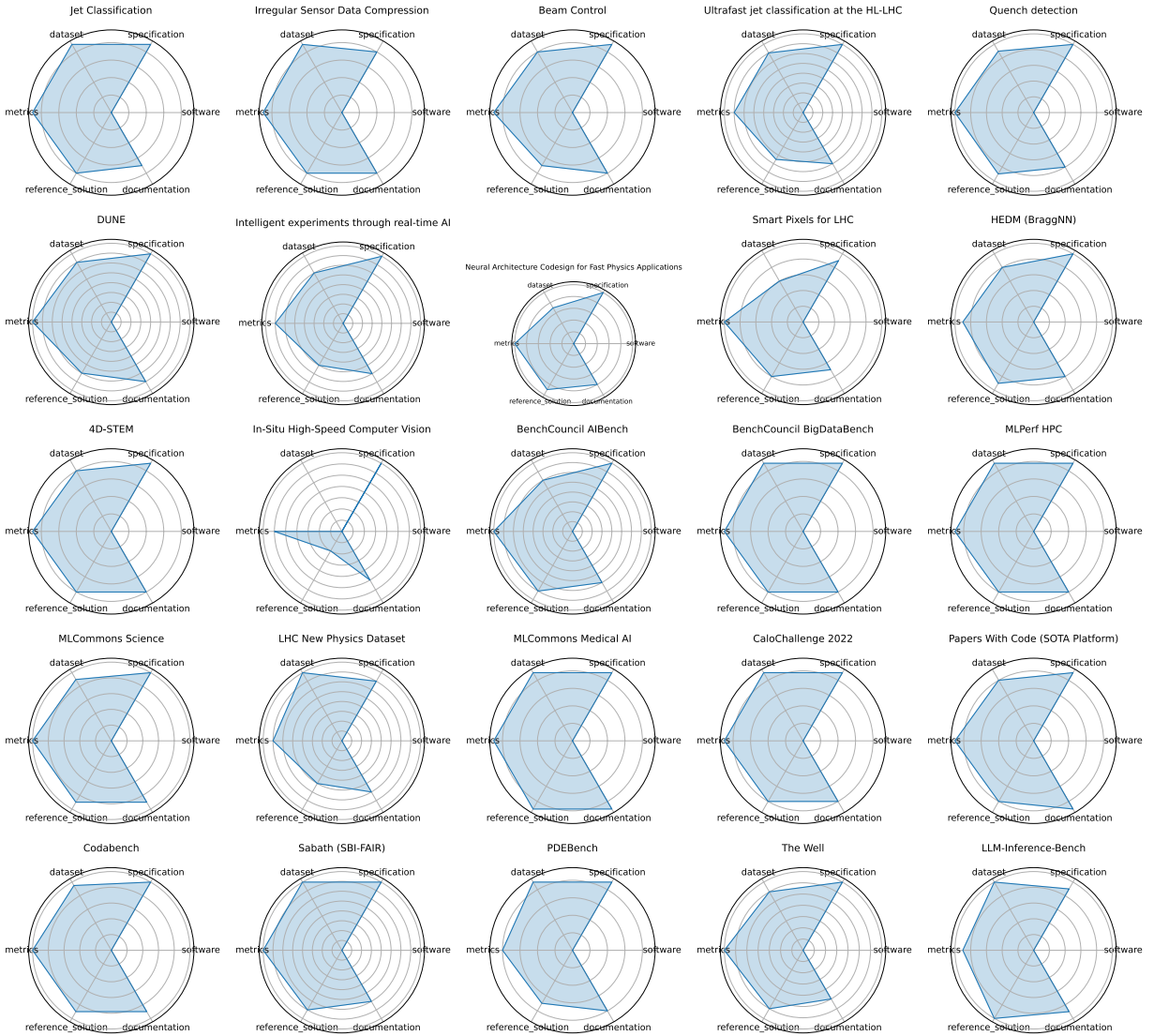


Figure 1: Radar chart overview (page 1)

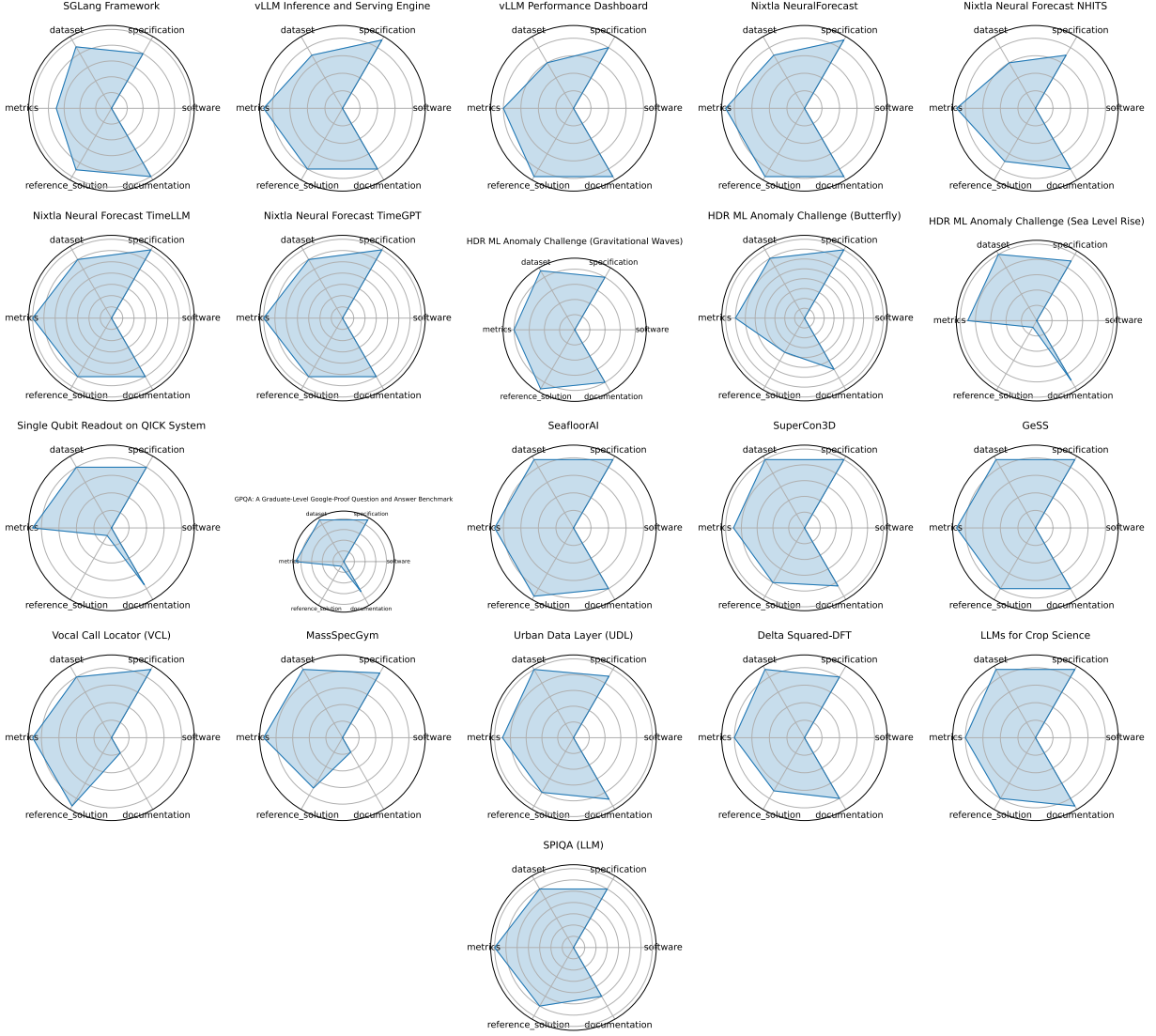
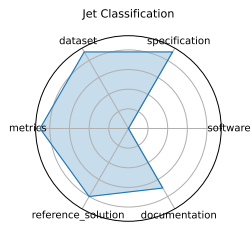


Figure 2: Radar chart overview (page 2)

### 3 Benchmark Details

#### 4 Jet Classification

**date:** 2024-05-01  
**last\_updated:** 2024-05  
**expired:** unknown  
**valid:** yes  
**url:** <https://github.com/fastmachinelearning/fastml-science/tree/main/jet-classify>  
**domain:** Particle Physics  
**focus:** Real-time classification of particle jets using HL-LHC simulation features  
**keywords:** - classification - real-time ML - jet tagging - QKeras  
**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  
**notes:** Includes both float and quantized models using QKeras  
**contact.name:** Jules Muhizi  
**contact.email:** unknown  
**dataset.name:** JetClass  
**dataset.url:** <https://zenodo.org/record/6619768>  
**results.name:** ChatGPT LLM  
**results.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  
**ratings.software.rating:** 0  
**ratings.software.reason:** Not analyzed.  
**ratings.specification.rating:** 9.0  
**ratings.specification.reason:** 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.  
**ratings.dataset.rating:** 9.0  
**ratings.dataset.reason:** Public, versioned, and FAIR-compliant; includes metadata, splits, and licensing; well-integrated with HuggingFace and other ML libraries.  
**ratings.metrics.rating:** 9.0  
**ratings.metrics.reason:** Accuracy is a simple, reproducible metric aligned with task goals; no ambiguity in evaluation.  
**ratings.reference\_solution.rating:** 8.0  
**ratings.reference\_solution.reason:** Several baseline models (e.g., BERT, RoBERTa) are reported with scores; implementations exist in public repos, but not bundled as an official starter kit.  
**ratings.documentation.rating:** 7.0  
**ratings.documentation.reason:** Clear paper, GitHub repo, and integration with HuggingFace Datasets; full reproducibility requires manually connecting models to dataset.  
**id:** jet\_classification  
**Citations:** [1]



**Ratings:**

## 5 Irregular Sensor Data Compression

**date:** 2024-05-01

**last\_updated:** 2024-05

**expired:** unknown

**valid:** yes

**url:** <https://github.com/fastmachinelearning/fastml-science/tree/main/sensor-data-compression>

**domain:** Particle Physics

**focus:** Real-time compression of sparse sensor data with autoencoders

**keywords:** - compression - autoencoder - sparse data - irregular sampling

**task\_types:** - Compression

**ai\_capability\_measured:** - Reconstruction quality - compression efficiency

**metrics:** - MSE - Compression ratio

**models:** - Autoencoder - Quantized autoencoder

**ml\_motif:** - Real-time, Image/CV

**type:** Benchmark

**ml\_task:** Unsupervised Learning

**notes:** Based on synthetic but realistic physics sensor data

**contact.name:** Ben Hawks, Nhan Tran

**contact.email:** unknown

**dataset.name:** Custom synthetic irregular sensor dataset

**dataset.url:** see GitHub repo

**results.name:** ChatGPT LLM

**fair.reproducible:** True

**fair.benchmark\_ready:** True

**ratings.software.rating:** 0

**ratings.software.reason:** Not analyzed.

**ratings.specification.rating:** 8.0

**ratings.specification.reason:** Classification is clearly defined for real-time inference on simulated LHC jets. Input features (HLFs) are documented, though exact latency or resource constraints are not numerically specified.

**ratings.dataset.rating:** 9.0

**ratings.dataset.reason:** Two datasets (OpenML and Zenodo) are public, well-formatted, and documented; FAIR principles are followed, though richer metadata would raise confidence to a 10.

**ratings.metrics.rating:** 9.0

**ratings.metrics.reason:** AUC and Accuracy are standard, quantitative, and well-aligned with goals of jet tagging and inference efficiency.

**ratings.reference\_solution.rating:** 8.0

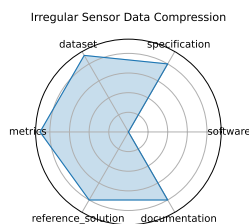
**ratings.reference\_solution.reason:** Float and quantized Keras/QKeras models are provided with results. Reproducibility is good, though full automation and documentation could be improved.

**ratings.documentation.rating:** 8.0

**ratings.documentation.reason:** GitHub contains baseline code, data loaders, and references, but setup for deployment (e.g., FPGA pipeline) requires familiarity with the tooling.

**id:** irregular\_sensor\_data\_compression

**Citations:** [2]



**Ratings:**

## 6 Beam Control

**date:** 2024-05-01

**last\_updated:** 2024-05

**expired:** unknown

**valid:** yes

**url:** <https://github.com/fastmachinelearning/fastml-science/tree/main/beam-control>

**domain:** Accelerators and Magnets

**focus:** Reinforcement learning control of accelerator beam position

**keywords:** - RL - beam stabilization - control systems - simulation

**task\_types:** - Control

**ai\_capability\_measured:** - Policy performance in simulated accelerator control

**metrics:** - Stability - Control loss

**models:** - DDPG - PPO (planned)

**ml\_motif:** - Real-time, RL

**type:** Benchmark

**ml\_task:** Reinforcement Learning

**notes:** Environment defined, baseline RL implementation is in progress

**contact.name:** Ben Hawks, Nhan Tran

**contact.email:** unknown

**results.name:** ChatGPT LLM

**fair.reproducible:** in progress

**fair.benchmark\_ready:** in progress

**ratings.software.rating:** 0

**ratings.software.reason:** Not analyzed.

**ratings.specification.rating:** 9.0

**ratings.specification.reason:** Task is well defined (real-time compression of sparse, irregular sensor data using autoencoders); latency constraints are implied but not fully quantified.

**ratings.dataset.rating:** 8.0

**ratings.dataset.reason:** Dataset is custom and synthetic but described well; FAIR-compliance is partial (reusable and accessible, but not externally versioned with rich metadata).

**ratings.metrics.rating:** 9.0

**ratings.metrics.reason:** Uses standard quantitative metrics (MSE, compression ratio) clearly aligned with compression and reconstruction goals.

**ratings.reference\_solution.rating:** 7.0

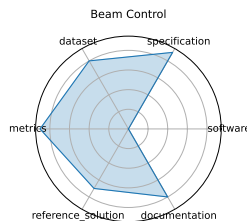
**ratings.reference\_solution.reason:** Baseline (autoencoder and quantized variant) is provided, but training/inference pipeline is minimally documented and needs user setup.

**ratings.documentation.rating:** 8.0

**ratings.documentation.reason:** GitHub repo contains core components, but more structured setup instructions and pre-trained weights would improve usability.

**id:** beam\_control

**Citations:** [2], [3]

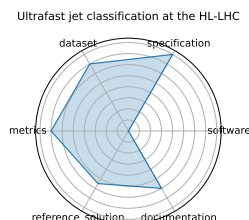


**Ratings:**



## 7 Ultrafast jet classification at the HL-LHC

**date:** 2024-07-08  
**last\_updated:** 2024-07  
**expired:** unknown  
**valid:** yes  
**url:** <https://arxiv.org/pdf/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  
**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  
**notes:** Uses quantization-aware training; hardware synthesis evaluated via hls4ml  
**contact.name:** Patrick Odagiu  
**contact.email:** unknown  
**dataset.name:** Zenodo DOI:10.5281/zenodo.3602260  
**dataset.url:** constituent-level jets  
**results.name:** ChatGPT LLM  
**results.url:** [https://docs.google.com/document/d/1gDf1CIYtfmfZ9urv1jCRZMYz\\_3WwEETkugUC65OZBdw](https://docs.google.com/document/d/1gDf1CIYtfmfZ9urv1jCRZMYz_3WwEETkugUC65OZBdw)  
**fair.reproducible:** True  
**fair.benchmark\_ready:** False  
**ratings.software.rating:** 0  
**ratings.software.reason:** Not analyzed.  
**ratings.specification.rating:** 8.0  
**ratings.specification.reason:** Task is clear (RL control of beam stability), with BOOSTR-based simulator; control objectives are well motivated, but system constraints and reward structure are still under refinement.  
**ratings.dataset.rating:** 7.0  
**ratings.dataset.reason:** BOOSTR dataset exists and is cited, but integration into the benchmark is in early stages; metadata and FAIR structure are limited.  
**ratings.metrics.rating:** 7.0  
**ratings.metrics.reason:** Stability and control loss are mentioned, but metrics are not yet formalized with clear definitions or baselines.  
**ratings.reference\_solution.rating:** 5.5  
**ratings.reference\_solution.reason:** DDPG baseline mentioned; PPO planned; implementation is still in progress with no reproducible results available yet.  
**ratings.documentation.rating:** 6.0  
**ratings.documentation.reason:** GitHub has a defined structure but is incomplete; setup and execution instructions for training/evaluation are not fully established.  
**id:** ultrafast\_jet\_classification\_at\_the\_hl-lhc  
**Citations:** [4]



**Ratings:**

## 8 Quench detection

**date:** 2024-10-15  
**last\_updated:** 2024-10  
**expired:** unknown  
**valid:** yes  
**url:** [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)  
**domain:** Accelerators and Magnets  
**focus:** Real-time detection of superconducting magnet quenches using ML  
**keywords:** - quench detection - autoencoder - anomaly detection - real-time  
**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  
**notes:** Precursor detection in progress; multi-modal and dynamic weighting methods  
**contact.name:** Maira Khan  
**contact.email:** unknown  
**dataset.name:** BPM and power supply data from BNL  
**dataset.url:** HDF5 preprocessed  
**results.name:** ChatGPT LLM  
**fair.reproducible:** in progress  
**fair.benchmark\_ready:** False  
**ratings.software.rating:** 0  
**ratings.software.reason:** Not analyzed.  
**ratings.specification.rating:** 10.0  
**ratings.specification.reason:** Real-time jet origin classification under FPGA constraints is clearly defined, with explicit latency targets (~100 ns) and I/O formats.  
**ratings.dataset.rating:** 9.0  
**ratings.dataset.reason:** Data available on Zenodo with DOI, includes constituent-level jets; accessible and well-documented, though not deeply versioned with full FAIR metadata.  
**ratings.metrics.rating:** 10.0  
**ratings.metrics.reason:** Accuracy, latency, and hardware resource usage (LUTs, DSPs) are rigorously measured and aligned with real-time goals.  
**ratings.reference\_solution.rating:** 9.0  
**ratings.reference\_solution.reason:** Includes models (MLP, Deep Sets, Interaction Networks) with quantization-aware training and synthesis results via hls4ml; reproducible but tightly coupled with specific toolchains.  
**ratings.documentation.rating:** 8.0  
**ratings.documentation.reason:** Paper and code (via hls4ml) are sufficient, but a centralized, standalone repo for reproducing all models would enhance accessibility.  
**id:** quench\_detection

## 9 DUNE

**date:** 2024-10-15

**last\_updated:** 2024-10

**expired:** unknown

**valid:** yes

**url:** [https://indico.fnal.gov/event/66520/contributions/301423/attachments/182439/250508/fast\\_ml\\_dunedaq\\_sonic\\_10\\_15\\_24.pdf](https://indico.fnal.gov/event/66520/contributions/301423/attachments/182439/250508/fast_ml_dunedaq_sonic_10_15_24.pdf)

**domain:** Particle Physics

**focus:** Real-time ML for DUNE DAQ time-series data

**keywords:** - DUNE - time-series - real-time - trigger

**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

**notes:** Prototype models demonstrated on SONIC platform

**contact.name:** Andrew J. Morgan

**contact.email:** unknown

**dataset.name:** DUNE SONIC data

**dataset.url:** via internal FNAL systems

**results.name:** ChatGPT LLM

**fair.reproducible:** in progress

**fair.benchmark\_ready:** False

**ratings.software.rating:** 0

**ratings.software.reason:** Not analyzed.

**ratings.specification.rating:** 8.0

**ratings.specification.reason:** Task (quench detection via anomaly detection) is clearly described; multi-modal sensors, streaming rates, and objective are provided, but constraints (latency thresholds) are qualitative.

**ratings.dataset.rating:** 7.0

**ratings.dataset.reason:** Custom dataset using real data from BNL; HDF5 formatted and structured, but access may be internal or limited, and not versioned for public FAIR use.

**ratings.metrics.rating:** 8.0

**ratings.metrics.reason:** ROC-AUC and detection latency are defined; relevant and quantitative but not yet paired with benchmark baselines.

**ratings.reference\_solution.rating:** 6.0

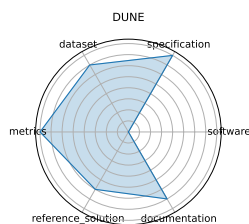
**ratings.reference\_solution.reason:** Autoencoder prototype exists; RL methods are in development; no fully reproducible pipeline is available yet.

**ratings.documentation.rating:** 7.0

**ratings.documentation.reason:** Slides and GDocs outline results; implementation is in progress with limited setup/code release.

**id:** dune

**Citations:** [5]



**Ratings:**

## 10 Intelligent experiments through real-time AI

**date:** 2025-01-08

**last\_updated:** 2025-01

**expired:** unknown

**valid:** yes

**url:** <https://arxiv.org/pdf/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

**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 ( $\mu$ s) - 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

**notes:** Achieved  $\sim 97.4\%$  accuracy for beauty decay triggers; sub-10  $\mu$ s latency on Alveo U280; hit-based FPGA design via hls4ml and FlowGNN.

**contact.name:** Jakub Kvapil (lanl.gov)

**contact.email:** unknown

**dataset.name:** Internal simulated tracking data

**dataset.url:** sPHENIX and EIC DIS-electron tagger

**results.name:** ChatGPT LLM

**fair.reproducible:** True

**fair.benchmark\_ready:** False

**ratings.software.rating:** 0

**ratings.software.reason:** Not analyzed.

**ratings.specification.rating:** 8.0

**ratings.specification.reason:** Task (trigger-level anomaly detection) is clearly defined for low-latency streaming input, but the problem framing lacks complete architectural/system specs.

**ratings.dataset.rating:** 6.0

**ratings.dataset.reason:** Internal DUNE SONIC data; not publicly released and no formal FAIR support; replicability is institutionally gated.

**ratings.metrics.rating:** 7.0

**ratings.metrics.reason:** Metrics include detection efficiency and latency, which are relevant, but only lightly supported by baselines or formal eval scripts.

**ratings.reference\_solution.rating:** 5.0

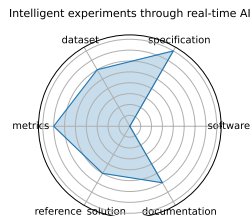
**ratings.reference\_solution.reason:** One CNN prototype demonstrated; LSTM planned. No public implementation or ready-to-run example yet.

**ratings.documentation.rating:** 6.0

**ratings.documentation.reason:** Slides and some internal documentation exist, but no full pipeline or public GitHub repo yet.

**id:** intelligent\_experiments\_through\_real-time\_ai

**Citations:** [6]



**Ratings:**

# 11 Neural Architecture Codesign for Fast Physics Applications

**date:** 2025-01-09

**last\_updated:** 2025-01

**expired:** unknown

**valid:** yes

**url:** <https://arxiv.org/abs/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

**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

**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.name:** ChatGPT LLM

**fair.reproducible:** Yes (nac-opt, hls4ml)

**fair.benchmark\_ready:** False

**ratings.software.rating:** 0

**ratings.software.reason:** Not analyzed.

**ratings.specification.rating:** 10.0

**ratings.specification.reason:** Task is clearly defined (triggering on rare events with sub-10  $\mu$ s latency); architecture, constraints, and system context (FPGA, Alveo) are well detailed.

**ratings.dataset.rating:** 7.0

**ratings.dataset.reason:** Simulated tracking data from sPHENIX and EIC; internally structured but not yet released in a public FAIR-compliant format.

**ratings.metrics.rating:** 10.0

**ratings.metrics.reason:** Accuracy, latency, and hardware resource utilization (LUTs, DSPs) are clearly defined and used in evaluation.

**ratings.reference\_solution.rating:** 9.0

**ratings.reference\_solution.reason:** Graph-based models (BGN-ST, GarNet) are implemented and tested on real hardware; reproducibility possible with hls4ml but full scripts not bundled.

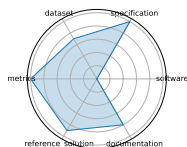
**ratings.documentation.rating:** 8.0

**ratings.documentation.reason:** Paper is detailed and tool usage (FlowGNN, hls4ml) is described, but repo release and dataset access remain in progress.

**id:** neural\_architecture\_codesign\_for\_fast\_physics\_applications

**Citations:** [7]

Neural Architecture Codesign for Fast Physics Applications



**Ratings:**

## 12 Smart Pixels for LHC

**date:** 2024-06-24

**last\_updated:** 2024-06

**expired:** unknown

**valid:** yes

**url:** <https://arxiv.org/abs/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

**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

**notes:** Prototype in CMOS 28 nm; proof-of-concept for Phase III pixel upgrades.

**contact.name:** Lindsey Gray; Jennet Dickinson

**contact.email:** unknown

**results.name:** ChatGPT LLM

**fair.reproducible:** True

**fair.benchmark\_ready:** Yes (Zenodo:7331128)

**ratings.software.rating:** 0

**ratings.software.reason:** Not analyzed.

**ratings.specification.rating:** 9.0

**ratings.specification.reason:** Task (automated neural architecture search for real-time physics) is well formulated with clear latency, model compression, and deployment goals.

**ratings.dataset.rating:** 6.0

**ratings.dataset.reason:** Internal Bragg and jet datasets used; not publicly hosted or FAIR-compliant, though mentioned in the paper.

**ratings.metrics.rating:** 10.0

**ratings.metrics.reason:** BOP reduction, latency, and accuracy are all quantitatively evaluated.

**ratings.reference\_solution.rating:** 8.0

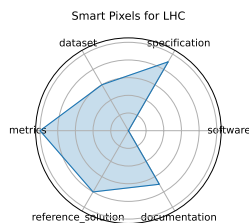
**ratings.reference\_solution.reason:** NAC-generated models for Bragg peak and jet classification are described, but pipeline requires integration of several tools and is not fully packaged.

**ratings.documentation.rating:** 7.0

**ratings.documentation.reason:** NAC pipeline, hls4ml usage, and results are discussed; code (e.g., nac-opt) referenced, but replication requires stitching together toolchain and data.

**id:** smart\_pixels\_for\_lhc

**Citations:** [8]



**Ratings:**

## 13 HEDM (BraggNN)

**date:** 2023-10-03

**last\_updated:** 2023-10

**expired:** unknown

**valid:** yes

**url:** <https://arxiv.org/abs/2008.08198>

**domain:** Material Science

**focus:** Fast Bragg peak analysis using deep learning in diffraction microscopy

**keywords:** - BraggNN - diffraction - peak finding - HEDM

**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

**notes:** Enables real-time HEDM workflows; basis for NAC case study.

**contact.name:** Jason Weitz (UCSD)

**contact.email:** unknown

**results.name:** ChatGPT LLM

**fair.reproducible:** True

**fair.benchmark\_ready:** True

**ratings.software.rating:** 0

**ratings.software.reason:** Not analyzed.

**ratings.specification.rating:** 10.0

**ratings.specification.reason:** Fully specified: describes task (data filtering/classification, system design (on-sensor inference), latency (25 ns), and power constraints.

**ratings.dataset.rating:** 8.0

**ratings.dataset.reason:** In-pixel charge cluster data used, but dataset release info is minimal; FAIR metadata/versioning limited.

**ratings.metrics.rating:** 9.0

**ratings.metrics.reason:** Data rejection rate and power per pixel are clearly defined and directly tied to hardware goals.

**ratings.reference\_solution.rating:** 9.0

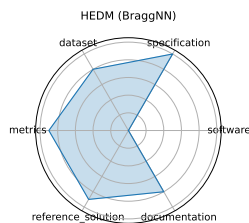
**ratings.reference\_solution.reason:** 2-layer NN implementation is evaluated in hardware; reproducible via hls4ml flow with results in paper.

**ratings.documentation.rating:** 8.0

**ratings.documentation.reason:** Paper is clear; Zenodo asset is referenced, but additional GitHub or setup repo would improve reproducibility.

**id:** hedm\_braggnn

**Citations:** [9]



**Ratings:**

## 14 4D-STEM

**date:** 2023-12-03

**last\_updated:** 2023-12

**expired:** unknown

**valid:** yes

**url:** <https://openreview.net/pdf?id=7yt3N0o0W9>

**domain:** Material Science

**focus:** Real-time ML for scanning transmission electron microscopy

**keywords:** - 4D-STEM - electron microscopy - real-time - image processing

**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

**notes:** In-progress; model design under development.

**contact.name:** —

**contact.email:** unknown

**results.name:** ChatGPT LLM

**fair.reproducible:** in progress

**fair.benchmark\_ready:** False

**ratings.software.rating:** 0

**ratings.software.reason:** Not analyzed.

**ratings.specification.rating:** 9.0

**ratings.specification.reason:** Peak localization task is well-defined for diffraction images; input/output described clearly, but no system constraints.

**ratings.dataset.rating:** 8.0

**ratings.dataset.reason:** Simulated diffraction images provided; reusable and downloadable, but not externally versioned or FAIR-structured.

**ratings.metrics.rating:** 9.0

**ratings.metrics.reason:** Inference speed and localization accuracy are standard and quantitatively reported.

**ratings.reference\_solution.rating:** 8.0

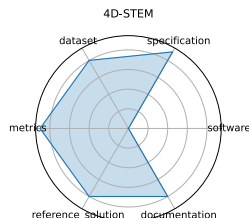
**ratings.reference\_solution.reason:** BraggNN model and training pipeline exist, but need stitching from separate repositories.

**ratings.documentation.rating:** 8.0

**ratings.documentation.reason:** Paper and codebase are available and usable, though not fully turnkey.

**id:** dstem

**Citations:** [10]



**Ratings:**



## 15 In-Situ High-Speed Computer Vision

**date:** 2023-12-05

**last\_updated:** 2023-12

**expired:** unknown

**valid:** yes

**url:** <https://arxiv.org/abs/2312.00128>

**domain:** Fusion/Plasma

**focus:** Real-time image classification for in-situ plasma diagnostics

**keywords:** - plasma - in-situ vision - real-time ML

**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

**notes:** Embedded/deployment details in progress.

**contact.name:** —

**contact.email:** unknown

**results.name:** ChatGPT LLM

**results.url:** [https://docs.google.com/document/d/1EqkRHuQslyQqMvZs\\_L6p9JAY2vKX5OCTubzttFBuRoQ/edit?usp=sharing](https://docs.google.com/document/d/1EqkRHuQslyQqMvZs_L6p9JAY2vKX5OCTubzttFBuRoQ/edit?usp=sharing)

**fair.reproducible:** in progress

**fair.benchmark\_ready:** False

**ratings.software.rating:** 0

**ratings.software.reason:** Not analyzed.

**ratings.specification.rating:** 7.0

**ratings.specification.reason:** General task defined (real-time microscopy inference), but no standardized I/O format, latency constraint, or complete problem framing yet.

**ratings.dataset.rating:** 0.0

**ratings.dataset.reason:** Dataset not provided or described in any formal way.

**ratings.metrics.rating:** 6.0

**ratings.metrics.reason:** Mentions throughput and accuracy, but metrics are not formally defined or benchmarked.

**ratings.reference\_solution.rating:** 2.0

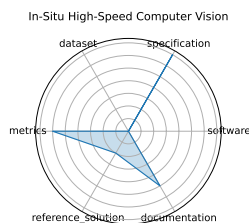
**ratings.reference\_solution.reason:** Prototype CNNs described; no baseline or implementation released.

**ratings.documentation.rating:** 5.0

**ratings.documentation.reason:** OpenReview paper and Gemini doc give some insight, but no working code, environment, or example.

**id:** in-situ\_high-speed\_computer\_vision

**Citations:** [11]



**Ratings:**

## 16 BenchCouncil AIBench

**date:** 2020-01-01

**last\_updated:** 2020-01

**expired:** unknown

**valid:** yes

**url:** <https://www.benchcouncil.org/AIBench/>

**domain:** General

**focus:** End-to-end AI benchmarking across micro, component, and application levels

**keywords:** - benchmarking - AI systems - application-level evaluation

**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

**notes:** Covers scenario-distilling, micro, component, and end-to-end benchmarks.

**contact.name:** Wanling Gao (BenchCouncil)

**contact.email:** unknown

**results.name:** ChatGPT LLM

**results.url:** unknown

**fair.reproducible:** Yes

**fair.benchmark\_ready:** Yes

**ratings.software.rating:** 0

**ratings.software.reason:** Not analyzed.

**ratings.specification.rating:** 8.0

**ratings.specification.reason:** Task (plasma diagnostic classification) and real-time deployment described; system specs (FPS targets) implied but not fully quantified.

**ratings.dataset.rating:** 6.0

**ratings.dataset.reason:** Dataset is sensor stream-based but not shared or FAIR-documented.

**ratings.metrics.rating:** 8.0

**ratings.metrics.reason:** FPS and classification accuracy reported and relevant.

**ratings.reference\_solution.rating:** 7.0

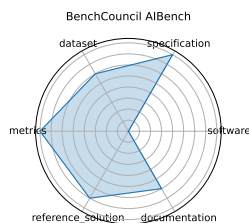
**ratings.reference\_solution.reason:** CNN model described and evaluated, but public implementation and benchmarks are not available yet.

**ratings.documentation.rating:** 6.0

**ratings.documentation.reason:** Paper and Gemini doc exist, but full setup instructions and tools are still in progress.

**id:** benchcouncil\_aibench

**Citations:** [12]



**Ratings:**

## 17 BenchCouncil BigDataBench

**date:** 2020-01-01

**last\_updated:** 2020-01

**expired:** unknown

**valid:** yes

**url:** <https://www.benchcouncil.org/BigDataBench/>

**domain:** General

**focus:** Big data and AI benchmarking across structured, semi-structured, and unstructured data workloads

**keywords:** - big data - AI benchmarking - data analytics

**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

**notes:** Built on eight data motifs; provides Hadoop, Spark, Flink, MPI implementations.

**contact.name:** Jianfeng Zhan (BenchCouncil)

**contact.email:** unknown

**results.name:** ChatGPT LLM

**results.url:** <https://docs.google.com/document/d/1VFRxhR2G5A83S8PqKBrP99LLVgcCGvX2WW4vTtwxmQ4/edit?usp=sharing>

**fair.reproducible:** Yes

**fair.benchmark\_ready:** Yes

**ratings.software.rating:** 0

**ratings.software.reason:** Not analyzed.

**ratings.specification.rating:** 9.0

**ratings.specification.reason:** Evaluates AI at multiple levels (micro to end-to-end); tasks and workloads are clearly defined, though specific I/O formats and constraints vary.

**ratings.dataset.rating:** 9.0

**ratings.dataset.reason:** Realistic datasets across diverse domains; FAIR structure for many components, but individual datasets may not all be versioned or richly annotated.

**ratings.metrics.rating:** 9.0

**ratings.metrics.reason:** Latency, throughput, and accuracy clearly defined for end-to-end tasks; consistent across models and setups.

**ratings.reference\_solution.rating:** 8.0

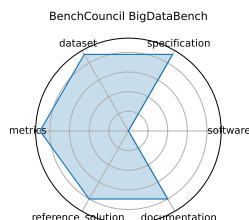
**ratings.reference\_solution.reason:** Reference implementations for several tasks exist, but setup across all tasks is complex and not fully streamlined.

**ratings.documentation.rating:** 8.0

**ratings.documentation.reason:** Central documentation exists, with detailed component breakdowns; environment setup across platforms (e.g., hardware variations) can require manual adjustment.

**id:** benchcouncil\_bigdatabench

**Citations:** [13]



**Ratings:**

## 18 MLPerf HPC

**date:** 2021-10-20

**last\_updated:** 2021-10

**expired:** unknown

**valid:** yes

**url:** <https://github.com/mlcommons/hpc>

**domain:** Cosmology, Climate, Protein Structure, Catalysis

**focus:** Scientific ML training and inference on HPC systems

**keywords:** - HPC - training - inference - scientific ML

**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

**notes:** Shared framework with MLCommons Science; reference implementations included.

**contact.name:** Steven Farrell (MLCommons)

**contact.email:** unknown

**results.name:** ChatGPT LLM

**results.url:** unknown

**fair.reproducible:** Yes

**fair.benchmark\_ready:** Yes

**ratings.software.rating:** 0

**ratings.software.reason:** Not analyzed.

**ratings.specification.rating:** 9.0

**ratings.specification.reason:** Focused on structured/unstructured data pipelines; clearly defined tasks spanning analytics to AI; some scenarios lack hardware constraint modeling.

**ratings.dataset.rating:** 9.0

**ratings.dataset.reason:** Built from 13 real-world sources; structured for realistic big data scenarios; partially FAIR-compliant with documented data motifs.

**ratings.metrics.rating:** 9.0

**ratings.metrics.reason:** Covers data throughput, latency, and accuracy; quantitative and benchmark-ready.

**ratings.reference\_solution.rating:** 8.0

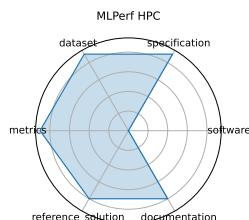
**ratings.reference\_solution.reason:** Many pipeline and model examples provided using Hadoop/Spark/Flink; setup effort varies by task and platform.

**ratings.documentation.rating:** 8.0

**ratings.documentation.reason:** Strong documentation with examples and task specifications; centralized support exists, but task-specific tuning may require domain expertise.

**id:** mlperf\_hpc

**Citations:** [14]



**Ratings:**

## 19 MLCommons Science

**date:** 2023-06-01

**last\_updated:** 2023-06

**expired:** unknown

**valid:** yes

**url:** <https://github.com/mlcommons/science>

**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

**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

**notes:** Joint national-lab effort under Apache-2.0 license.

**contact.name:** MLCommons Science Working Group

**contact.email:** unknown

**results.name:** ChatGPT LLM

**results.url:** unknown

**fair.reproducible:** Yes

**fair.benchmark\_ready:** Yes

**ratings.software.rating:** 0

**ratings.software.reason:** Not analyzed.

**ratings.specification.rating:** 10.0

**ratings.specification.reason:** Scientific ML tasks (e.g., CosmoFlow, DeepCAM) are clearly defined with HPC system-level constraints and targets.

**ratings.dataset.rating:** 9.0

**ratings.dataset.reason:** Public scientific datasets (e.g., cosmology, weather); used consistently, though FAIR-compliance of individual datasets varies slightly.

**ratings.metrics.rating:** 10.0

**ratings.metrics.reason:** Training time, GPU utilization, and accuracy are all directly measured and benchmarked across HPC systems.

**ratings.reference\_solution.rating:** 9.0

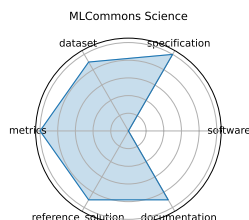
**ratings.reference\_solution.reason:** Reference implementations available and actively maintained; HPC setup may require domain-specific environment.

**ratings.documentation.rating:** 9.0

**ratings.documentation.reason:** GitHub repo and papers provide detailed instructions; reproducibility supported across multiple institutions.

**id:** mlcommons\_science

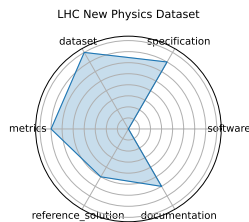
**Citations:** [15]



**Ratings:**

## 20 LHC New Physics Dataset

**date:** 2021-07-05  
**last\_updated:** 2021-07  
**expired:** unknown  
**valid:** yes  
**url:** <https://arxiv.org/pdf/2107.02157>  
**domain:** Particle Physics; Real-time Triggering  
**focus:** Real-time LHC event filtering for anomaly detection using proton collision data  
**keywords:** - anomaly detection - proton collision - real-time inference - event filtering - unsupervised ML  
**task\_types:** - Anomaly detection - Event classification  
**ai\_capability\_measured:** - Unsupervised signal detection under latency and bandwidth constraints  
**metrics:** - ROC-AUC - Detection efficiency  
**models:** - Autoencoder - Variational autoencoder - Isolation forest  
**ml\_motif:** - Multiple  
**type:** Framework  
**ml\_task:** NA  
**notes:** Includes electron/muon-filtered background and black-box signal benchmarks; 1M events per black box.  
**contact.name:** Ema Puljak (ema.puljak@cern.ch)  
**contact.email:** unknown  
**dataset.name:** Zenodo stores: background + 3 black-box signal sets  
**dataset.url:** 1M events each  
**results.name:** ChatGPT LLM  
**results.url:** unknown  
**fair.reproducible:** Yes  
**fair.benchmark\_ready:** Yes  
**ratings.software.rating:** 0  
**ratings.software.reason:** Not analysed.  
**ratings.specification.rating:** 7.0  
**ratings.specification.reason:** The problem (anomaly detection for new physics at LHC) is clearly described with goals and background, but lacks a formal task specification or constraints.  
**ratings.dataset.rating:** 8.0  
**ratings.dataset.reason:** Large-scale, public dataset derived from LHC simulations; well-documented and available via Zenodo.  
**ratings.metrics.rating:** 7.0  
**ratings.metrics.reason:** Provides AUROC, accuracy, and anomaly detection metrics but lacks standardized evaluation script.  
**ratings.reference\_solution.rating:** 5.0  
**ratings.reference\_solution.reason:** Baseline models (autoencoders, GANs) are described in associated papers, but implementations vary across papers.  
**ratings.documentation.rating:** 6.0  
**ratings.documentation.reason:** Publicly available papers and datasets with descriptions, but no unified README or training setup.  
**id:** lhc\_new\_physics\_dataset  
**Citations:** [16]



**Ratings:**

## 21 MLCommons Medical AI

**date:** 2023-07-17

**last\_updated:** 2023-07

**expired:** unknown

**valid:** yes

**url:** <https://github.com/mlcommons/medical>

**domain:** Healthcare; Medical AI

**focus:** Federated benchmarking and evaluation of medical AI models across diverse real-world clinical data

**keywords:** - medical AI - federated evaluation - privacy-preserving - fairness - healthcare benchmarks

**task\_types:** - Federated evaluation - Model validation

**ai\_capability\_measured:** - Clinical accuracy - fairness - generalizability - privacy compliance

**metrics:** - ROC AUC - Accuracy - Fairness metrics

**models:** - MedPerf-validated CNNs - GaNDLF workflows

**ml\_motif:** - Multiple

**type:** Platform

**ml\_task:** NA

**notes:** Open-source platform under Apache-2.0; used across 20+ institutions and hospitals :contentReference[oaicite:2]{index=2}.

**contact.name:** Alex Karargyris (MLCommons Medical AI)

**contact.email:** unknown

**dataset.name:** Multi-institutional clinical datasets

**dataset.url:** radiology

**results.name:** ChatGPT LLM

**results.url:** unknown

**fair.reproducible:** Yes

**fair.benchmark\_ready:** Yes

**ratings.software.rating:** 0

**ratings.software.reason:** Not analyzed.

**ratings.specification.rating:** 9.0

**ratings.specification.reason:** Diverse scientific tasks (earthquake, CFD, microscopy) with detailed problem statements and goals; system constraints not uniformly applied.

**ratings.dataset.rating:** 9.0

**ratings.dataset.reason:** Domain-specific datasets (e.g., microscopy, climate); mostly public and structured, but FAIR annotations are not always explicit.

**ratings.metrics.rating:** 9.0

**ratings.metrics.reason:** Task-specific metrics (MAE, speedup, accuracy) are clear and reproducible.

**ratings.reference\_solution.rating:** 9.0

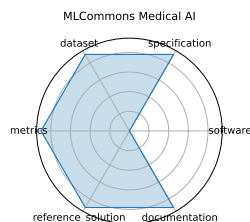
**ratings.reference\_solution.reason:** Reference models (CNN, GNN, Transformer) provided with training/evaluation pipelines.

**ratings.documentation.rating:** 9.0

**ratings.documentation.reason:** Well-documented, open-sourced, and maintained with examples; strong community support and reproducibility focus.

**id:** mlcommons\_medical\_ai

**Citations:** [17]

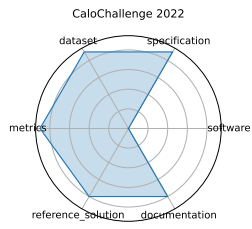


**Ratings:**

## 22 CaloChallenge 2022

**date:** 2024-10-28  
**last\_updated:** 2024-10  
**expired:** unknown  
**valid:** yes  
**url:** <http://arxiv.org/abs/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  
**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  
**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  
**dataset.name:** Four LHC calorimeter shower datasets  
**dataset.url:** various voxel resolutions  
**results.name:** ChatGPT LLM  
**results.url:** unknown  
**fair.reproducible:** Yes  
**fair.benchmark\_ready:** Yes  
**ratings.software.rating:** 0  
**ratings.software.reason:** Not analyzed.  
**ratings.specification.rating:** 9.0  
**ratings.specification.reason:** Task is clearly defined: real-time anomaly detection from high-rate LHC collisions. Latency and bandwidth constraints are mentioned, though not numerically enforced.  
**ratings.dataset.rating:** 9.0  
**ratings.dataset.reason:** Publicly available via Zenodo, with structured signal/background splits, and rich metadata; nearly fully FAIR.  
**ratings.metrics.rating:** 9.0  
**ratings.metrics.reason:** ROC-AUC and detection efficiency are clearly defined and appropriate for unsupervised anomaly detection.  
**ratings.reference\_solution.rating:** 8.0  
**ratings.reference\_solution.reason:** Several baseline methods (autoencoder, VAE, isolation forest) are evaluated; runnable versions available via community repos but not tightly bundled.  
**ratings.documentation.rating:** 8.0  
**ratings.documentation.reason:** Paper and data documentation are clear, and the dataset is widely reused. Setup requires some manual effort to reproduce full pipelines.  
**id:** calochallenge\_  
**Citations:** [18]

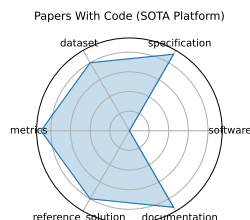




**Ratings:**

## 23 Papers With Code (SOTA Platform)

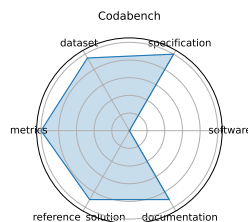
**date:** ongoing  
**last\_updated:** 2025-06  
**expired:** unknown  
**valid:** yes  
**url:** <https://paperswithcode.com/sota>  
**domain:** General ML; All domains  
**focus:** Open platform tracking state-of-the-art results, benchmarks, and implementations across ML tasks and papers  
**keywords:** - leaderboard - benchmarking - reproducibility - open-source  
**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  
**notes:** Community-driven open platform; automatic data extraction and versioning.  
**contact.name:** Papers With Code Team  
**contact.email:** unknown  
**results.name:** ChatGPT LLM  
**results.url:** unknown  
**fair.reproducible:** Yes  
**fair.benchmark\_ready:** Yes  
**ratings.software.rating:** 0  
**ratings.software.reason:** Not analyzed.  
**ratings.specification.rating:** 9.0  
**ratings.specification.reason:** Evaluation setting (federated clinical benchmarking) is well-defined; I/O interfaces vary slightly by task but are standardized in MedPerf platform.  
**ratings.dataset.rating:** 8.0  
**ratings.dataset.reason:** Uses distributed, real-world clinical datasets across institutions; FAIR compliance varies across hospitals and data hosts.  
**ratings.metrics.rating:** 9.0  
**ratings.metrics.reason:** ROC AUC, accuracy, and fairness metrics are explicitly defined and task-dependent; consistently tracked across institutions.  
**ratings.reference\_solution.rating:** 8.0  
**ratings.reference\_solution.reason:** Validated CNNs and GaNDLF pipelines are used and shared via the MedPerf tool, but some implementations are abstracted behind the platform.  
**ratings.documentation.rating:** 9.0  
**ratings.documentation.reason:** Excellent documentation across MedPerf, GaNDLF, and COFE; reproducibility handled via containerized flows and task templates.  
**id:** papers\_with\_code\_sota\_platform  
**Citations:** [19]



**Ratings:**

## 24 Codabench

**date:** 2022-01-01  
**last\_updated:** 2025-03  
**expired:** unknown  
**valid:** yes  
**url:** <https://www.codabench.org/>  
**domain:** General ML; Multiple  
**focus:** Open-source platform for organizing reproducible AI benchmarks and competitions  
**keywords:** - benchmark platform - code submission - competitions - meta-benchmark  
**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  
**notes:** Hosts 51 public competitions, ~26 k users, 177 k submissions :contentReference[oaicite:2]{index=2}  
**contact.name:** Isabelle Guyon (Université Paris-Saclay)  
**contact.email:** unknown  
**results.name:** ChatGPT LLM  
**results.url:** unknown  
**fair.reproducible:** Yes  
**fair.benchmark\_ready:** Yes  
**ratings.software.rating:** 0  
**ratings.software.reason:** Not analyzed.  
**ratings.specification.rating:** 10.0  
**ratings.specification.reason:** Simulation task (generative calorimeter showers) is clearly stated with multiple datasets, fidelity requirements, and performance constraints.  
**ratings.dataset.rating:** 9.5  
**ratings.dataset.reason:** Public datasets available in multiple sizes and formats; well-documented; not versioned  
**ratings.metrics.rating:** 10.0  
**ratings.metrics.reason:** Histogram similarity, classifier AUC, and generation latency are clearly defined and benchmarked across all submissions.  
**ratings.reference\_solution.rating:** 9.0  
**ratings.reference\_solution.reason:** 31 model implementations submitted; some made public and reproducible, though others remain undocumented or private.  
**ratings.documentation.rating:** 9.0  
**ratings.documentation.reason:** Paper, leaderboard, and Gemini doc are comprehensive; unified repo or launchable baseline kit would push this to a 10.  
**id:** codabench  
**Citations:** [20]



**Ratings:**

## 25 Sabath (SBI-FAIR)

**date:** 2021-09-27

**last\_updated:** 2023-07

**expired:** unknown

**valid:** yes

**url:** <https://sbi-fair.github.io/docs/software/sabath/>

**domain:** Systems; Metadata

**focus:** FAIR metadata framework for ML-driven surrogate workflows in HPC systems

**keywords:** - meta-benchmark - metadata - HPC - surrogate modeling

**task\_types:** - Systems benchmarking

**ai\_capability\_measured:** - Metadata tracking - reproducible HPC workflows

**metrics:** - Metadata completeness - FAIR compliance

**models:** - N/A

**ml\_motif:** - Systems

**type:** Platform

**ml\_task:** NA

**notes:** Developed by PI Piotr Luszczek at UTK; integrates with MiniWeatherML, AutoPhaseNN, Cosmoflow, etc. :contentReference[oaicite:4]{index=4}

**contact.name:** Piotr Luszczek (luszczek@utk.edu)

**contact.email:** unknown

**results.name:** ChatGPT LLM

**results.url:** unknown

**fair.reproducible:** Yes

**fair.benchmark\_ready:** N/A

**ratings.software.rating:** 0

**ratings.software.reason:** Not analyzed.

**ratings.specification.rating:** 8.0

**ratings.specification.reason:** The benchmark defines simulation-based inference (SBI) tasks clearly with FAIR principles applied to particle physics datasets.

**ratings.dataset.rating:** 8.0

**ratings.dataset.reason:** Data is well-structured for SBI and publicly available with clear licensing.

**ratings.metrics.rating:** 8.0

**ratings.metrics.reason:** Includes likelihood and posterior accuracy; metrics well-matched to SBI.

**ratings.reference\_solution.rating:** 7.0

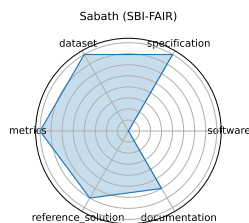
**ratings.reference\_solution.reason:** Baseline SBI models are implemented and reproducible.

**ratings.documentation.rating:** 6.0

**ratings.documentation.reason:** GitHub repo includes code and instructions, but lacks full tutorials or walkthroughs.

**id:** sabath\_sbi-fair

**Citations:** [21]



**Ratings:**

## 26 PDEBench

**date:** 2022-10-13

**last\_updated:** 2025-05

**expired:** unknown

**valid:** yes

**url:** <https://github.com/pdebench/PDEBench>

**domain:** CFD; Weather Modeling

**focus:** Benchmark suite for ML-based surrogates solving time-dependent PDEs

**keywords:** - PDEs - CFD - scientific ML - surrogate modeling - NeurIPS

**task\_types:** - Supervised Learning

**ai\_capability\_measured:** - Time-dependent PDE modeling; physical accuracy

**metrics:** - RMSE - boundary RMSE - Fourier RMSE

**models:** - FNO - U-Net - PINN - Gradient-Based inverse methods

**ml\_motif:** - Multiple

**type:** Framework

**ml\_task:** Supervised Learning

**notes:** Datasets hosted on DaRUS (DOI:10.18419/darus-2986); contact maintainers by email :contentReference[oaicite:6]{index=6}

**contact.name:** Makoto Takamoto (makoto.takamoto@neclab.eu)

**contact.email:** unknown

**results.name:** ChatGPT LLM

**results.url:** unknown

**fair.reproducible:** Yes

**fair.benchmark\_ready:** Yes

**ratings.software.rating:** 0

**ratings.software.reason:** Not analyzed.

**ratings.specification.rating:** 9.0

**ratings.specification.reason:** Clearly defined PDE-solving tasks with well-specified constraints and solution formats.

**ratings.dataset.rating:** 9.0

**ratings.dataset.reason:** Includes synthetic and real-world PDE datasets with detailed format descriptions.

**ratings.metrics.rating:** 8.0

**ratings.metrics.reason:** Uses L2 error and other norms relevant to PDE solutions.

**ratings.reference\_solution.rating:** 7.0

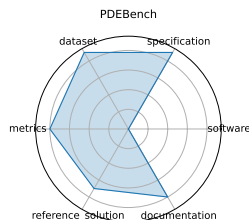
**ratings.reference\_solution.reason:** Includes baseline solvers and trained models across multiple PDE tasks.

**ratings.documentation.rating:** 8.0

**ratings.documentation.reason:** Well-organized GitHub with examples, dataset loading scripts, and training configs.

**id:** pdebench

**Citations:** [22]



**Ratings:**

## 27 The Well

**date:** 2024-12-03

**last\_updated:** 2025-06

**expired:** unknown

**valid:** yes

**url:** [https://polymathic-ai.org/the\\_well/](https://polymathic-ai.org/the_well/)

**domain:** biological systems, fluid dynamics, acoustic scattering, astrophysical MHD

**focus:** Foundation model + surrogate dataset spanning 16 physical simulation domains

**keywords:** - surrogate modeling - foundation model - physics simulations - spatiotemporal dynamics

**task\_types:** - Supervised Learning

**ai\_capability\_measured:** - Surrogate modeling - physics-based prediction

**metrics:** - Dataset size - Domain breadth

**models:** - FNO baselines - U-Net baselines

**ml\_motif:** - Foundation model, Surrogate

**type:** Dataset

**ml\_task:** Supervised Learning

**notes:** Includes unified API and dataset metadata; see 2025 NeurIPS paper for full benchmark details. Size: 15 TB. :contentReference[oaicite:2]{index=2}

**contact.name:** Wes Brewer

**contact.email:** unknown

**dataset.name:** 16 simulation datasets

**dataset.url:** HDF5) via PyPI/GitHub

**results.name:** ChatGPT LLM

**results.url:** unknown

**fair.reproducible:** Yes

**fair.benchmark\_ready:** Yes

**ratings.software.rating:** 0

**ratings.software.reason:** Not analyzed.

**ratings.specification.rating:** 7.0

**ratings.specification.reason:** Explores LLM understanding of mental health scenarios; framing is creative but loosely defined.

**ratings.dataset.rating:** 6.0

**ratings.dataset.reason:** Dataset is described in concept but not released; privacy limits public access though synthetic proxies are referenced.

**ratings.metrics.rating:** 7.0

**ratings.metrics.reason:** Uses manual annotation and quality scores, but lacks standardized automatic metrics.

**ratings.reference\_solution.rating:** 6.0

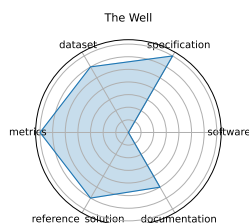
**ratings.reference\_solution.reason:** Provides few-shot prompt examples and human rating calibration details.

**ratings.documentation.rating:** 5.0

**ratings.documentation.reason:** Paper gives use cases, but code and data are not yet public.

**id:** the\_well

**Citations:** [23]



**Ratings:**

## 28 LLM-Inference-Bench

**date:** 2024-10-31

**last\_updated:** 2024-11

**expired:** unknown

**valid:** yes

**url:** <https://github.com/argonne-lcf/LLM-Inference-Bench>

**domain:** LLM; HPC/inference

**focus:** Hardware performance benchmarking of LLMs on AI accelerators

**keywords:** - LLM - inference benchmarking - GPU - accelerator - throughput

**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

**notes:** Licensed under BSD-3, maintained by Argonne; supports GPUs and accelerators. :contentReference[oaicite:4]{index=4}

**contact.name:** Krishna Teja Chitty-Venkata (Argonne LCF)

**contact.email:** unknown

**results.name:** ChatGPT LLM

**results.url:** unknown

**fair.reproducible:** Yes

**fair.benchmark\_ready:** Yes

**ratings.software.rating:** 0

**ratings.software.reason:** Not analyzed.

**ratings.specification.rating:** 9.0

**ratings.specification.reason:** PDE tasks (forward/inverse) and I/O structures are clearly specified with detailed PDE context and constraints.

**ratings.dataset.rating:** 10.0

**ratings.dataset.reason:** Hosted via DaRUS with a DOI, well-documented, versioned, and FAIR-compliant.

**ratings.metrics.rating:** 9.0

**ratings.metrics.reason:** Uses RMSE variants and Fourier-based errors.

**ratings.reference\_solution.rating:** 10.0

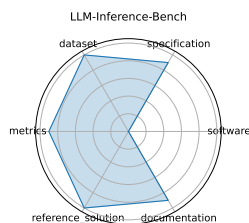
**ratings.reference\_solution.reason:** Baselines (FNO, U-Net, PINN) implemented and ready-to-run; strong community adoption.

**ratings.documentation.rating:** 9.0

**ratings.documentation.reason:** Clean GitHub with usage, dataset links, and tutorial notebooks.

**id:** llm-inference-bench

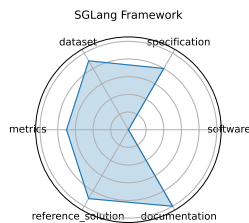
**Citations:** [24]



**Ratings:**

## 29 SGLang Framework

**date:** 2023-12-12  
**last\_updated:** 2025-06  
**expired:** unkown  
**valid:** yes  
**url:** <https://github.com/sgl-project/sglang/tree/main/benchmark>  
**domain:** LLM Vision  
**focus:** Fast serving framework for LLMs and vision-language models  
**keywords:** - LLM serving - vision-language - RadixAttention - performance - JSON decoding  
**task\_types:** - Model serving framework  
**ai\_capability\_measured:** - Serving throughput - JSON/task-specific latency  
**metrics:** - Tokens/sec - Time-to-first-token - Throughput gain vs baseline  
**models:** - LLaVA - DeepSeek - Llama  
**ml\_motif:** - LLM Vision  
**type:** Framework  
**ml\_task:** Model serving  
**notes:** Deployed in production (xAI, NVIDIA, Google Cloud); v0.4.8 release June 2025. :contentReference[oaicite:6]{index=6}  
**contact.name:** SGLang Team  
**contact.email:** unkown  
**dataset.name:** Benchmark configs  
**dataset.url:** dummy or real  
**results.name:** ChatGPT LLM  
**results.url:** unkown  
**fair.reproducible:** Yes  
**fair.benchmark\_ready:** Yes  
**ratings.software.rating:** 0  
**ratings.software.reason:** Not analyzed.  
**ratings.specification.rating:** 8.0  
**ratings.specification.reason:** Clearly framed around surrogate learning across 16 domains, but not all tasks are formally posed or constrained in a unified benchmark protocol. Paper mentions performance on NVIDIA H100.  
**ratings.dataset.rating:** 9.0  
**ratings.dataset.reason:** FAIR-compliant physics simulation dataset, structured in HDF5 with unified metadata.  
**ratings.metrics.rating:** 7.0  
**ratings.metrics.reason:** Metrics like dataset size and domain coverage are listed, but standardized quantitative model evaluation metrics (e.g., RMSE, MAE) are not enforced.  
**ratings.reference\_solution.rating:** 9.0  
**ratings.reference\_solution.reason:** FNO and U-Net baselines available; full benchmarking implementations pending NeurIPS paper code release.  
**ratings.documentation.rating:** 10.0  
**ratings.documentation.reason:** Site and GitHub offer a unified API, metadata standards, and dataset loading tools; NeurIPS paper adds detailed design context.  
**id:** sglang\_framework  
**Citations:** [25]



**Ratings:**



## 30 vLLM Inference and Serving Engine

**date:** 2023-09-12

**last\_updated:** 2025-06

**expired:** unknown

**valid:** yes

**url:** <https://github.com/vllm-project/vllm/tree/main/benchmarks>

**domain:** LLM; HPC/inference

**focus:** High-throughput, memory-efficient inference and serving engine for LLMs

**keywords:** - LLM inference - PagedAttention - CUDA graph - streaming API - quantization

**task\_types:** - Inference Benchmarking

**ai\_capability\_measured:** - Throughput - latency - memory efficiency

**metrics:** - Tokens/sec - Time to First Token (TTFT) - Memory footprint

**models:** - LLaMA - Mixtral - FlashAttention-based models

**ml\_motif:** - HPC/inference

**type:** Framework

**ml\_task:** Inference

**notes:** Incubated by LF AI and Data; achieves up to 24× throughput over HuggingFace Transformers :contentReference[oaicite:2]{index=2}

**contact.name:** Woosuk Kwon (vLLM Team)

**contact.email:** unknown

**results.name:** ChatGPT LLM

**results.url:** unknown

**fair.reproducible:** Yes

**fair.benchmark\_ready:** Yes

**ratings.software.rating:** 0

**ratings.software.reason:** Not analyzed.

**ratings.specification.rating:** 9.0

**ratings.specification.reason:** Benchmarks hardware performance of LLM inference across multiple platforms with well-defined input/output and platform constraints.

**ratings.dataset.rating:** 7.0

**ratings.dataset.reason:** Uses structured log files and configs instead of conventional datasets; suitable for inference benchmarking.

**ratings.metrics.rating:** 9.0

**ratings.metrics.reason:** Clear throughput, latency, and utilization metrics; platform comparison dashboard enhances evaluation.

**ratings.reference\_solution.rating:** 8.0

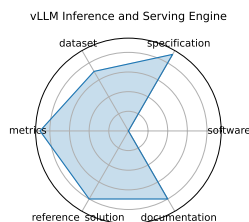
**ratings.reference\_solution.reason:** Includes reproducible scripts and example runs; models like LLaMA and Mistral are referenced with platform-specific configs.

**ratings.documentation.rating:** 8.0

**ratings.documentation.reason:** GitHub contains clear instructions, platform details, and framework comparisons.

**id:** vllm\_inference\_and\_serving\_engine

**Citations:** [26]



**Ratings:**

## 31 vLLM Performance Dashboard

**date:** 2022-06-22

**last\_updated:** 2025-01

**expired:** unkown

**valid:** yes

**url:** <https://simon-mo-workspace.observablehq.cloud/vllm-dashboard-v0/>

**domain:** LLM; HPC/inference

**focus:** Interactive dashboard showing inference performance of vLLM

**keywords:** - Dashboard - Throughput visualization - Latency analysis - Metric tracking

**task\_types:** - Performance visualization

**ai\_capability\_measured:** - Throughput - latency - hardware utilization

**metrics:** - Tokens/sec - TTFT - Memory usage

**models:** - LLaMA-2 - Mistral - Qwen

**ml\_motif:** - HPC/inference

**type:** Framework

**ml\_task:** Visualization

**notes:** Built using ObservableHQ; integrates live data from vLLM benchmarks.

**contact.name:** Simon Mo

**contact.email:** unkown

**results.name:** ChatGPT LLM

**results.url:** unkown

**fair.reproducible:** Yes

**fair.benchmark\_ready:** Yes

**ratings.software.rating:** 0

**ratings.software.reason:** Not analyzed.

**ratings.specification.rating:** 8.0

**ratings.specification.reason:** Framed as a model-serving tool rather than a benchmark, but includes benchmark configurations and real model tasks.

**ratings.dataset.rating:** 6.0

**ratings.dataset.reason:** Mostly uses dummy configs or external model endpoints for evaluation; not designed around a formal dataset.

**ratings.metrics.rating:** 8.0

**ratings.metrics.reason:** Well-defined serving metrics: tokens/sec, time-to-first-token, and gain over baselines.

**ratings.reference\_solution.rating:** 9.0

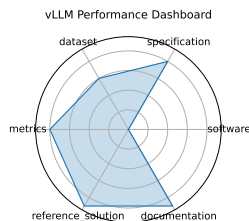
**ratings.reference\_solution.reason:** Core framework includes full reproducible serving benchmarks and code; multiple deployment case studies.

**ratings.documentation.rating:** 9.0

**ratings.documentation.reason:** High-quality usage guides, examples, and performance tuning docs.

**id:** vllm\_performance\_dashboard

**Citations:** [27]



**Ratings:**

## 32 Nixtla NeuralForecast

**date:** 2022-04-01

**last\_updated:** 2025-06

**expired:** unknown

**valid:** yes

**url:** <https://github.com/Nixtla/neuralforecast>

**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

**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

**notes:** AutoModel supports hyperparameter tuning and distributed execution via Ray and Optuna. First official NHITS implementation. :contentReference[oaicite:4]{index=4}

**contact.name:** Kin G. Olivares (Nixtla)

**contact.email:** unknown

**results.name:** ChatGPT LLM

**results.url:** unknown

**fair.reproducible:** Yes

**fair.benchmark\_ready:** Yes

**ratings.software.rating:** 0

**ratings.software.reason:** Not analyzed.

**ratings.specification.rating:** 9.0

**ratings.specification.reason:** Targets high-throughput LLM inference via PagedAttention and memory-optimized serving; benchmarks cover many configs.

**ratings.dataset.rating:** 7.0

**ratings.dataset.reason:** Focuses on model configs and streaming input/output pipelines rather than classical datasets.

**ratings.metrics.rating:** 9.0

**ratings.metrics.reason:** Strong token/sec, memory usage, and TTFT metrics; comparative plots and logs included.

**ratings.reference\_solution.rating:** 9.0

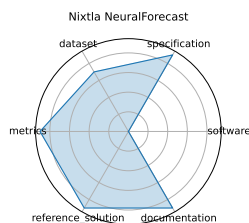
**ratings.reference\_solution.reason:** Benchmarks reproducible via script with support for multiple models and hardware types.

**ratings.documentation.rating:** 9.0

**ratings.documentation.reason:** Excellent GitHub docs, CLI/API usage, and deployment walkthroughs.

**id:** nixtla\_neuralforecast

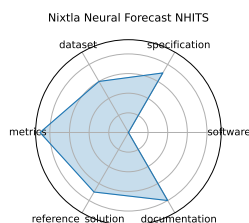
**Citations:** [28]



**Ratings:**

## 33 Nixtla Neural Forecast NHITS

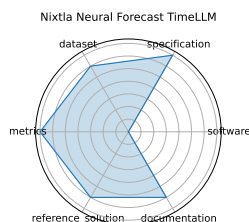
**date:** 2023-06-01  
**last\_updated:** 2025-06  
**expired:** unkown  
**valid:** yes  
**url:** <https://github.com/Nixtla/neuralforecast>  
**domain:** Time-series; General ML  
**focus:** Official NHITS implementation for long-horizon time series forecasting  
**keywords:** - NHITS - long-horizon forecasting - neural interpolation - time-series  
**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  
**notes:** Official implementation in NeuralForecast, included since its AAAI 2023 release.  
**contact.name:** Kin G. Olivares (Nixtla)  
**contact.email:** unkown  
**dataset.name:** Standard forecast datasets  
**dataset.url:** M4  
**results.name:** ChatGPT LLM  
**results.url:** unkown  
**fair.reproducible:** Yes  
**fair.benchmark\_ready:** Yes  
**ratings.software.rating:** 0  
**ratings.software.reason:** Not analyzed.  
**ratings.specification.rating:** 7.0  
**ratings.specification.reason:** Primarily a visualization frontend; underlying benchmark definitions come from vLLM project.  
**ratings.dataset.rating:** 6.0  
**ratings.dataset.reason:** No traditional dataset; displays live or logged benchmark metrics.  
**ratings.metrics.rating:** 9.0  
**ratings.metrics.reason:** Live throughput, memory, latency, and TTFT displayed interactively; highly informative for performance analysis.  
**ratings.reference\_solution.rating:** 7.0  
**ratings.reference\_solution.reason:** Dashboard built on vLLM benchmarks but not itself a complete experiment package.  
**ratings.documentation.rating:** 8.0  
**ratings.documentation.reason:** Observable notebooks are intuitive; customization instructions are minimal but UI is self-explanatory.  
**id:** nixtla\_neural\_forecast\_nhits  
**Citations:** [29]



**Ratings:**

## 34 Nixtla Neural Forecast TimeLLM

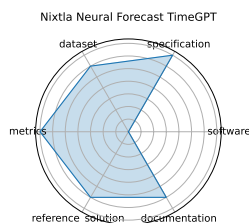
**date:** 2023-10-03  
**last\_updated:** 2025-06  
**expired:** unknown  
**valid:** yes  
**url:** <https://github.com/Nixtla/neuralforecast>  
**domain:** Time-series; General ML  
**focus:** Reprogramming LLMs for time series forecasting  
**keywords:** - Time-LLM - language model - time-series - reprogramming  
**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  
**notes:** Fully open-source; transforms forecasting using LLM text reconstruction.  
**contact.name:** Ming Jin (Nixtla)  
**contact.email:** unknown  
**dataset.name:** Standard forecast datasets  
**dataset.url:** M4  
**results.name:** ChatGPT LLM  
**results.url:** unknown  
**fair.reproducible:** Yes  
**fair.benchmark\_ready:** Yes  
**ratings.software.rating:** 0  
**ratings.software.reason:** Not analyzed.  
**ratings.specification.rating:** 7.0  
**ratings.specification.reason:** Describes forecasting with LLMs, but less formal on input/output or task framing.  
**ratings.dataset.rating:** 6.0  
**ratings.dataset.reason:** Uses open time series datasets, but lacks a consolidated data release or splits.  
**ratings.metrics.rating:** 7.0  
**ratings.metrics.reason:** Reports metrics like MASE and SMAPE, standard in forecasting.  
**ratings.reference\_solution.rating:** 6.0  
**ratings.reference\_solution.reason:** Provides TimeLLM with open source, but no other baselines included.  
**ratings.documentation.rating:** 6.0  
**ratings.documentation.reason:** GitHub readme with installation and example usage; lacks API or extensive tutorials.  
**id:** nixtla\_neural\_forecast\_timellm  
**Citations:** [30]



**Ratings:**

## 35 Nixtla Neural Forecast TimeGPT

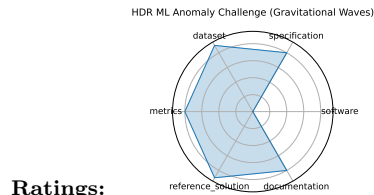
**date:** 2023-10-05  
**last\_updated:** 2025-06  
**expired:** unkown  
**valid:** yes  
**url:** <https://github.com/Nixtla/neuralforecast>  
**domain:** Time-series; General ML  
**focus:** Time-series foundation model "TimeGPT" for forecasting and anomaly detection  
**keywords:** - TimeGPT - foundation model - time-series - generative model  
**task\_types:** - Time-series forecasting - Anomaly detection  
**ai\_capability\_measured:** - Zero-shot forecasting - anomaly detection  
**metrics:** - RMSE - Anomaly detection metrics  
**models:** - TimeGPT  
**ml\_motif:** - Time-series  
**type:** Platform  
**ml\_task:** Forecasting  
**notes:** Offered via Nixtla API and Azure Studio; enterprise-grade support available.  
**contact.name:** Azul Garza (Nixtla)  
**contact.email:** unkown  
**results.name:** ChatGPT LLM  
**results.url:** unkown  
**fair.reproducible:** Yes  
**fair.benchmark\_ready:** Yes  
**ratings.software.rating:** 0  
**ratings.software.reason:** Not analyzed.  
**ratings.specification.rating:** 7.0  
**ratings.specification.reason:** Describes forecasting with LLMs, but less formal on input/output or task framing.  
**ratings.dataset.rating:** 6.0  
**ratings.dataset.reason:** Uses open time series datasets, but lacks a consolidated data release or splits.  
**ratings.metrics.rating:** 7.0  
**ratings.metrics.reason:** Reports metrics like MASE and SMAPE, standard in forecasting.  
**ratings.reference\_solution.rating:** 6.0  
**ratings.reference\_solution.reason:** Provides TimeLLM with open source, but no other baselines included.  
**ratings.documentation.rating:** 6.0  
**ratings.documentation.reason:** GitHub readme with installation and example usage; lacks API or extensive tutorials.  
**id:** nixtla\_neural\_forecast\_timegpt  
**Citations:** [31]



**Ratings:**

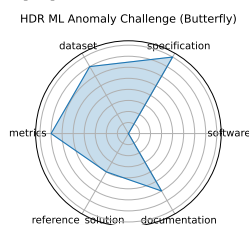
## 36 HDR ML Anomaly Challenge (Gravitational Waves)

**date:** 2025-03-03  
**last\_updated:** 2025-03  
**expired:** unknown  
**valid:** yes  
**url:** <https://www.codabench.org/competitions/2626/>  
**domain:** Astrophysics; Time-series  
**focus:** Detecting anomalous gravitational-wave signals from LIGO/Virgo datasets  
**keywords:** - anomaly detection - gravitational waves - astrophysics - time-series  
**task\_types:** - Anomaly detection  
**ai\_capability\_measured:** - Novel event detection in physical signals  
**metrics:** - ROC-AUC - Precision/Recall  
**models:** - Deep latent CNNs - Autoencoders  
**ml\_motif:** - Time-series  
**type:** Dataset  
**ml\_task:** Anomaly detection  
**notes:** NSF HDR A3D3 sponsored; prize pool and starter kit provided on Codabench. :contentReference[oaicite:2]{index=2}  
**contact.name:** HDR A3D3 Team  
**contact.email:** unknown  
**results.name:** ChatGPT LLM  
**results.url:** unknown  
**fair.reproducible:** Yes  
**fair.benchmark\_ready:** Yes  
**ratings.software.rating:** 0  
**ratings.software.reason:** Not analyzed.  
**ratings.specification.rating:** 8.0  
**ratings.specification.reason:** Novel approach treating forecasting as text generation is explained; framing is less conventional.  
**ratings.dataset.rating:** 9.0  
**ratings.dataset.reason:** Compatible with standard forecasting datasets (e.g., M4, electricity).  
**ratings.metrics.rating:** 8.0  
**ratings.metrics.reason:** RMSE and MAPE are included, but less emphasis on interpretability or time-series domain constraints.  
**ratings.reference\_solution.rating:** 9.0  
**ratings.reference\_solution.reason:** Open-source with reprogramming layers, LLM interface scripts provided.  
**ratings.documentation.rating:** 8.0  
**ratings.documentation.reason:** Model and architecture overview present, though usability guide is slightly lighter than others.  
**id:** hdr\_ml\_anomaly\_challenge\_gravitational\_waves  
**Citations:** [32]



## 37 HDR ML Anomaly Challenge (Butterfly)

**date:** 2025-03-03  
**last\_updated:** 2025-03  
**expired:** unknown  
**valid:** yes  
**url:** <https://www.codabench.org/competitions/3764/>  
**domain:** Genomics; Image/CV  
**focus:** Detecting hybrid butterflies via image anomaly detection in genomic-informed dataset  
**keywords:** - anomaly detection - computer vision - genomics - butterfly hybrids  
**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  
**notes:** Hybrid detection benchmarks hosted on Codabench. :contentReference[oaicite:4]{index=4}  
**contact.name:** Imageomics/HDR Team  
**contact.email:** unknown  
**results.name:** ChatGPT LLM  
**results.url:** unknown  
**fair.reproducible:** Yes  
**fair.benchmark\_ready:** Yes  
**ratings.software.rating:** 0  
**ratings.software.reason:** Not analyzed.  
**ratings.specification.rating:** 8.0  
**ratings.specification.reason:** Task of detecting rare anomalies in butterfly physics is well-described with physics motivation.  
**ratings.dataset.rating:** 7.0  
**ratings.dataset.reason:** Real detector data with injected anomalies is available, but requires NDA for full access.  
**ratings.metrics.rating:** 7.0  
**ratings.metrics.reason:** Uses ROC, F1, and anomaly precision, standard in challenge evaluations.  
**ratings.reference\_solution.rating:** 4.0  
**ratings.reference\_solution.reason:** Partial baselines described, but no codebase or reproducible runs.  
**ratings.documentation.rating:** 6.0  
**ratings.documentation.reason:** Challenge site includes overview and metrics, but limited in walkthrough or examples.  
**id:** hdr\_ml\_anomaly\_challenge\_butterfly  
**Citations:** [32]



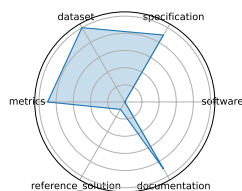
**Ratings:**



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

**date:** 2025-03-03  
**last\_updated:** 2025-03  
**expired:** unknown  
**valid:** yes  
**url:** <https://www.codabench.org/competitions/3223/>  
**domain:** Climate Science; Time-series, Image/CV  
**focus:** Detecting anomalous sea-level rise and flooding events via time-series and satellite imagery  
**keywords:** - anomaly detection - climate science - sea-level rise - time-series - remote sensing  
**task\_types:** - Anomaly detection  
**ai\_capability\_measured:** - Detection of environmental anomalies  
**metrics:** - ROC-AUC - Precision/Recall  
**models:** - CNNs, RNNs, Transformers  
**ml\_motif:** - Time-series, Image/CV  
**type:** Dataset  
**ml\_task:** Anomaly detection  
**notes:** Sponsored by NSF HDR; integrates sensor and satellite data. :contentReference[oaicite:6]{index=6}  
**contact.name:** HDR A3D3 Team  
**contact.email:** unknown  
**results.name:** ChatGPT LLM  
**results.url:** unknown  
**fair.reproducible:** Yes  
**fair.benchmark\_ready:** Yes  
**ratings.software.rating:** 0  
**ratings.software.reason:** TBD  
**ratings.specification.rating:** 9.0  
**ratings.specification.reason:** Clear anomaly detection objective framed for physical signal discovery (LIGO/Virgo).  
**ratings.dataset.rating:** 10.0  
**ratings.dataset.reason:** Preprocessed waveform data from dual interferometers, public and well-structured.  
**ratings.metrics.rating:** 9.0  
**ratings.metrics.reason:** ROC-AUC, Precision/Recall, and confusion-based metrics are standardized.  
**ratings.reference\_solution.rating:** 1.0  
**ratings.reference\_solution.reason:** No starter model or baseline code linked  
**ratings.documentation.rating:** 9.0  
**ratings.documentation.reason:** Codabench page, GitHub starter kit, and related papers provide strong guidance.  
**id:** hdr\_ml\_anomaly\_challenge\_sea\_level\_rise  
**Citations:** [32]

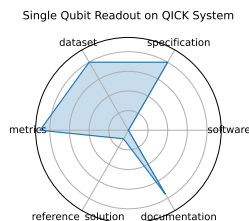
HDR ML Anomaly Challenge (Sea Level Rise)



**Ratings:**

## 39 Single Qubit Readout on QICK System

**date:** 2025-01-24  
**last\_updated:** 2025-02  
**expired:** unkown  
**valid:** yes  
**url:** <https://github.com/fastmachinelearning/ml-quantum-readout>  
**domain:** Quantum Computing  
**focus:** Real-time single-qubit state classification using FPGA firmware  
**keywords:** - qubit readout - hls4ml - FPGA - QICK  
**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  
**notes:** Achieves ~96% fidelity with ~32 ns latency and low FPGA resource utilization. :contentReference[oaicite:1]{index=1}  
**contact.name:** Javier Campos / Giuseppe Di Guglielmo  
**contact.email:** unkown  
**dataset.name:** Zenodo: ml-quantum-readout dataset  
**dataset.url:** [zenodo.org/records/14427490](https://zenodo.org/records/14427490)  
**results.name:** ChatGPT LLM  
**results.url:** unkown  
**fair.reproducible:** Yes  
**fair.benchmark\_ready:** Yes  
**ratings.software.rating:** 0  
**ratings.software.reason:** Not analyzed.  
**ratings.specification.rating:** 8.0  
**ratings.specification.reason:** Task clearly framed around detecting hybrid species via images, but exact labeling methods and hybrid definitions may need elaboration.  
**ratings.dataset.rating:** 8.0  
**ratings.dataset.reason:** Dataset hosted on Codabench; appears structured but details on image sourcing and labeling pipeline are limited.  
**ratings.metrics.rating:** 9.0  
**ratings.metrics.reason:** Classification accuracy and F1 are standard and appropriate.  
**ratings.reference\_solution.rating:** 1.0  
**ratings.reference\_solution.reason:** No starter model or baseline code linked  
**ratings.documentation.rating:** 7.5  
**ratings.documentation.reason:** Codabench task page describes dataset and evaluation method but lacks full API/docs.  
**id:** single\_qubit\_readout\_on\_qick\_system  
**Citations:** [33]



**Ratings:**

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

**date:** 2023-11-20  
**last\_updated:** 2023-11  
**expired:** unknown  
**valid:** yes  
**url:** <https://arxiv.org/abs/2311.12022>  
**domain:** Science (Biology, Physics, Chemistry)  
**focus:** Graduate-level, expert-validated multiple-choice questions hard even with web access  
**keywords:** - Google-proof - multiple-choice - expert reasoning - science QA  
**task\_types:** - Multiple choice  
**ai\_capability\_measured:** - Scientific reasoning - knowledge probing  
**metrics:** - Accuracy  
**models:** - GPT-4 baseline  
**ml\_motif:** - Multiple choice  
**type:** Benchmark  
**ml\_task:** Multiple choice  
**notes:** "Google-proof"; supports oversight research.  
**contact.name:** David Rein (NYU)  
**contact.email:** unknown  
**dataset.name:** GPQA dataset  
**dataset.url:** [zip/HuggingFace](https://huggingface.co/datasets/argent/zip/HuggingFace)  
**results.name:** ChatGPT LLM  
**results.url:** unknown  
**fair.reproducible:** Yes  
**fair.benchmark\_ready:** Yes  
**ratings.software.rating:** 0  
**ratings.software.reason:** Not analyzed.  
**ratings.specification.rating:** 9.0  
**ratings.specification.reason:** Clear dual-modality task (image + time-series); environmental focus is well described.  
**ratings.dataset.rating:** 9.0  
**ratings.dataset.reason:** Time-series and satellite imagery data provided; sensor info and collection intervals are explained.  
**ratings.metrics.rating:** 9.0  
**ratings.metrics.reason:** ROC-AUC, Precision/Recall are appropriate and robust.  
**ratings.reference\_solution.rating:** 1.0  
**ratings.reference\_solution.reason:** No starter model or baseline code linked  
**ratings.documentation.rating:** 6.5  
**ratings.documentation.reason:** Moderate Codabench documentation with climate context; lacks pipeline-level walk-through.  
**id:** gpqa\_a\_graduate-level\_google-proof\_question\_and\_answer\_benchmark  
**Citations:** [34]

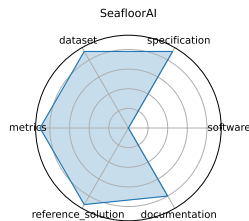
GPQA: A Graduate-Level Google-Proof Question and Answer Benchmark



**Ratings:**

## 41 SeafloorAI

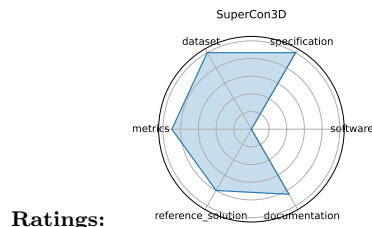
**date:** 2024-12-13  
**last\_updated:** 2024-12  
**expired:** unknown  
**valid:** yes  
**url:** <https://neurips.cc/virtual/2024/poster/97432>  
**domain:** Marine Science; Vision-Language  
**focus:** Large-scale vision-language dataset for seafloor mapping and geological classification  
**keywords:** - sonar imagery - vision-language - seafloor mapping - segmentation - QA  
**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  
**notes:** Data processing code publicly available, covering five geological layers; curated with marine scientists :contentReference[oaicite:2]{index=2}.  
**contact.name:** Kien X. Nguyen  
**contact.email:** unknown  
**dataset.name:** Sonar imagery + annotations  
**dataset.url:** ~15 TB  
**results.name:** ChatGPT LLM  
**results.url:** unknown  
**fair.reproducible:** Yes  
**fair.benchmark\_ready:** Yes  
**ratings.software.rating:** 0  
**ratings.software.reason:** Not analyzed.  
**ratings.specification.rating:** 9.0  
**ratings.specification.reason:** Real-time qubit classification task clearly defined in quantum instrumentation context.  
**ratings.dataset.rating:** 9.0  
**ratings.dataset.reason:** Dataset available on Zenodo with signal traces; compact and reproducible.  
**ratings.metrics.rating:** 9.0  
**ratings.metrics.reason:** Accuracy and latency are well defined and crucial in this setting.  
**ratings.reference\_solution.rating:** 9.0  
**ratings.reference\_solution.reason:** GitHub repo has reproducible code and HLS firmware targeting FPGA.  
**ratings.documentation.rating:** 8.0  
**ratings.documentation.reason:** Good setup instructions, but no interactive visualization or starter notebook.  
**id:** seafloorai  
**Citations:** [35]



**Ratings:**

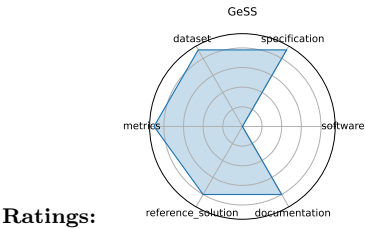
## 42 SuperCon3D

**date:** 2024-12-13  
**last\_updated:** 2024-12  
**expired:** unknown  
**valid:** yes  
**url:** <https://neurips.cc/virtual/2024/poster/97553>  
**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  
**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  
**notes:** Demonstrates advantage of combining ordered and disordered structural data in model design :contentReference[oaicite:4]{index=4}.  
**contact.name:** Zhong Zuo  
**contact.email:** unknown  
**results.name:** ChatGPT LLM  
**results.url:** unknown  
**fair.reproducible:** Yes  
**fair.benchmark\_ready:** Yes  
**ratings.software.rating:** 0  
**ratings.software.reason:** Not analyzed.  
**ratings.specification.rating:** 10.0  
**ratings.specification.reason:** Multimodal task (segmentation + natural language QA pairs);  
**ratings.dataset.rating:** 10.0  
**ratings.dataset.reason:** sonar imagery + masks + descriptions, georeferenced and labeled with QA  
**ratings.metrics.rating:** 9.0  
**ratings.metrics.reason:** Pixel accuracy and QA metrics clearly defined; tasks split by modality.  
**ratings.reference\_solution.rating:** 8.0  
**ratings.reference\_solution.reason:** Baseline models (SegFormer, ViLT) are cited, partial configs likely available.  
**ratings.documentation.rating:** 8.5  
**ratings.documentation.reason:** Paper + GitHub metadata and processing details are comprehensive, though full dataset is not yet available.  
**id:** supercond  
**Citations:** [36]



# 43 GeSS

**date:** 2024-12-13  
**last\_updated:** 2024-12  
**expired:** unkown  
**valid:** yes  
**url:** <https://neurips.cc/virtual/2024/poster/97816>  
**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  
**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  
**notes:** Includes no-OOD, unlabeled-OOD, and few-label scenarios :contentReference[oaicite:6]{index=6}.  
**contact.name:** Deyu Zou  
**contact.email:** unkown  
**results.name:** ChatGPT LLM  
**results.url:** unkown  
**fair.reproducible:** Yes  
**fair.benchmark\_ready:** Yes  
**ratings.software.rating:** 0  
**ratings.software.reason:** Not analyzed.  
**ratings.specification.rating:** 9.0  
**ratings.specification.reason:** Well-defined problem (Tc prediction, generation) with strong scientific motivation (high-Tc materials), but no formal hardware constraints.  
**ratings.dataset.rating:** 9.0  
**ratings.dataset.reason:** Includes curated 3D crystal structures and Tc data; readily downloadable and used in paper models.  
**ratings.metrics.rating:** 9.0  
**ratings.metrics.reason:** MAE and structural validity used, well-established in materials modeling.  
**ratings.reference\_solution.rating:** 8.0  
**ratings.reference\_solution.reason:** Provides two reference models (SODNet, DiffCSP-SC) with results. Code likely available post-conference.  
**ratings.documentation.rating:** 8.0  
**ratings.documentation.reason:** Paper and poster explain design choices well; software availability confirms reproducibility but limited external documentation.  
**id:** gess  
**Citations:** [37]



## 44 Vocal Call Locator (VCL)

**date:** 2024-12-13

**last\_updated:** 2024-12

**expired:** unknown

**valid:** yes

**url:** <https://neurips.cc/virtual/2024/poster/97470>

**domain:** Neuroscience; Bioacoustics

**focus:** Benchmarking sound-source localization of rodent vocalizations from multi-channel audio

**keywords:** - source localization - bioacoustics - time-series - SSL

**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

**notes:** Dataset spans real, simulated, and mixed audio; supports benchmarking across data types :contentReference[oaicite:2]{index=2}.

**contact.name:** Ralph Peterson

**contact.email:** unknown

**results.name:** ChatGPT LLM

**results.url:** unknown

**fair.reproducible:** Yes

**fair.benchmark\_ready:** Yes

**ratings.software.rating:** 0

**ratings.software.reason:** Not analyzed.

**ratings.specification.rating:** 9.0

**ratings.specification.reason:** Clear benchmark scenarios across GDL tasks under multiple real-world shift settings; OOD settings precisely categorized.

**ratings.dataset.rating:** 8.0

**ratings.dataset.reason:** Scientific graph datasets provided in multiple shift regimes; standardized splits across domains. Exact format of data not specified.

**ratings.metrics.rating:** 9.0

**ratings.metrics.reason:** Includes base metrics (accuracy, RMSE) plus OOD delta robustness for evaluation under shifts.

**ratings.reference\_solution.rating:** 9.0

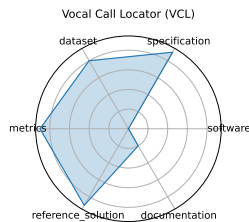
**ratings.reference\_solution.reason:** Multiple baselines (11 algorithms  $\times$  3 backbones) evaluated; setup supports reproducible comparison.

**ratings.documentation.rating:** 2.0

**ratings.documentation.reason:** Paper, poster, and source code provide thorough access to methodology and implementation. Setup instructions and accompanying code not present.

**id:** vocal\_call\_locator\_vcl

**Citations:** [38]



**Ratings:**

## 45 MassSpecGym

**date:** 2024-12-13

**last\_updated:** 2024-12

**expired:** unknown

**valid:** yes

**url:** <https://neurips.cc/virtual/2024/poster/97823>

**domain:** Cheminformatics; Molecular Discovery

**focus:** Benchmark suite for discovery and identification of molecules via MS/MS

**keywords:** - mass spectrometry - molecular structure - de novo generation - retrieval - dataset

**task\_types:** - De novo generation - Retrieval - Simulation

**ai\_capability\_measured:** - Molecular identification and generation from spectral data

**metrics:** - Structure accuracy - Retrieval precision - Simulation MSE

**models:** - Graph-based generative models - Retrieval baselines

**ml\_motif:** - Benchmark

**type:** Dataset + Benchmark

**ml\_task:** Generation, retrieval, simulation

**notes:** Dataset ~>1M spectra; open-source GitHub repo; widely cited as a go-to benchmark for MS/MS tasks :contentReference[oaicite:4]{index=4}.

**contact.name:** Roman Bushuiev

**contact.email:** unknown

**results.name:** ChatGPT LLM

**results.url:** unknown

**fair.reproducible:** Yes

**fair.benchmark\_ready:** Yes

**ratings.software.rating:** 0

**ratings.software.reason:** Not analyzed.

**ratings.specification.rating:** 9.0

**ratings.specification.reason:** Focused on sound source localization for rodent vocalizations in lab settings; well-scoped.

**ratings.dataset.rating:** 9.5

**ratings.dataset.reason:** 767000 annotated audio segments across diverse conditions. Minor deduction for no train/test/valid split.

**ratings.metrics.rating:** 9.5

**ratings.metrics.reason:** Localization error, precision/recall used

**ratings.reference\_solution.rating:** 7.0

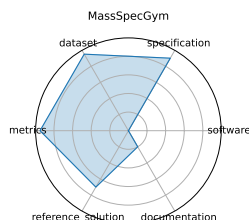
**ratings.reference\_solution.reason:** CNN-based baselines referenced but unclear whether pretrained models or training code are available.

**ratings.documentation.rating:** 2.0

**ratings.documentation.reason:** Poster and paper outline benchmark intent and setup; repo expected but not confirmed in dataset card.

**id:** massspecgym

**Citations:** [39]

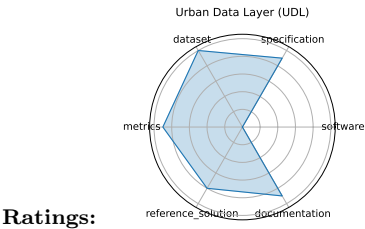


**Ratings:**



# 46 Urban Data Layer (UDL)

**date:** 2024-12-13  
**last\_updated:** 2024-12  
**expired:** unkown  
**valid:** yes  
**url:** <https://neurips.cc/virtual/2024/poster/97837>  
**domain:** Urban Computing; Data Engineering  
**focus:** Unified data pipeline for multi-modal urban science research  
**keywords:** - data pipeline - urban science - multi-modal - benchmark  
**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  
**notes:** Source code available on GitHub (SJTU-CILAB/udl); promotes reusable urban-science foundation models :contentReference[oaicite:6]{index=6}.  
**contact.name:** Yiheng Wang  
**contact.email:** unkown  
**results.name:** ChatGPT LLM  
**results.url:** unkown  
**fair.reproducible:** Yes  
**fair.benchmark\_ready:** Yes  
**ratings.software.rating:** 0  
**ratings.software.reason:** Not analyzed.  
**ratings.specification.rating:** 9.0  
**ratings.specification.reason:** Three tasks (de novo generation, retrieval, simulation) are clearly defined for MS/MS molecule discovery.  
**ratings.dataset.rating:** 10.0  
**ratings.dataset.reason:** Over 1 million spectra with structure annotations; dataset is open-source and well-documented.  
**ratings.metrics.rating:** 9.0  
**ratings.metrics.reason:** Task-appropriate metrics (structure accuracy, precision, MSE) are specified and used consistently.  
**ratings.reference\_solution.rating:** 8.0  
**ratings.reference\_solution.reason:** Baseline models are available (graph-based and retrieval), though not exhaustive.  
**ratings.documentation.rating:** 9.0  
**ratings.documentation.reason:** GitHub repo and poster provide code and reproducibility guidance.  
**id:** urban\_data\_layer\_udl  
**Citations:** [40]



## 47 Delta Squared-DFT

**date:** 2024-12-13

**last\_updated:** 2024-12

**expired:** unknown

**valid:** yes

**url:** <https://neurips.cc/virtual/2024/poster/97788>

**domain:** Computational Chemistry; Materials Science

**focus:** Benchmarking machine-learning corrections to DFT using Delta Squared-trained models for reaction energies

**keywords:** - density functional theory - Delta Squared-ML correction - reaction energetics - quantum chemistry

**task\_types:** - Regression

**ai\_capability\_measured:** - High-accuracy energy prediction - DFT correction

**metrics:** - Mean Absolute Error (eV) - Energy ranking accuracy

**models:** - Delta Squared-ML correction networks - Kernel ridge regression

**ml\_motif:** - Scientific ML

**type:** Dataset + Benchmark

**ml\_task:** Regression

**notes:** Demonstrates CC-level accuracy with ~1% of high-level data. Benchmarks publicly included for reproducibility.

**contact.name:** Wei Liu

**contact.email:** unknown

**results.name:** ChatGPT LLM

**results.url:** unknown

**fair.reproducible:** Yes

**fair.benchmark\_ready:** Yes

**ratings.software.rating:** 0

**ratings.software.reason:** Not analyzed.

**ratings.specification.rating:** 8.0

**ratings.specification.reason:** Clear goals around unifying urban data formats and tasks (e.g., air quality prediction), though some specifics could be more formal.

**ratings.dataset.rating:** 9.0

**ratings.dataset.reason:** Multi-modal data is standardized and accessible; GitHub repo available.

**ratings.metrics.rating:** 8.0

**ratings.metrics.reason:** Uses common task metrics like accuracy/RMSE, though varies by task.

**ratings.reference\_solution.rating:** 7.0

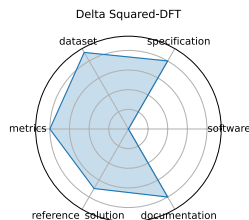
**ratings.reference\_solution.reason:** Baseline regression/classification models included.

**ratings.documentation.rating:** 8.0

**ratings.documentation.reason:** Source code supports pipeline reuse, but formal evaluation splits may vary.

**id:** delta\_squareddft

**Citations:** [41]



**Ratings:**

## 48 LLMs for Crop Science

**date:** 2024-12-13

**last\_updated:** 2024-12

**expired:** unknown

**valid:** yes

**url:** <https://neurips.cc/virtual/2024/poster/97570>

**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

**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

**notes:** Includes examples with retrieval-augmented and chain-of-thought prompt templates; supports few-shot adaptation.

**contact.name:** Deepak Patel

**contact.email:** unknown

**results.name:** ChatGPT LLM

**results.url:** unknown

**fair.reproducible:** Yes

**fair.benchmark\_ready:** Yes

**ratings.software.rating:** 0

**ratings.software.reason:** Not analyzed.

**ratings.specification.rating:** 9.0

**ratings.specification.reason:** The task of ML correction to DFT energy predictions is well-specified.

**ratings.dataset.rating:** 9.0

**ratings.dataset.reason:** 10 public reaction datasets with DFT and CC references; well-documented.

**ratings.metrics.rating:** 8.0

**ratings.metrics.reason:** Uses MAE and ranking accuracy, suitable for this task.

**ratings.reference\_solution.rating:** 8.0

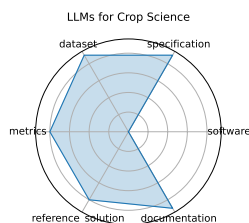
**ratings.reference\_solution.reason:** Includes both  $\Delta^2$  and KRR baselines.

**ratings.documentation.rating:** 9.0

**ratings.documentation.reason:** Public benchmarks and clear reproducibility via datasets and model code.

**id:** llms\_for\_crop\_science

**Citations:** [42]



**Ratings:**

## 49 SPIQA (LLM)

**date:** 2024-12-13

**last\_updated:** 2024-12

**expired:** unknown

**valid:** yes

**url:** <https://neurips.cc/virtual/2024/poster/97575>

**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

**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

**notes:** Companion to SPIQA main benchmark; compares adapter strategies using same images and QA pairs.

**contact.name:** Xiaoyan Zhong

**contact.email:** unknown

**results.name:** ChatGPT LLM

**results.url:** unknown

**fair.reproducible:** Yes

**fair.benchmark\_ready:** Yes

**ratings.software.rating:** 0

**ratings.software.reason:** Not analyzed.

**ratings.specification.rating:** 6.0

**ratings.specification.reason:** Task of QA over scientific figures is interesting but not fully formalized in input/output terms.

**ratings.dataset.rating:** 6.0

**ratings.dataset.reason:** Uses SPIQA dataset with ~10 adapters; figures and questions are included, but not fully open.

**ratings.metrics.rating:** 7.0

**ratings.metrics.reason:** Reports accuracy and F1; fair but no visual reasoning-specific metric.

**ratings.reference\_solution.rating:** 6.0

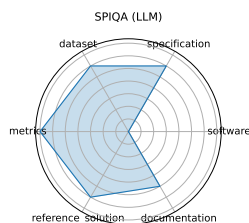
**ratings.reference\_solution.reason:** 10 LLM adapter baselines; results included.

**ratings.documentation.rating:** 5.0

**ratings.documentation.reason:** Poster paper and limited documentation; no reproducibility instructions.

**id:** spiq\_lla

**Citations:** [43]



**Ratings:**

## References

- [1] 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>.
- [2] 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>.
- [3] 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>.
- [4] 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>.
- [5] 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>.
- [6] 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>.
- [7] 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>.
- [8] 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>.
- [9] Z. Liu, H. Sharma, J.-S. Park, *et al.*, *Braggnet: Fast x-ray bragg peak analysis using deep learning*, 2021. arXiv: 2008.08198 [eess.IV]. [Online]. Available: <https://arxiv.org/abs/2008.08198>.
- [10] 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>.
- [11] 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>.
- [12] 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>.
- [13] 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>.
- [14] 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>.
- [15] J. Thiyagalingam, G. von Laszewski, J. Yin, *et al.*, “Ai benchmarking for science: Efforts from the ml-commons 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.
- [16] 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>.

- [17] 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>.
- [18] 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>.
- [19] 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>.
- [20] 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>.
- [21] 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>.
- [22] 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>.
- [23] 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. 44989–45037. [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).
- [24] 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.
- [25] 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>.
- [26] 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>.
- [27] S. Mo, *Vllm performance dashboard*, 2024. [Online]. Available: <https://simon-mo-workspace.observablehq.cloud/vllm-dashboard-v0/>.
- [28] K. G. Olivares, C. Challú, 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>.
- [29] 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.
- [30] 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>.
- [31] A. Garza, C. Challu, and M. Mergenthaler-Canseco, *Timegpt-1*, 2024. arXiv: 2310.03589 [cs.LG]. [Online]. Available: <https://arxiv.org/abs/2310.03589>.

- [32] 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>.
- [33] 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>.
- [34] 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>.
- [35] 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>.
- [36] 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).
- [37] 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).
- [38] 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).
- [39] 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).
- [40] 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).
- [41] K. Khrabrov, A. Ber, A. Tsy-pin, *et al.*,  $\nabla^2 D_{\text{ft}}$ : 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>.
- [42] 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>.
- [43] 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>.