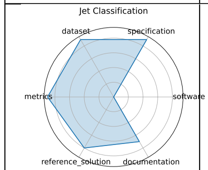
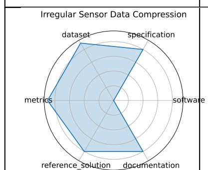
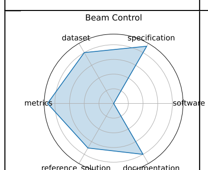
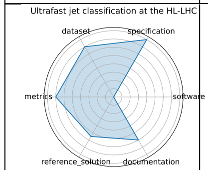
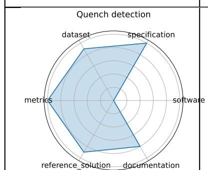
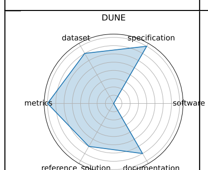
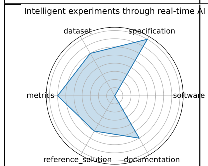
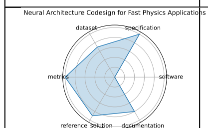
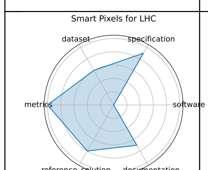
# 1 Benchmark Overview Table

| 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 learning control of accelerator beam position | RL, beam stabilization, control systems, simulation | Control | Policy performance in simulated accelerator control | Stability, Control loss | DDPG, PPO (planned) | [2], [3]⇒ |

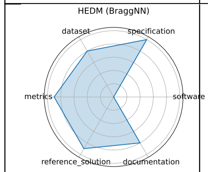
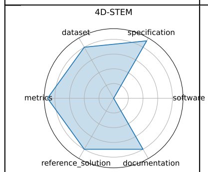
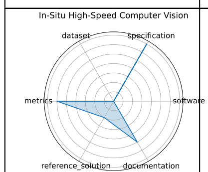
Continued on next page

| 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]⇒ |

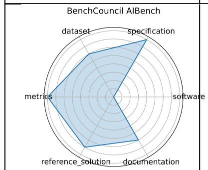
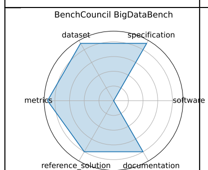
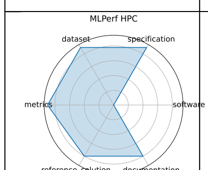
Continued on next page

| 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 (micros), Resource utilization (LUT/FF/BRAM/DSP) | Bipartite Graph Network with Set Transformers (BGN-ST), GarNet (edge-classifier) | [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]⇒ |

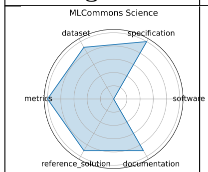
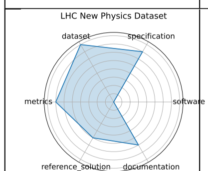
Continued on next page

| 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]⇒ |

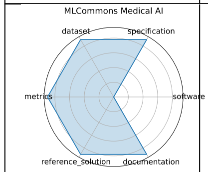
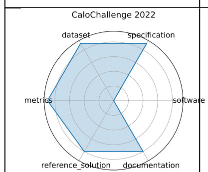
Continued on next page

| 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 AI workloads | System-level AI workload performance | Throughput, Latency, Accuracy | ResNet, BERT, GANs, Recommendation systems | [12]⇒ |
|  | BenchCouncil BigDataBench | General | Big data and AI benchmarking across structured, semi-structured, and unstructured data workloads | big data, AI benchmarking, data analytics | Data preprocessing, Inference, End-to-end data 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]⇒ |

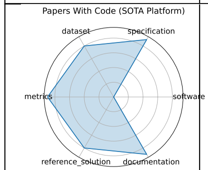
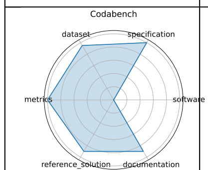
Continued on next page

| 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]⇒ |

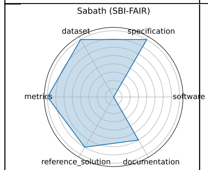
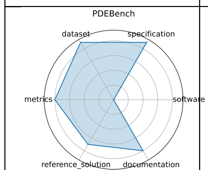
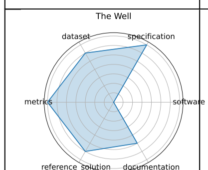
Continued on next page

| 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]⇒ |

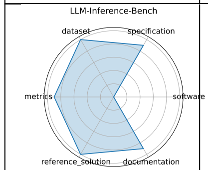
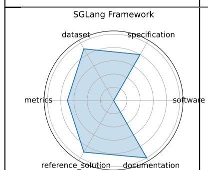
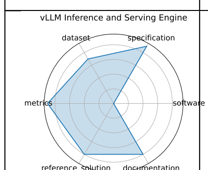
Continued on next page

| Ratings | Name | Domain | Focus | Keywords | Task Types | AI Capability | Metrics | Models | Citation |
|---|---|---|---|---|---|---|---|---|---|
|  | 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]⇒ |
|  | 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]⇒ |

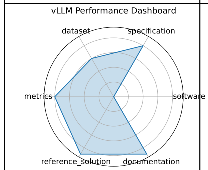
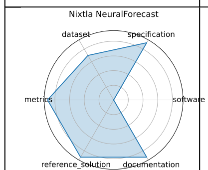
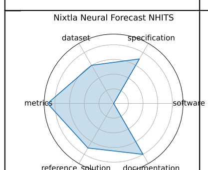
Continued on next page

| 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]⇒ |

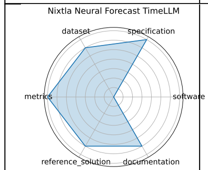
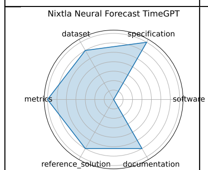
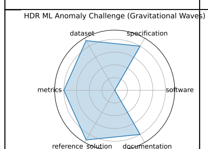
Continued on next page

| 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]⇒ |

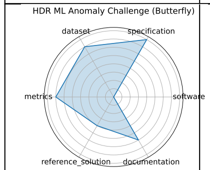
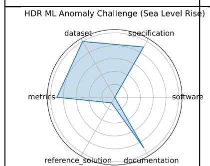
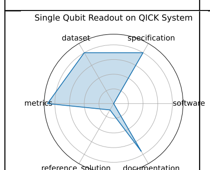
Continued on next page

| 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]⇒ |

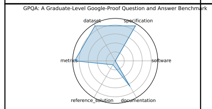
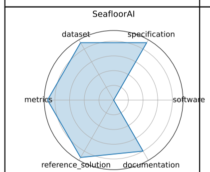
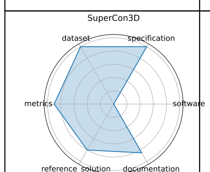
Continued on next page

| 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]⇒ |

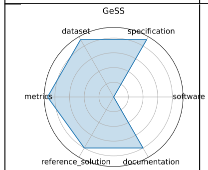
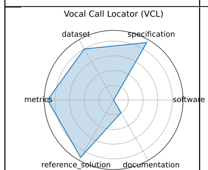
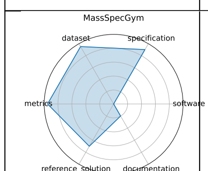
Continued on next page

| 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]⇒ |

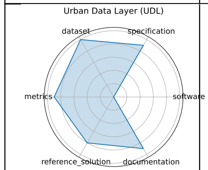
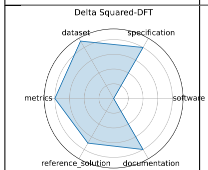
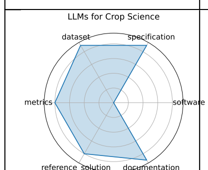
Continued on next page

| 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 multimodal 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, DiffCSP-SC | [36]⇒ |

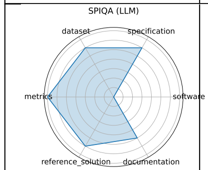
Continued on next page

| 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 machine-learning corrections to DFT using Delta Squared-trained models for reaction energies | density functional theory, Delta Squared-ML correction, reaction energetics, quantum chemistry | Regression | High-accuracy energy prediction, DFT correction | Mean Absolute Error (eV), Energy ranking accuracy | Delta Squared-ML correction networks, Kernel ridge regression | [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

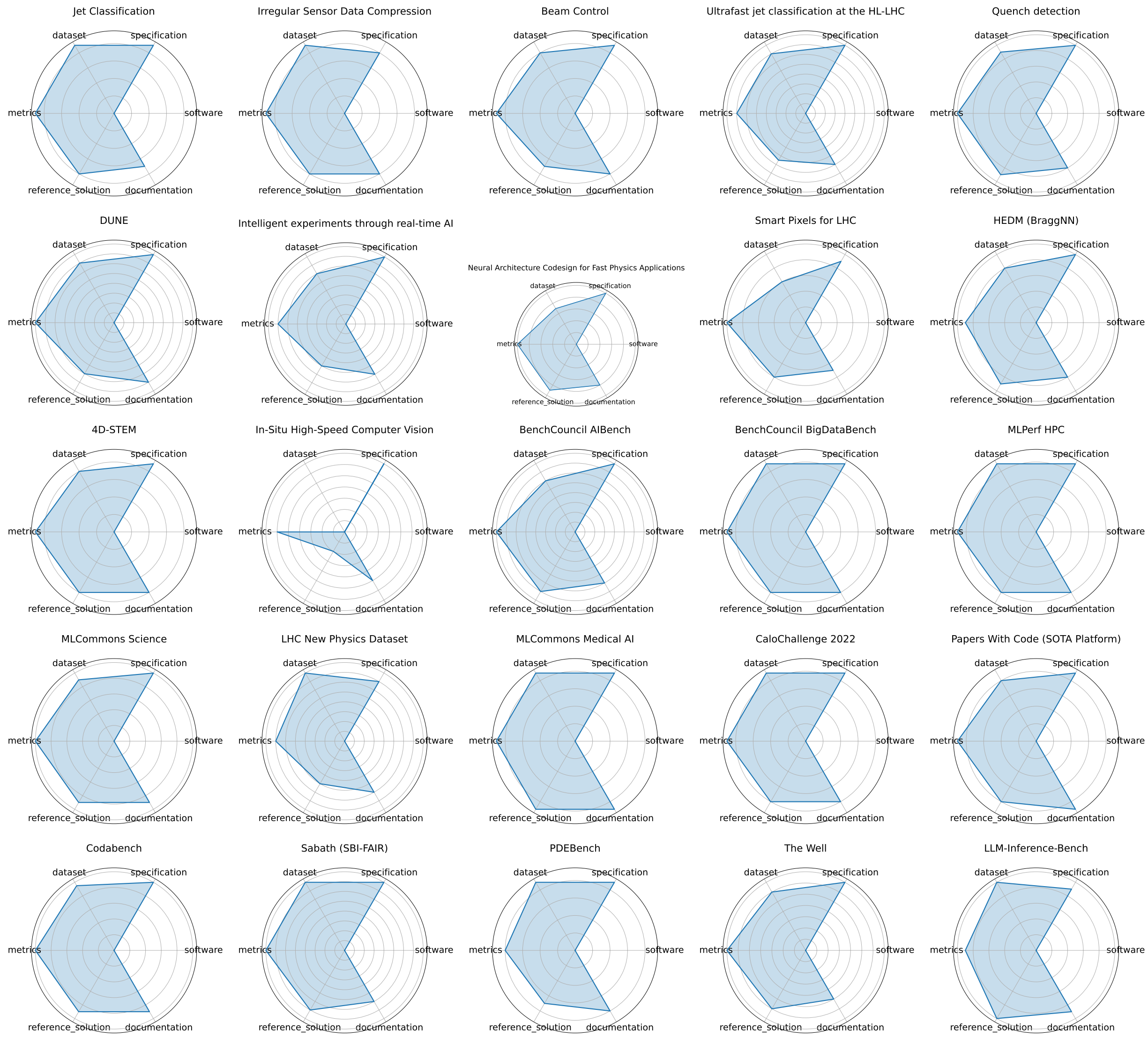



Figure 1: Radar chart overview (page 1)

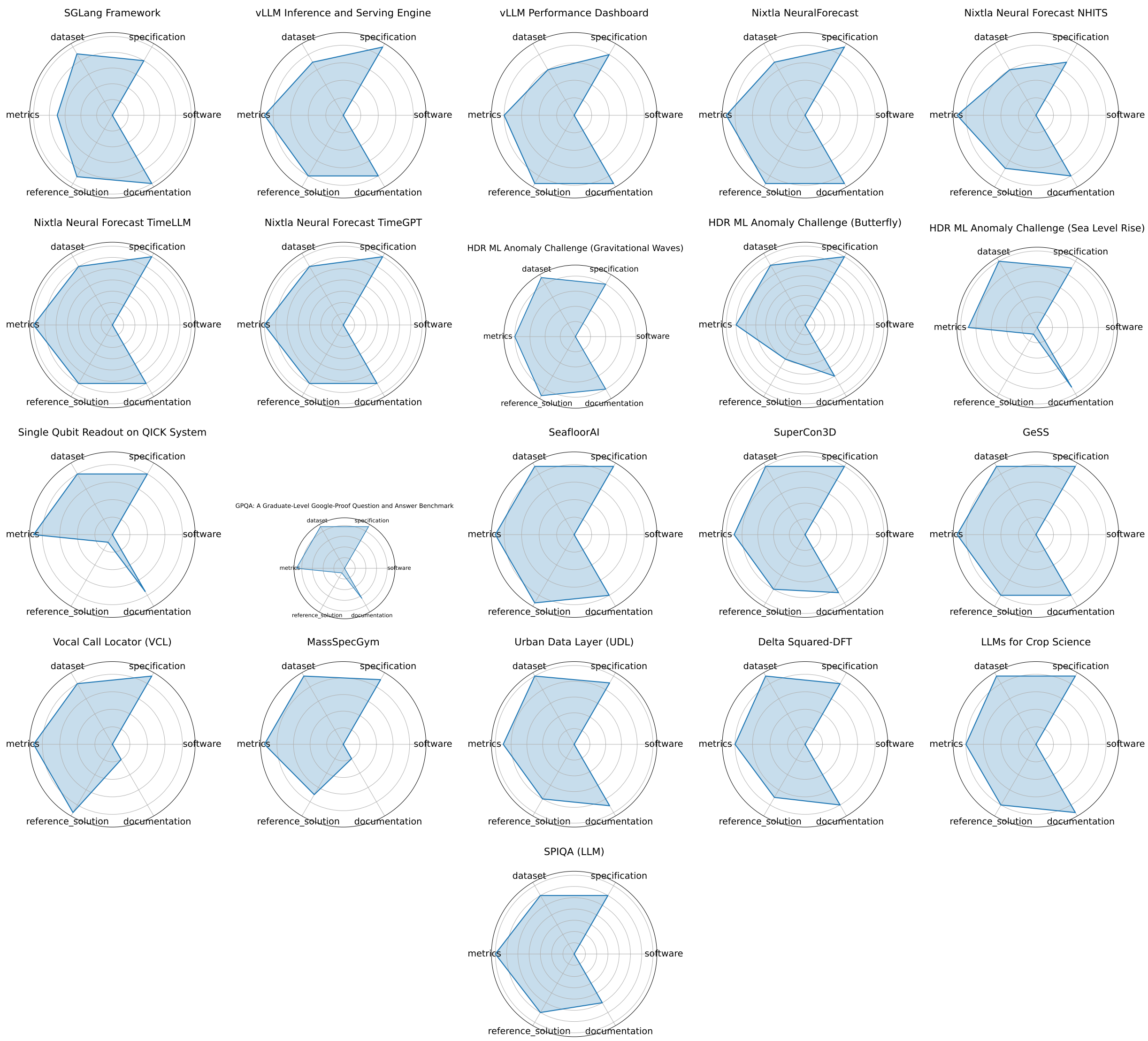



Figure 2: Radar chart overview (page 2)

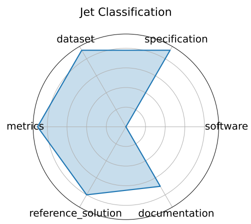
# 3 Benchmark Details

# 4 Jet Classification

**date:** 2024-05-01
**version:** TODO
**last_updated:** 2024-05
**expired:** unknown
**valid:** yes
**valid_date:** TODO
**url:** https://github.com/fastmachinelearning/fastml-science/tree/main/jet-classify
**doi:** TODO
**domain:** Particle Physics
**focus:** Real-time classification of particle jets using HL-LHC simulation features
**keywords:** - classification - real-time ML - jet tagging - QKeras
**summary:** This benchmark evaluates ML models for real-time classification of particle jets using high-level features derived from simulated LHC data. It includes both full-precision and quantized models optimized for FPGA deployment.
**licensing:** TODO
**task_types:** - Classification
**ai_capability_measured:** - Real-time inference - model compression performance
**metrics:** - Accuracy - AUC
**models:** - Keras DNN - QKeras quantized DNN
**ml_motif:** - Real-time
**type:** Benchmark
**ml_task:** - Supervised Learning
**solutions:** TODO
**notes:** Includes both float and quantized models using QKeras
**contact.name:** Jules Muhizi
**contact.email:** unknown
**datasets.links.name:** JetClass
**datasets.links.url:** https://zenodo.org/record/6619768
**results.links.name:** ChatGPT LLM
**results.links.url:** https://docs.google.com/document/d/1runrcij-eoH3_lgGZ8wm2z1YbL1Qf5cSNbVbHyWFDs4
**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]

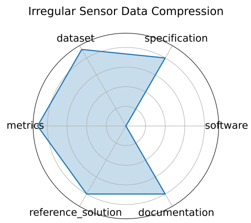
**Ratings:**



# 5 Irregular Sensor Data Compression

**date:** 2024-05-01
**version:** TODO
**last_updated:** 2024-05
**expired:** unknown
**valid:** yes
**valid_date:** TODO
**url:** https://github.com/fastmachinelearning/fastml-science/tree/main/sensor-data-compression
**doi:** TODO
**domain:** Particle Physics
**focus:** Real-time compression of sparse sensor data with autoencoders
**keywords:** - compression - autoencoder - sparse data - irregular sampling
**summary:** This benchmark addresses lossy compression of irregularly sampled sensor data from particle detectors using real-time autoencoder architectures, targeting latency-critical applications in physics experiments.
**licensing:** TODO
**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
**solutions:** TODO
**notes:** Based on synthetic but realistic physics sensor data
**contact.name:** Ben Hawks, Nhan Tran
**contact.email:** unknown
**datasets.links.name:** Custom synthetic irregular sensor dataset
**datasets.links.url:** https://github.com/fastmachinelearning/fastml-science/tree/main/sensor-data-compression
**results.links.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]

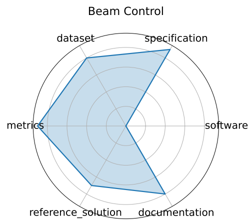
**Ratings:**



Irregular Sensor Data Compression
dataset
specification
metrics
software
reference_solution
documentation

# 6 Beam Control

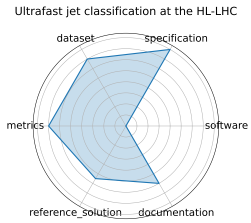
**date:** 2024-05-01
**version:** TODO
**last_updated:** 2024-05
**expired:** unknown
**valid:** yes
**valid_date:** TODO
**url:** https://github.com/fastmachinelearning/fastml-science/tree/main/beam-control
**doi:** TODO
**domain:** Accelerators and Magnets
**focus:** Reinforcement learning control of accelerator beam position
**keywords:** - RL - beam stabilization - control systems - simulation
**summary:** Beam Control explores real-time reinforcement learning strategies for maintaining stable beam trajectories in particle accelerators. The benchmark is based on the BOOSTR environment for accelerator simulation.
**licensing:** TODO
**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
**solutions:** TODO
**notes:** Environment defined, baseline RL implementation is in progress
**contact.name:** Ben Hawks, Nhan Tran
**contact.email:** unknown
**results.links.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 pretrained weights would improve usability.
**id:** beam_control
**Citations:** [2], [3]

**Ratings:** 
Beam Control
dataset
specification
metrics
software
reference_solution
documentation

# 7 Ultrafast jet classification at the HL-LHC

**date:** 2024-07-08
**version:** TODO
**last_updated:** 2024-07
**expired:** unknown
**valid:** yes
**valid_date:** TODO
**url:** https://arxiv.org/pdf/2402.01876
**doi:** TODO
**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
**summary:** Demonstrates three ML models (MLP, Deep Sets, Interaction Networks) optimized for FPGA deployment with O(100 ns) inference using quantized models and hls4ml, targeting real-time jet tagging in the L1 trigger environment at the high-luminosity LHC. Data is available on Zenodo DOI:10.5281/zenodo.3602260. :contentReference[oaicite:1]{index=1}
**licensing:** TODO
**task_types:** - Classification
**ai_capability_measured:** - Real-time inference under FPGA constraints
**metrics:** - Accuracy - Latency - Resource utilization
**models:** - MLP - Deep Sets - Interaction Network
**ml_motif:** - Real-time
**type:** Model
**ml_task:** - Supervised Learning
**solutions:** TODO
**notes:** Uses quantization-aware training; hardware synthesis evaluated via hls4ml
**contact.name:** Patrick Odagiu
**contact.email:** unknown
**datasets.links.name:** Zenodo dataset
**datasets.links.url:** https://zenodo.org/records/3602260
**results.links.name:** ChatGPT LLM
**results.links.url:** https://docs.google.com/document/d/1gDf1CIYtfmfZ9urv1jCRZMYz_3WwEETkugUC65OZBdw
**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:**



Ultrafast jet classification at the HL-LHC
dataset
specification
metrics
software
reference_solution
documentation

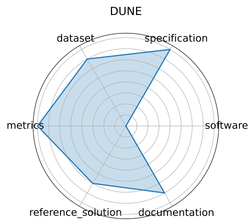
# 8 Quench detection

**date:** 2024-10-15
**version:** TODO
**last_updated:** 2024-10
**expired:** unknown
**valid:** yes
**valid_date:** TODO
**url:** https://indico.cern.ch/event/1387540/contributions/6153618/attachments/2948441/5182077/fast_ml_magnets_2024_final.pdf
**doi:** TODO
**domain:** Accelerators and Magnets
**focus:** Real-time detection of superconducting magnet quenches using ML
**keywords:** - quench detection - autoencoder - anomaly detection - real-time
**summary:** Exploration of real-time quench detection using unsupervised and RL approaches, combining multi-modal sensor data (BPM, power supply, acoustic), operating on kHz-MHz streams with anomaly detection and frequency-domain features. :contentReference[oaicite:2]{index=2}
**licensing:** TODO
**task_types:** - Anomaly detection - Quench localization
**ai_capability_measured:** - Real-time anomaly detection with multi-modal sensors
**metrics:** - ROC-AUC - Detection latency
**models:** - Autoencoder - RL agents (in development)
**ml_motif:** - Real-time, RL
**type:** Benchmark
**ml_task:** - Reinforcement + Unsupervised Learning
**solutions:** TODO
**notes:** Precursor detection in progress; multi-modal and dynamic weighting methods
**contact.name:** Maira Khan
**contact.email:** unknown
**datasets.links.name:** BPM and power supply data from BNL
**results.links.name:** ChatGPT LLM
**fair.reproducible:** in progress
**fair.benchmark_ready:** False
**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
**version:** TODO
**last_updated:** 2024-10
**expired:** unknown
**valid:** yes
**valid_date:** TODO
**url:** https://indico.fnal.gov/event/66520/contributions/301423/attachments/182439/250508/fast_ml_dunedaq_sonic_10_15_24.pdf
**doi:** TODO
**domain:** Particle Physics
**focus:** Real-time ML for DUNE DAQ time-series data
**keywords:** - DUNE - time-series - real-time - trigger
**summary:** Applying real-time ML methods to time-series data from DUNE detectors, exploring trigger-level anomaly detection and event selection with low latency constraints.
**licensing:** TODO
**task_types:** - Trigger selection - Time-series anomaly detection
**ai_capability_measured:** - Low-latency event detection
**metrics:** - Detection efficiency - Latency
**models:** - CNN - LSTM (planned)
**ml_motif:** - Real-time, Time-series
**type:** Benchmark (in progress)
**ml_task:** - Supervised Learning
**solutions:** TODO
**notes:** Prototype models demonstrated on SONIC platform
**contact.name:** Andrew J. Morgan
**contact.email:** unknown
**datasets.links.name:** DUNE SONIC data
**results.links.name:** ChatGPT LLM
**fair.reproducible:** in progress
**fair.benchmark_ready:** False
**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]

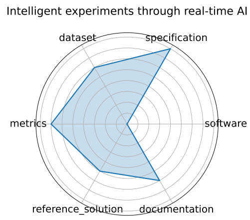
**Ratings:**


DUNE
dataset
specification
metrics
software
reference_solution
documentation

# 10 Intelligent experiments through real-time AI

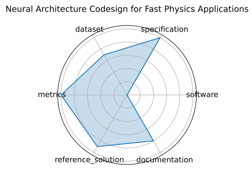
**date:** 2025-01-08
**version:** TODO
**last_updated:** 2025-01
**expired:** unknown
**valid:** yes
**valid_date:** TODO
**url:** https://arxiv.org/pdf/2501.04845
**doi:** TODO
**domain:** Instrumentation and Detectors; Nuclear Physics; Particle Physics
**focus:** Real-time FPGA-based triggering and detector control for sPHENIX and future EIC
**keywords:** - FPGA - Graph Neural Network - hls4ml - real-time inference - detector control
**summary:** Resaerch and Development demonstrator for real-time processing of high-rate tracking data from the sPHENIX detector (RHIC) and future EIC systems. Uses GNNs with hls4ml for FPGA-based trigger generation to identify rare events (heavy flavor, DIS electrons) within 10 micros latency. Demonstrated improved accuracy and latency on Alveo/FELIX platforms.
**licensing:** TODO
**task_types:** - Trigger classification - Detector control - Real-time inference
**ai_capability_measured:** - Low-latency GNN inference on FPGA
**metrics:** - Accuracy (charm and beauty detection) - Latency (micros) - Resource utilization (LUT/FF/BRAM/DSP)
**models:** - Bipartite Graph Network with Set Transformers (BGN-ST) - GarNet (edge-classifier)
**ml_motif:** - Real-time
**type:** Model
**ml_task:** - Supervised Learning
**solutions:** TODO
**notes:** Achieved ~97.4% accuracy for beauty decay triggers; sub-10 micros latency on Alveo U280; hit-based FPGA design via hls4ml and FlowGNN.
**contact.name:** Jakub Kvapil (lanl.gov)
**contact.email:** unknown
**datasets.links.name:** Internal simulated tracking data (sPHENIX and EIC DIS-electron tagger)
**results.links.name:** ChatGPT LLM
**fair.reproducible:** True
**fair.benchmark_ready:** False
**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:**


Intelligent experiments through real-time AI
dataset
specification
metrics
software
reference_solution
documentation

# 11 Neural Architecture Codesign for Fast Physics Applications

**date:** 2025-01-09
**version:** TODO
**last_updated:** 2025-01
**expired:** unknown
**valid:** yes
**valid_date:** TODO
**url:** https://arxiv.org/abs/2501.05515
**doi:** TODO
**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
**summary:** Introduces a two-stage neural architecture codesign (NAC) pipeline combining global and local search, quantization-aware training, and pruning to design efficient models for fast Bragg peak finding and jet classification, synthesized for FPGA deployment with hls4ml. Achieves >30x reduction in BOPs and sub-100 ns inference latency on FPGA.
**licensing:** TODO
**task_types:** - Classification - Peak finding
**ai_capability_measured:** - Hardware-aware model optimization; low-latency inference
**metrics:** - Accuracy - Latency - Resource utilization
**models:** - NAC-based BraggNN - NAC-optimized Deep Sets (jet)
**ml_motif:** - Real-time, Image/CV
**type:** Framework
**ml_task:** - Supervised Learning
**solutions:** TODO
**notes:** Demonstrated two case studies (materials science, HEP); pipeline and code open-sourced.
**contact.name:** Jason Weitz (UCSD), Nhan Tran (FNAL)
**contact.email:** unknown
**results.links.name:** ChatGPT LLM
**fair.reproducible:** Yes (nac-opt, hls4ml)
**fair.benchmark_ready:** False
**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 micros 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]

**Ratings:** 
Neural Architecture Codesign for Fast Physics Applications
dataset
specification
metrics
software
reference_solution
documentation

# 12 Smart Pixels for LHC

**date:** 2024-06-24

**version:** TODO

**last_updated:** 2024-06

**expired:** unknown

**valid:** yes

**valid_date:** TODO

**url:** https://arxiv.org/abs/2406.14860

**doi:** TODO

**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

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

**licensing:** TODO

**task_types:** - Image Classification - Data filtering

**ai_capability_measured:** - On-chip - low-power inference; data reduction

**metrics:** - Data rejection rate - Power per pixel

**models:** - 2-layer pixel NN

**ml_motif:** - Real-time, Image/CV

**type:** Benchmark

**ml_task:** - Image Classification

**solutions:** TODO

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

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

**contact.email:** unknown

**results.links.name:** ChatGPT LLM

**fair.reproducible:** True

**fair.benchmark_ready:** Yes (Zenodo:7331128)

**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.

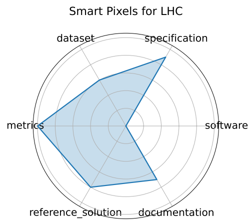
**ratings.documentation.rating:** 7.0

**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]

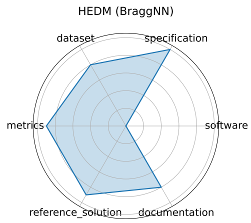
**Ratings:**


Smart Pixels for LHC
dataset
specification
metrics
software
reference_solution
documentation

# 13 HEDM (BraggNN)

**date:** 2023-10-03
**version:** TODO
**last_updated:** 2023-10
**expired:** unknown
**valid:** yes
**valid_date:** TODO
**url:** https://arxiv.org/abs/2008.08198
**doi:** TODO
**domain:** Material Science
**focus:** Fast Bragg peak analysis using deep learning in diffraction microscopy
**keywords:** - BraggNN - diffraction - peak finding - HEDM
**summary:** Uses BraggNN, a deep neural network, for rapid Bragg peak localization in high-energy diffraction microscopy, achieving about 13x speedup compared to Voigt-based methods while maintaining sub-pixel accuracy.
**licensing:** TODO
**task_types:** - Peak detection
**ai_capability_measured:** - High-throughput peak localization
**metrics:** - Localization accuracy - Inference time
**models:** - BraggNN
**ml_motif:** - Real-time, Image/CV
**type:** Framework
**ml_task:** - Peak finding
**solutions:** TODO
**notes:** Enables real-time HEDM workflows; basis for NAC case study.
**contact.name:** Jason Weitz (UCSD)
**contact.email:** unknown
**results.links.name:** ChatGPT LLM
**fair.reproducible:** True
**fair.benchmark_ready:** True
**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:**

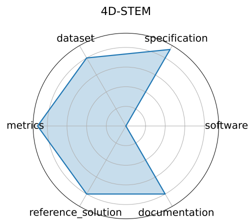


HEDM (BraggNN)
dataset
specification
metrics
software
reference_solution
documentation

# 14 4D-STEM

**date:** 2023-12-03
**version:** TODO
**last_updated:** 2023-12
**expired:** unknown
**valid:** yes
**valid_date:** TODO
**url:** https://openreview.net/pdf?id=7yt3N0o0W9
**doi:** TODO
**domain:** Material Science
**focus:** Real-time ML for scanning transmission electron microscopy
**keywords:** - 4D-STEM - electron microscopy - real-time - image processing
**summary:** Proposes ML methods for real-time analysis of 4D scanning transmission electron microscopy datasets; framework details in progress.
**licensing:** TODO
**task_types:** - Image Classification - Streamed data inference
**ai_capability_measured:** - Real-time large-scale microscopy inference
**metrics:** - Classification accuracy - Throughput
**models:** - CNN models (prototype)
**ml_motif:** - Real-time, Image/CV
**type:** Model
**ml_task:** - Image Classification
**solutions:** TODO
**notes:** In-progress; model design under development.
**contact.name:** unknown
**contact.email:** unknown
**results.links.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:** d-stem
**Citations:** [10]

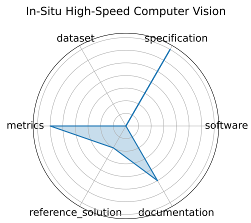
**Ratings:**



4D-STEM
dataset
specification
metrics
software
reference_solution
documentation

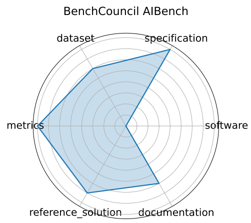
# 15 In-Situ High-Speed Computer Vision

**date:** 2023-12-05
**version:** TODO
**last_updated:** 2023-12
**expired:** unknown
**valid:** yes
**valid_date:** TODO
**url:** https://arxiv.org/abs/2312.00128
**doi:** TODO
**domain:** Fusion/Plasma
**focus:** Real-time image classification for in-situ plasma diagnostics
**keywords:** - plasma - in-situ vision - real-time ML
**summary:** Applies low-latency CNN models for image classification of plasma diagnostics streams; supports deployment on embedded platforms.
**licensing:** TODO
**task_types:** - Image Classification
**ai_capability_measured:** - Real-time diagnostic inference
**metrics:** - Accuracy - FPS
**models:** - CNN
**ml_motif:** - Real-time, Image/CV
**type:** Model
**ml_task:** - Image Classification
**solutions:** TODO
**notes:** Embedded/deployment details in progress.
**contact.name:** unknown
**contact.email:** unknown
**results.links.name:** ChatGPT LLM
**results.links.url:** https://docs.google.com/document/d/1EqkRHuQs1yQqMvZs_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:** 
In-Situ High-Speed Computer Vision
dataset
specification
metrics
software
reference_solution
documentation

## 16 BenchCouncil AIBench

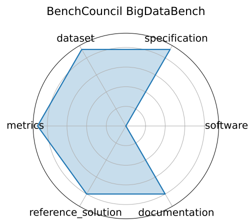
**date:** 2020-01-01
**version:** TODO
**last_updated:** 2020-01
**expired:** unknown
**valid:** yes
**valid_date:** TODO
**url:** https://www.benchcouncil.org/AIBench/
**doi:** TODO
**domain:** General
**focus:** End-to-end AI benchmarking across micro, component, and application levels
**keywords:** - benchmarking - AI systems - application-level evaluation
**summary:** AIBench is a comprehensive benchmark suite that evaluates AI workloads at different levels (micro, component, application) across hardware systems-covering image generation, object detection, translation, recommendation, video prediction, etc.
**licensing:** TODO
**task_types:** - Training - Inference - End-to-end AI workloads
**ai_capability_measured:** - System-level AI workload performance
**metrics:** - Throughput - Latency - Accuracy
**models:** - ResNet - BERT - GANs - Recommendation systems
**ml_motif:** - General
**type:** Benchmark
**ml_task:** - NA
**solutions:** TODO
**notes:** Covers scenario-distilling, micro, component, and end-to-end benchmarks.
**contact.name:** Wanling Gao (BenchCouncil)
**contact.email:** unknown
**results.links.name:** ChatGPT LLM
**fair.reproducible:** Yes
**fair.benchmark_ready:** Yes
**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:** 
BenchCouncil AIBench
dataset
specification
metrics
software
reference_solution
documentation

## 17 BenchCouncil BigDataBench

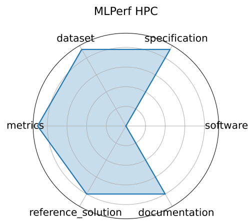
**date:** 2020-01-01
**version:** TODO
**last_updated:** 2020-01
**expired:** unknown
**valid:** yes
**valid_date:** TODO
**url:** https://www.benchcouncil.org/BigDataBench/
**doi:** TODO
**domain:** General
**focus:** Big data and AI benchmarking across structured, semi-structured, and unstructured data workloads
**keywords:** - big data - AI benchmarking - data analytics
**summary:** BigDataBench provides benchmarks for evaluating big data and AI workloads with realistic datasets (13 sources) and pipelines across analytics, graph, warehouse, NoSQL, streaming, and AI.
**licensing:** TODO
**task_types:** - Data preprocessing - Inference - End-to-end data pipelines
**ai_capability_measured:** - Data processing and AI model inference performance at scale
**metrics:** - Data throughput - Latency - Accuracy
**models:** - CNN - LSTM - SVM - XGBoost
**ml_motif:** - General
**type:** Benchmark
**ml_task:** - NA
**solutions:** TODO
**notes:** Built on eight data motifs; provides Hadoop, Spark, Flink, MPI implementations.
**contact.name:** Jianfeng Zhan (BenchCouncil)
**contact.email:** unknown
**results.links.name:** ChatGPT LLM
**results.links.url:** https://docs.google.com/document/d/1VFRxhR2G5A83S8PqKBrP99LLVgcCGvX2WW4vTtwxmQ4/edit?usp=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:**


BenchCouncil BigDataBench
dataset
specification
metrics
software
reference_solution
documentation

# 18 MLPerf HPC

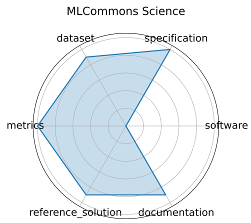
**date:** 2021-10-20
**version:** TODO
**last_updated:** 2021-10
**expired:** unknown
**valid:** yes
**valid_date:** TODO
**url:** https://github.com/mlcommons/hpc
**doi:** TODO
**domain:** Cosmology, Climate, Protein Structure, Catalysis
**focus:** Scientific ML training and inference on HPC systems
**keywords:** - HPC - training - inference - scientific ML
**summary:** MLPerf HPC introduces scientific model benchmarks (e.g., CosmoFlow, DeepCAM) aimed at large-scale HPC evaluation with >10x performance scaling through system-level optimizations.
**licensing:** TODO
**task_types:** - Training - Inference
**ai_capability_measured:** - Scaling efficiency - training time - model accuracy on HPC
**metrics:** - Training time - Accuracy - GPU utilization
**models:** - CosmoFlow - DeepCAM - OpenCatalyst
**ml_motif:** - HPC/inference, HPC/training
**type:** Framework
**ml_task:** - NA
**solutions:** TODO
**notes:** Shared framework with MLCommons Science; reference implementations included.
**contact.name:** Steven Farrell (MLCommons)
**contact.email:** unknown
**results.links.name:** ChatGPT LLM
**fair.reproducible:** Yes
**fair.benchmark_ready:** Yes
**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:** 
MLPerf HPC
dataset
specification
metrics
software
reference_solution
documentation

## 19 MLCommons Science

**date:** 2023-06-01
**version:** TODO
**last_updated:** 2023-06
**expired:** unknown
**valid:** yes
**valid_date:** TODO
**url:** https://github.com/mlcommons/science
**doi:** TODO
**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
**summary:** MLCommons Science assembles benchmark tasks with datasets, targets, and implementations across earthquake forecasting, satellite imagery, drug screening, electron microscopy, and CFD to drive scientific ML reproducibility.
**licensing:** TODO
**task_types:** - Time-series analysis - Image classification - Simulation surrogate modeling
**ai_capability_measured:** - Inference accuracy - simulation speed-up - generalization
**metrics:** - MAE - Accuracy - Speedup vs simulation
**models:** - CNN - GNN - Transformer
**ml_motif:** - Time-series, Image/CV, HPC/inference
**type:** Framework
**ml_task:** - NA
**solutions:** TODO
**notes:** Joint national-lab effort under Apache-2.0 license.
**contact.name:** MLCommons Science Working Group
**contact.email:** unknown
**results.links.name:** ChatGPT LLM
**fair.reproducible:** Yes
**fair.benchmark_ready:** Yes
**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]

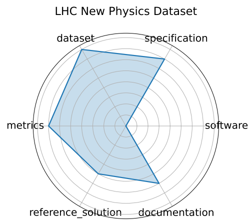
**Ratings:**


MLCommons Science
dataset
specification
metrics
software
reference_solution
documentation

# 20 LHC New Physics Dataset

**date:** 2021-07-05
**version:** TODO
**last_updated:** 2021-07
**expired:** unknown
**valid:** yes
**valid_date:** TODO
**url:** https://arxiv.org/pdf/2107.02157
**doi:** TODO
**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
**summary:** A dataset of proton-proton collision events emulating a 40 MHz real-time data stream from LHC detectors, pre-filtered on electron or muon presence. Designed for unsupervised new-physics detection algorithms under latency/bandwidth constraints.
**licensing:** TODO
**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
**solutions:** TODO
**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
**datasets.links.name:** Zenodo stores, background + 3 black-box signal sets. 1M events each
**results.links.name:** ChatGPT LLM
**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:**


LHC New Physics Dataset
dataset
specification
software
metrics
reference_solution
documentation

# 21 MLCommons Medical AI

**date:** 2023-07-17

**version:** TODO

**last_updated:** 2023-07

**expired:** unknown

**valid:** yes

**valid_date:** TODO

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

**doi:** TODO

**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

**summary:** The MLCommons Medical AI working group develops benchmarks, best practices, and platforms (MedPerf, GaNDLF, COFE) to accelerate robust, privacy-preserving AI development for healthcare. MedPerf enables federated testing of clinical models on diverse datasets, improving generalizability and equity while keeping data onsite :contentReference[oaicite:1]{index=1}.

**licensing:** TODO

**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

**solutions:** TODO

**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

**datasets.links.name:** Multi-institutional clinical datasets, radiology

**results.links.name:** ChatGPT LLM

**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.

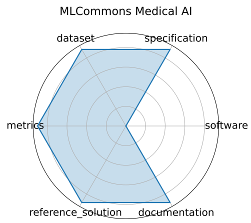
**ratings.documentation.rating:** 9.0

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

**id:** mlcommons_medical_ai

**Citations:** [17]

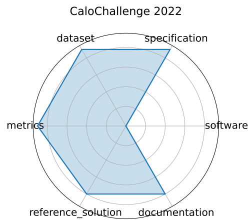
**Ratings:**


MLCommons Medical AI
dataset
specification
metrics
software
reference_solution
documentation

## 22 CaloChallenge 2022

**date:** 2024-10-28
**version:** TODO
**last_updated:** 2024-10
**expired:** unknown
**valid:** yes
**valid_date:** TODO
**url:** http://arxiv.org/abs/2410.21611
**doi:** TODO
**domain:** LHC Calorimeter; Particle Physics
**focus:** Fast generative-model-based calorimeter shower simulation evaluation
**keywords:** - calorimeter simulation - generative models - surrogate modeling - LHC - fast simulation
**summary:** The Fast Calorimeter Simulation Challenge 2022 assessed 31 generative-model submissions (VAEs, GANs, Flows, Diffusion) on four calorimeter shower datasets; benchmarking shower quality, generation speed, and model complexity :contentReference[oaicite:3]{index=3}.
**licensing:** TODO
**task_types:** - Surrogate modeling
**ai_capability_measured:** - Simulation fidelity - speed - efficiency
**metrics:** - Histogram similarity - Classifier AUC - Generation latency
**models:** - VAE variants - GAN variants - Normalizing flows - Diffusion models
**ml_motif:** - Surrogate
**type:** Dataset
**ml_task:** - Surrogate Modeling
**solutions:** TODO
**notes:** The most comprehensive survey to date on ML-based calorimeter simulation; 31 submissions over different dataset sizes.
**contact.name:** Claudius Krause (CaloChallenge Lead)
**contact.email:** unknown
**datasets.links.name:** Four LHC calorimeter shower datasets
**datasets.links.url:** various voxel resolutions
**results.links.name:** ChatGPT LLM
**fair.reproducible:** Yes
**fair.benchmark_ready:** Yes
**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:**

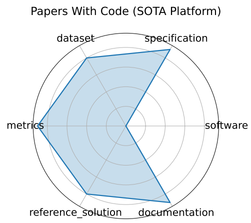


CaloChallenge 2022
dataset
specification
metrics
software
reference_solution
documentation

# 23 Papers With Code (SOTA Platform)

**date:** ongoing
**version:** TODO
**last_updated:** 2025-06
**expired:** unknown
**valid:** yes
**valid_date:** TODO
**url:** https://paperswithcode.com/sota
**doi:** TODO
**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
**summary:** Papers With Code (PWC) aggregates benchmark suites, tasks, and code across ML research: 12,423 benchmarks, 5,358 unique tasks, and 154,766 papers with code links. It tracks SOTA metrics and fosters reproducibility.
**licensing:** TODO
**task_types:** - Multiple (Classification, Detection, NLP, etc.)
**ai_capability_measured:** - Model performance across tasks (accuracy - F1 - BLEU - etc.)
**metrics:** - Task-specific (Accuracy, F1, BLEU, etc.)
**models:** - All published models with code
**ml_motif:** - Multiple
**type:** Platform
**ml_task:** - Multiple
**solutions:** TODO
**notes:** Community-driven open platform; automatic data extraction and versioning.
**contact.name:** Papers With Code Team
**contact.email:** unknown
**results.links.name:** ChatGPT LLM
**fair.reproducible:** Yes
**fair.benchmark_ready:** Yes
**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:**

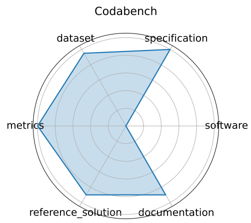


Papers With Code (SOTA Platform)
dataset
specification
metrics
software
reference_solution
documentation

# 24 Codabench

**date:** 2022-01-01
**version:** TODO
**last_updated:** 2025-03
**expired:** unknown
**valid:** yes
**valid_date:** TODO
**url:** https://www.codabench.org/
**doi:** TODO
**domain:** General ML; Multiple
**focus:** Open-source platform for organizing reproducible AI benchmarks and competitions
**keywords:** - benchmark platform - code submission - competitions - meta-benchmark
**summary:** Codabench (successor to CodaLab) is a flexible, easy-to-use, reproducible API platform for hosting AI benchmarks and code-submission challenges. It supports custom scoring, inverted benchmarks, and scalable public or private queues :contentReference[oaicite:1]{index=1}.
**licensing:** TODO
**task_types:** - Multiple
**ai_capability_measured:** - Model reproducibility - performance across datasets
**metrics:** - Submission count - Leaderboard ranking - Task-specific metrics
**models:** - Arbitrary code submissions
**ml_motif:** - Multiple
**type:** Platform
**ml_task:** - Multiple
**solutions:** TODO
**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.links.name:** ChatGPT LLM
**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]

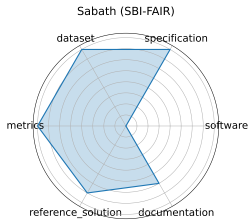
**Ratings:**


Codabench
dataset
specification
metrics
software
reference_solution
documentation

# 25 Sabath (SBI-FAIR)

**date:** 2021-09-27
**version:** TODO
**last_updated:** 2023-07
**expired:** unknown
**valid:** yes
**valid_date:** TODO
**url:** https://sbi-fair.github.io/docs/software/sabath/
**doi:** TODO
**domain:** Systems; Metadata
**focus:** FAIR metadata framework for ML-driven surrogate workflows in HPC systems
**keywords:** - meta-benchmark - metadata - HPC - surrogate modeling
**summary:** Sabath is a metadata framework from the SBI-FAIR group (UTK, Argonne, Virginia) facilitating FAIR-compliant benchmarking and surrogate execution logging across HPC systems :contentReference[oaicite:3]{index=3}.
**licensing:** TODO
**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
**solutions:** TODO
**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.links.name:** ChatGPT LLM
**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]

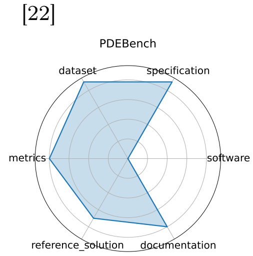
**Ratings:**


Sabath (SBI-FAIR)
dataset
specification
metrics
software
reference_solution
documentation

# 26 PDEBench

**date:** 2022-10-13
**version:** TODO
**last_updated:** 2025-05
**expired:** unknown
**valid:** yes
**valid_date:** TODO
**url:** https://github.com/pdebench/PDEBench
**doi:** TODO
**domain:** CFD; Weather Modeling
**focus:** Benchmark suite for ML-based surrogates solving time-dependent PDEs
**keywords:** - PDEs - CFD - scientific ML - surrogate modeling - NeurIPS
**summary:** PDEBench offers forward/inverse PDE tasks with large ready-to-use datasets and baselines (FNO, U-Net, PINN), packaged via a unified API. It won the SimTech Best Paper Award 2023 :contentReference[oaicite:5]{index=5}.
**licensing:** TODO
**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
**solutions:** TODO
**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.links.name:** ChatGPT LLM
**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]

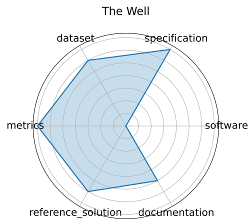
**Ratings:**



# 27 The Well

**date:** 2024-12-03
**version:** TODO
**last_updated:** 2025-06
**expired:** unknown
**valid:** yes
**valid_date:** TODO
**url:** https://polymathic-ai.org/the_well/
**doi:** TODO
**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
**summary:** A 15 TB collection of ML-ready physics simulation datasets (HDF5), covering 16 domains-from biology to astrophysical magnetohydrodynamic simulations-with unified API and metadata. Ideal for training surrogate and foundation models on scientific data. :contentReference[oaicite:1]{index=1}
**licensing:** TODO
**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
**solutions:** TODO
**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
**datasets.links.name:** 16 simulation datasets
**datasets.links.url:** HDF5) via PyPI/GitHub
**results.links.name:** ChatGPT LLM
**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:**


The Well
dataset
specification
metrics
software
reference_solution
documentation

# 28 LLM-Inference-Bench

**date:** 2024-10-31

**version:** TODO

**last_updated:** 2024-11

**expired:** unknown

**valid:** yes

**valid_date:** TODO

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

**doi:** TODO

**domain:** LLM; HPC/inference

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

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

**summary:** A suite evaluating inference performance of LLMs (LLaMA, Mistral, Qwen) across diverse accelerators (NVIDIA, AMD, Intel, SambaNova) and frameworks (vLLM, DeepSpeed-MII, etc.), with an interactive dashboard and per-platform metrics. :contentReference[oaicite:3]{index=3}

**licensing:** TODO

**task_types:** - Inference Benchmarking

**ai_capability_measured:** - Inference throughput - latency - hardware utilization

**metrics:** - Token throughput (tok/s) - Latency - Framework-hardware mix performance

**models:** - LLaMA-2-7B - LLaMA-2-70B - Mistral-7B - Qwen-7B

**ml_motif:** - HPC/inference

**type:** Dataset

**ml_task:** - Inference Benchmarking

**solutions:** TODO

**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.links.name:** ChatGPT LLM

**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.

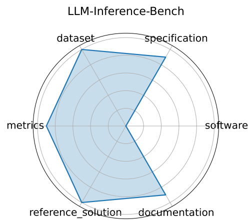
**ratings.documentation.rating:** 9.0

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

**id:** llm-inference-bench

**Citations:** [24]

**Ratings:**



LLM-Inference-Bench
dataset
specification
metrics
software
reference_solution
documentation

# 29 SGLang Framework

**date:** 2023-12-12

**version:** TODO

**last_updated:** 2025-06

**expired:** unknown

**valid:** yes

**valid_date:** TODO

**url:** https://github.com/sgl-project/sglang/tree/main/benchmark

**doi:** TODO

**domain:** LLM Vision

**focus:** Fast serving framework for LLMs and vision-language models

**keywords:** - LLM serving - vision-language - RadixAttention - performance - JSON decoding

**summary:** A high-performance open-source serving framework combining efficient backend runtime (RadixAttention, batching, quantization) and expressive frontend language, boosting LLM/VLM inference throughput up to ~3x over alternatives. :contentReference[oaicite:5]{index=5}

**licensing:** TODO

**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

**solutions:** TODO

**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:** unknown

**datasets.links.name:** Benchmark configs

**results.links.name:** ChatGPT LLM

**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.

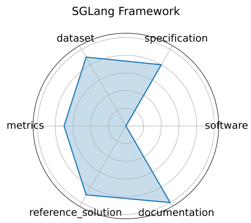
**ratings.documentation.rating:** 10.0

**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:**


SGLang Framework
dataset
specification
metrics
software
reference_solution
documentation

# 30 vLLM Inference and Serving Engine

**date:** 2023-09-12

**version:** TODO

**last_updated:** 2025-06

**expired:** unknown

**valid:** yes

**valid_date:** TODO

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

**doi:** TODO

**domain:** LLM; HPC/inference

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

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

**summary:** vLLM is a fast, high-throughput, memory-efficient inference and serving engine for large language models, featuring PagedAttention, continuous batching, and support for quantized and pipelined model execution. Benchmarks compare it to TensorRT-LLM, SGLang, and others. :contentReference[oaicite:1]{index=1}

**licensing:** TODO

**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

**solutions:** TODO

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

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

**contact.email:** unknown

**results.links.name:** ChatGPT LLM

**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.

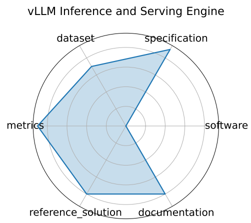
**ratings.documentation.rating:** 8.0

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

**id:** vllm_inference_and_serving_engine

**Citations:** [26]

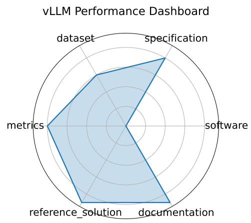
**Ratings:**


vLLM Inference and Serving Engine
dataset
specification
metrics
software
reference_solution
documentation

# 31 vLLM Performance Dashboard

**date:** 2022-06-22
**version:** TODO
**last_updated:** 2025-01
**expired:** unknown
**valid:** yes
**valid_date:** TODO
**url:** https://simon-mo-workspace.observablehq.cloud/vllm-dashboard-v0/
**doi:** TODO
**domain:** LLM; HPC/inference
**focus:** Interactive dashboard showing inference performance of vLLM
**keywords:** - Dashboard - Throughput visualization - Latency analysis - Metric tracking
**summary:** A live visual dashboard for vLLM showcasing throughput, latency, and other inference metrics across models and hardware configurations.
**licensing:** TODO
**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
**solutions:** TODO
**notes:** Built using ObservableHQ; integrates live data from vLLM benchmarks. The URL requires a login to access the content.
**contact.name:** Simon Mo
**contact.email:** unknown
**results.links.name:** ChatGPT LLM
**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:**

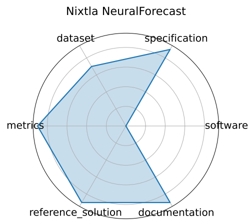


vLLM Performance Dashboard
dataset
specification
metrics
software
reference_solution
documentation

# 32 Nixtla NeuralForecast

**date:** 2022-04-01
**version:** TODO
**last_updated:** 2025-06
**expired:** unknown
**valid:** yes
**valid_date:** TODO
**url:** https://github.com/Nixtla/neuralforecast
**doi:** TODO
**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
**summary:** NeuralForecast offers scalable, user-friendly implementations of over 30 neural forecasting models (NBEATS, NHITS, TFT, DeepAR, etc.), emphasizing quality, usability, interpretability, and performance.
**licensing:** TODO
**task_types:** - Time-series forecasting
**ai_capability_measured:** - Forecast accuracy - interpretability - speed
**metrics:** - RMSE - MAPE - CRPS
**models:** - NBEATS - NHITS - TFT - DeepAR
**ml_motif:** - Time-series
**type:** Platform
**ml_task:** - Forecasting
**solutions:** TODO
**notes:** AutoModel supports hyperparameter tuning and distributed execution via Ray and Optuna. First official NHITS implementation. contentReference oaicite:4 ndex=4
**contact.name:** Kin G. Olivares (Nixtla)
**contact.email:** unknown
**results.links.name:** ChatGPT LLM
**fair.reproducible:** Yes
**fair.benchmark_ready:** Yes
**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]

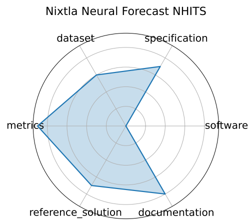
**Ratings:**


Nixtla NeuralForecast
dataset
specification
metrics
software
reference_solution
documentation

# 33 Nixtla Neural Forecast NHITS

**date:** 2023-06-01
**version:** TODO
**last_updated:** 2025-06
**expired:** unknown
**valid:** yes
**valid_date:** TODO
**url:** https://github.com/Nixtla/neuralforecast
**doi:** TODO
**domain:** Time-series; General ML
**focus:** Official NHITS implementation for long-horizon time series forecasting
**keywords:** - NHITS - long-horizon forecasting - neural interpolation - time-series
**summary:** NHITS (Neural Hierarchical Interpolation for Time Series) is a state-of-the-art model that improved accuracy by ~25% and reduced compute by 50x compared to Transformer baselines, using hierarchical interpolation and multi-rate sampling :contentReference[oaicite:1]{index=1}.
**licensing:** TODO
**task_types:** - Time-series forecasting
**ai_capability_measured:** - Accuracy - compute efficiency for long series
**metrics:** - RMSE - MAPE
**models:** - NHITS
**ml_motif:** - Time-series
**type:** Platform
**ml_task:** - Forecasting
**solutions:** TODO
**notes:** Official implementation in NeuralForecast, included since its AAAI 2023 release.
**contact.name:** Kin G. Olivares (Nixtla)
**contact.email:** unknown
**datasets.links.name:** Standard forecast datasets, M4
**results.links.name:** ChatGPT LLM
**fair.reproducible:** Yes
**fair.benchmark_ready:** Yes
**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:**

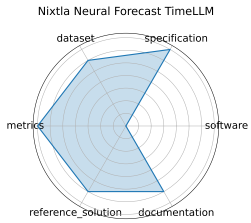


Nixtla Neural Forecast NHITS
dataset
specification
metrics
software
reference_solution
documentation

## 34 Nixtla Neural Forecast TimeLLM

**date:** 2023-10-03
**version:** TODO
**last_updated:** 2025-06
**expired:** unknown
**valid:** yes
**valid_date:** TODO
**url:** https://github.com/Nixtla/neuralforecast
**doi:** TODO
**domain:** Time-series; General ML
**focus:** Reprogramming LLMs for time series forecasting
**keywords:** - Time-LLM - language model - time-series - reprogramming
**summary:** Time-LLM uses reprogramming layers to adapt frozen LLMs for time series forecasting, treating forecasting as a language task :contentReference[oaicite:2]{index=2}.
**licensing:** TODO
**task_types:** - Time-series forecasting
**ai_capability_measured:** - Model reuse via LLM - few-shot forecasting
**metrics:** - RMSE - MAPE
**models:** - Time-LLM
**ml_motif:** - Time-series
**type:** Platform
**ml_task:** - Forecasting
**solutions:** TODO
**notes:** Fully open-source; transforms forecasting using LLM text reconstruction.
**contact.name:** Ming Jin (Nixtla)
**contact.email:** unknown
**datasets.links.name:** Standard forecast datasets, M4
**results.links.name:** ChatGPT LLM
**fair.reproducible:** Yes
**fair.benchmark_ready:** Yes
**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]

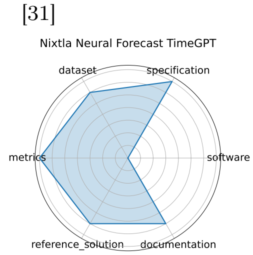
**Ratings:**



Nixtla Neural Forecast TimeLLM
dataset
specification
metrics
software
reference_solution
documentation

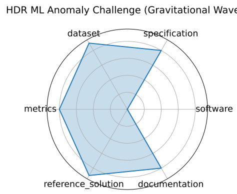
# 35 Nixtla Neural Forecast TimeGPT

**date:** 2023-10-05
**version:** TODO
**last_updated:** 2025-06
**expired:** unknown
**valid:** yes
**valid_date:** TODO
**url:** https://github.com/Nixtla/neuralforecast
**doi:** TODO
**domain:** Time-series; General ML
**focus:** Time-series foundation model "TimeGPT" for forecasting and anomaly detection
**keywords:** - TimeGPT - foundation model - time-series - generative model
**summary:** TimeGPT is a transformer-based generative pretrained model on 100B+ time series data for zero-shot forecasting and anomaly detection via API :contentReference[oaicite:3]{index=3}.
**licensing:** TODO
**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
**solutions:** TODO
**notes:** Offered via Nixtla API and Azure Studio; enterprise-grade support available.
**contact.name:** Azul Garza (Nixtla)
**contact.email:** unknown
**results.links.name:** ChatGPT LLM
**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)

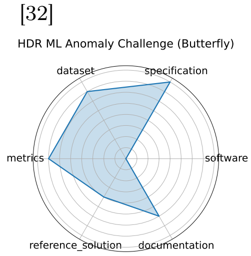
**date:** 2025-03-03
**version:** TODO
**last_updated:** 2025-03
**expired:** unknown
**valid:** yes
**valid_date:** TODO
**url:** https://www.codabench.org/competitions/2626/
**doi:** TODO
**domain:** Astrophysics; Time-series
**focus:** Detecting anomalous gravitational-wave signals from LIGO/Virgo datasets
**keywords:** - anomaly detection - gravitational waves - astrophysics - time-series
**summary:** A benchmark for detecting anomalous transient gravitational-wave signals, including "unknown-unknowns," using preprocessed LIGO time-series at 4096 Hz. Competitors submit inference models on Codabench for continuous 50 ms segments from dual interferometers. :contentReference[oaicite:1]{index=1}
**licensing:** TODO
**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
**solutions:** TODO
**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.links.name:** ChatGPT LLM
**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]

**Ratings:** 
HDR ML Anomaly Challenge (Gravitational Waves)
dataset
specification
metrics
software
reference_solution
documentation

# 37 HDR ML Anomaly Challenge (Butterfly)

**date:** 2025-03-03
**version:** TODO
**last_updated:** 2025-03
**expired:** unknown
**valid:** yes
**valid_date:** TODO
**url:** https://www.codabench.org/competitions/3764/
**doi:** TODO
**domain:** Genomics; Image/CV
**focus:** Detecting hybrid butterflies via image anomaly detection in genomic-informed dataset
**keywords:** - anomaly detection - computer vision - genomics - butterfly hybrids
**summary:** Image-based challenge for detecting butterfly hybrids in microscopy-driven species data. Participants evaluate models on Codabench using image segmentation/classification. :contentReference[oaicite:3]{index=3}
**licensing:** TODO
**task_types:** - Anomaly detection
**ai_capability_measured:** - Hybrid detection in biological systems
**metrics:** - Classification accuracy - F1 score
**models:** - CNN-based detectors
**ml_motif:** - Image/CV
**type:** Dataset
**ml_task:** - Anomaly detection
**solutions:** TODO
**notes:** Hybrid detection benchmarks hosted on Codabench. :contentReference[oaicite:4]{index=4}
**contact.name:** Imageomics/HDR Team
**contact.email:** unknown
**results.links.name:** ChatGPT LLM
**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]

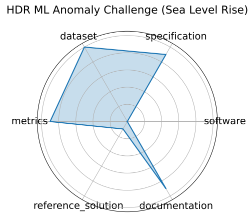
**Ratings:**



# 38 HDR ML Anomaly Challenge (Sea Level Rise)

**date:** 2025-03-03
**version:** TODO
**last_updated:** 2025-03
**expired:** unknown
**valid:** yes
**valid_date:** TODO
**url:** https://www.codabench.org/competitions/3223/
**doi:** TODO
**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
**summary:** A challenge combining North Atlantic sea-level time-series and satellite imagery to detect flooding anomalies. Models submitted via Codabench. :contentReference[oaicite:5]{index=5}
**licensing:** TODO
**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
**solutions:** TODO
**notes:** Sponsored by NSF HDR; integrates sensor and satellite data. :contentReference[oaicite:6]{index=6}
**contact.name:** HDR A3D3 Team
**contact.email:** unknown
**results.links.name:** ChatGPT LLM
**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]

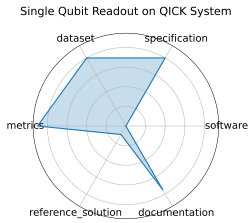
**Ratings:**



# 39 Single Qubit Readout on QICK System

**date:** 2025-01-24
**version:** TODO
**last_updated:** 2025-02
**expired:** unknown
**valid:** yes
**valid_date:** TODO
**url:** https://github.com/fastmachinelearning/ml-quantum-readout
**doi:** TODO
**domain:** Quantum Computing
**focus:** Real-time single-qubit state classification using FPGA firmware
**keywords:** - qubit readout - hls4ml - FPGA - QICK
**summary:** Implements real-time ML models for single-qubit readout on the Quantum Instrumentation Control Kit (QICK), using hls4ml to deploy quantized neural networks on RFSoC FPGAs. Offers high-fidelity, low-latency quantum state discrimination. :contentReference[oaicite:0]{index=0}
**licensing:** TODO
**task_types:** - Classification
**ai_capability_measured:** - Single-shot fidelity - inference latency
**metrics:** - Accuracy - Latency
**models:** - hls4ml quantized NN
**ml_motif:** - Real-time
**type:** Benchmark
**ml_task:** - Supervised Learning
**solutions:** TODO
**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:** unknown
**datasets.links.name:** Zenodo: ml-quantum-readout dataset
**datasets.links.url:** zenodo.org/records/14427490
**results.links.name:** ChatGPT LLM
**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:**

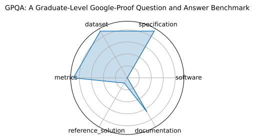


Single Qubit Readout on QICK System
dataset
specification
metrics
software
reference_solution
documentation

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

**date:** 2023-11-20
**version:** TODO
**last_updated:** 2023-11
**expired:** unknown
**valid:** yes
**valid_date:** TODO
**url:** https://arxiv.org/abs/2311.12022
**doi:** TODO
**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
**summary:** Contains 448 challenging questions written by domain experts, with expert accuracy at 65% (74% discounting clear errors) and non-experts reaching just 34%. GPT-4 baseline scores ~39%-designed for scalable oversight evaluation. :contentReference[oaicite:2]{index=2}
**licensing:** TODO
**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
**solutions:** TODO
**notes:** Google-proof, supports oversight research.
**contact.name:** David Rein (NYU)
**contact.email:** unknown
**datasets.links.name:** GPQA dataset
**datasets.links.url:** zip/HuggingFace
**results.links.name:** ChatGPT LLM
**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]

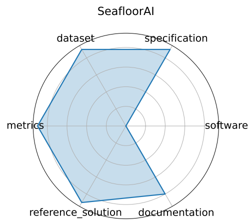
**Ratings:**


GPQA: A Graduate-Level Google-Proof Question and Answer Benchmark
dataset
specification
metrics
software
reference_solution
documentation

# 41 SeafloorAI

**date:** 2024-12-13
**version:** TODO
**last_updated:** 2024-12
**expired:** unknown
**valid:** yes
**valid_date:** TODO
**url:** https://neurips.cc/virtual/2024/poster/97432
**doi:** TODO
**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
**summary:** A first-of-its-kind dataset covering 17,300 sq.km of seafloor with 696K sonar images, 827K segmentation masks, and 696K natural-language descriptions plus ~7M QA pairs-designed for both vision and language-based ML models in marine science :contentReference[oaicite:1]{index=1}.
**licensing:** TODO
**task_types:** - Image segmentation - Vision-language QA
**ai_capability_measured:** - Geospatial understanding - multimodal reasoning
**metrics:** - Segmentation pixel accuracy - QA accuracy
**models:** - SegFormer - ViLT-style multimodal models
**ml_motif:** - Vision-Language
**type:** Dataset
**ml_task:** - Segmentation, QA
**solutions:** TODO
**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
**datasets.links.name:** Sonar imagery + annotations
**datasets.links.url:** ~15 TB
**results.links.name:** ChatGPT LLM
**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]

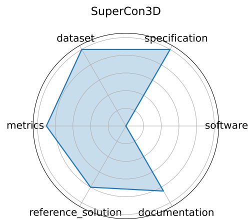
**Ratings:**


SeafloorAI
dataset
specification
metrics
software
reference_solution
documentation

## 42 SuperCon3D

**date:** 2024-12-13
**version:** TODO
**last_updated:** 2024-12
**expired:** unknown
**valid:** yes
**valid_date:** TODO
**url:** https://neurips.cc/virtual/2024/poster/97553
**doi:** TODO
**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
**summary:** SuperCon3D introduces 3D crystal structures with associated critical temperatures (Tc) and two deep-learning models: SODNet (equivariant graph model) and DiffCSP-SC (diffusion generator) designed to screen and synthesize high-Tc candidates :contentReference[oaicite:3]{index=3}.
**licensing:** TODO
**task_types:** - Regression (Tc prediction) - Generative modeling
**ai_capability_measured:** - Structure-to-property prediction - structure generation
**metrics:** - MAE (Tc) - Validity of generated structures
**models:** - SODNet - DiffCSP-SC
**ml_motif:** - Materials Modeling
**type:** Dataset + Models
**ml_task:** - Regression, Generation
**solutions:** TODO
**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.links.name:** ChatGPT LLM
**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]

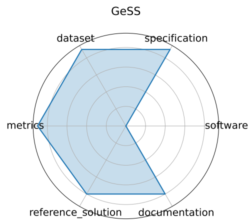
**Ratings:**


SuperCon3D
dataset
specification
metrics
software
reference_solution
documentation

# 43 GeSS

**date:** 2024-12-13
**version:** TODO
**last_updated:** 2024-12
**expired:** unknown
**valid:** yes
**valid_date:** TODO
**url:** https://neurips.cc/virtual/2024/poster/97816
**doi:** TODO
**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
**summary:** GeSS provides 30 benchmark scenarios across particle physics, materials science, and biochemistry, evaluating 3 GDL backbones and 11 algorithms under covariate, concept, and conditional shifts, with varied OOD access :contentReference[oaicite:5]{index=5}.
**licensing:** TODO
**task_types:** - Classification - Regression
**ai_capability_measured:** - OOD performance in scientific settings
**metrics:** - Accuracy - RMSE - OOD robustness delta
**models:** - GCN - EGNN - DimeNet++
**ml_motif:** - Geometric DL
**type:** Benchmark
**ml_task:** - Classification, Regression
**solutions:** TODO
**notes:** Includes no-OOD, unlabeled-OOD, and few-label scenarios :contentReference[oaicite:6]{index=6}.
**contact.name:** Deyu Zou
**contact.email:** unknown
**results.links.name:** ChatGPT LLM
**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]

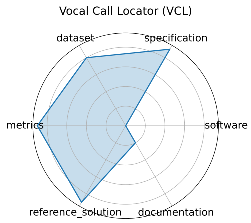
**Ratings:**


GeSS
dataset
specification
metrics
software
reference_solution
documentation

# 44 Vocal Call Locator (VCL)

**date:** 2024-12-13
**version:** TODO
**last_updated:** 2024-12
**expired:** unknown
**valid:** yes
**valid_date:** TODO
**url:** https://neurips.cc/virtual/2024/poster/97470
**doi:** TODO
**domain:** Neuroscience; Bioacoustics
**focus:** Benchmarking sound-source localization of rodent vocalizations from multi-channel audio
**keywords:** - source localization - bioacoustics - time-series - SSL
**summary:** The first large-scale benchmark (767K sounds across 9 conditions) for localizing rodent vocal calls using synchronized audio and video in standard lab environments, enabling systematic evaluation of sound-source localization algorithms in bioacoustics :contentReference[oaicite:1]{index=1}.
**licensing:** TODO
**task_types:** - Sound source localization
**ai_capability_measured:** - Source localization accuracy in bioacoustic settings
**metrics:** - Localization error (cm) - Recall/Precision
**models:** - CNN-based SSL models
**ml_motif:** - Real-time
**type:** Dataset
**ml_task:** - Anomaly detection / localization
**solutions:** TODO
**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.links.name:** ChatGPT LLM
**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 x 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:**

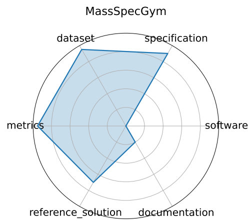


Vocal Call Locator (VCL)
dataset
specification
metrics
software
reference_solution
documentation

# 45 MassSpecGym

**date:** 2024-12-13
**version:** TODO
**last_updated:** 2024-12
**expired:** unknown
**valid:** yes
**valid_date:** TODO
**url:** https://neurips.cc/virtual/2024/poster/97823
**doi:** TODO
**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
**summary:** MassSpecGym curates the largest public MS/MS dataset with three standardized tasks-de novo structure generation, molecule retrieval, and spectrum simulation-using challenging generalization splits to propel ML-driven molecule discovery :contentReference[oaicite:3]{index=3}.
**licensing:** TODO
**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
**solutions:** TODO
**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.links.name:** ChatGPT LLM
**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]

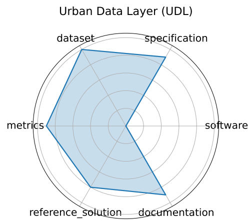
**Ratings:**


MassSpecGym
dataset
specification
metrics
software
reference_solution
documentation

# 46 Urban Data Layer (UDL)

**date:** 2024-12-13
**version:** TODO
**last_updated:** 2024-12
**expired:** unknown
**valid:** yes
**valid_date:** TODO
**url:** https://neurips.cc/virtual/2024/poster/97837
**doi:** TODO
**domain:** Urban Computing; Data Engineering
**focus:** Unified data pipeline for multi-modal urban science research
**keywords:** - data pipeline - urban science - multi-modal - benchmark
**summary:** UrbanDataLayer standardizes heterogeneous urban data formats and provides pipelines for tasks like air quality prediction and land-use classification, enabling the rapid creation of multi-modal urban benchmarks :contentReference[oaicite:5]{index=5}.
**licensing:** TODO
**task_types:** - Prediction - Classification
**ai_capability_measured:** - Multi-modal urban inference - standardization
**metrics:** - Task-specific accuracy or RMSE
**models:** - Baseline regression/classification pipelines
**ml_motif:** - Data engineering
**type:** Framework
**ml_task:** - Prediction, classification
**solutions:** TODO
**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:** unknown
**results.links.name:** ChatGPT LLM
**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]

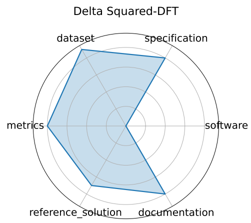
**Ratings:**


Urban Data Layer (UDL)
dataset
specification
metrics
software
reference_solution
documentation

## 47 Delta Squared-DFT

**date:** 2024-12-13
**version:** TODO
**last_updated:** 2024-12
**expired:** unknown
**valid:** yes
**valid_date:** TODO
**url:** https://neurips.cc/virtual/2024/poster/97788
**doi:** TODO
**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
**summary:** Introduces the Delta Squared-ML paradigm-using ML corrections to DFT to predict reaction energies with accuracy comparable to CCSD(T), while training on small CC datasets. Evaluated across 10 reaction datasets covering organic and organometallic transformations.
**licensing:** TODO
**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
**solutions:** TODO
**notes:** Demonstrates CC-level accuracy with ~1% of high-level data. Benchmarks publicly included for reproducibility.
**contact.name:** Wei Liu
**contact.email:** unknown
**results.links.name:** ChatGPT LLM
**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_squared-dft
**Citations:** [41]

**Ratings:**

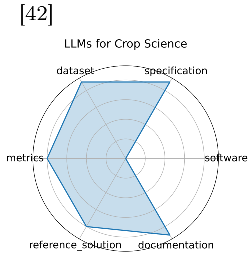


Delta Squared-DFT
dataset
specification
metrics
software
reference_solution
documentation

# 48 LLMs for Crop Science

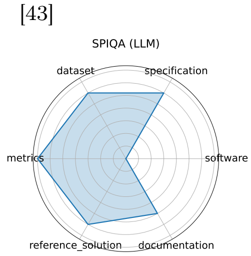
**date:** 2024-12-13
**version:** TODO
**last_updated:** 2024-12
**expired:** unknown
**valid:** yes
**valid_date:** TODO
**url:** https://neurips.cc/virtual/2024/poster/97570
**doi:** TODO
**domain:** Agricultural Science; NLP
**focus:** Evaluating LLMs on crop trait QA and textual inference tasks with domain-specific prompts
**keywords:** - crop science - prompt engineering - domain adaptation - question answering
**summary:** Establishes a benchmark of 3,500 expert-annotated prompts and QA pairs covering crop traits, growth stages, and environmental interactions. Tests GPT-style LLMs on accuracy and domain reasoning using in-context, chain-of-thought, and retrieval-augmented prompts.
**licensing:** TODO
**task_types:** - Question Answering - Inference
**ai_capability_measured:** - Scientific knowledge - crop reasoning
**metrics:** - Accuracy - F1 score
**models:** - GPT-4 - LLaMA-2-13B - T5-XXL
**ml_motif:** - NLP
**type:** Dataset
**ml_task:** - QA, inference
**solutions:** TODO
**notes:** Includes examples with retrieval-augmented and chain-of-thought prompt templates; supports few-shot adaptation.
**contact.name:** Deepak Patel
**contact.email:** unknown
**results.links.name:** ChatGPT LLM
**fair.reproducible:** Yes
**fair.benchmark_ready:** Yes
**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
**version:** TODO
**last_updated:** 2024-12
**expired:** unknown
**valid:** yes
**valid_date:** TODO
**url:** https://neurips.cc/virtual/2024/poster/97575
**doi:** TODO
**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
**summary:** A workshop version of SPIQA comparing 10 LLM adapter methods on the SPIQA benchmark with scientific diagram/questions. Highlights performance differences between chain-of-thought and end-to-end adapter models.
**licensing:** TODO
**task_types:** - Multimodal QA
**ai_capability_measured:** - Visual reasoning - scientific figure understanding
**metrics:** - Accuracy - F1 score
**models:** - LLaVA - MiniGPT-4 - Owl-LLM adapter variants
**ml_motif:** - Multimodal QA
**type:** Benchmark
**ml_task:** - Multimodal QA
**solutions:** TODO
**notes:** Companion to SPIQA main benchmark; compares adapter strategies using same images and QA pairs.
**contact.name:** Xiaoyan Zhong
**contact.email:** unknown
**results.links.name:** ChatGPT LLM
**fair.reproducible:** Yes
**fair.benchmark_ready:** Yes
**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:** spiqa_llm
**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.*, *Braggnn: 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=7yt3N0oOW9`.

[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 mlcommons science working group," in *High Performance Computing. ISC High Performance 2022 International Workshops*, H. Anzt, A. Bienz, P. Luszczek, and M. Baboulin, Eds., Cham: Springer International Publishing, 2022, pp. 47–64, ISBN: 978-3-031-23220-6.

[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. 100 543, 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. 44 989–45 037. [Online]. Available: `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.

[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.

[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.

[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.

[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.

[41] K. Khrabrov, A. Ber, A. Tsypin, *et al.*, *$\nabla^2$Dft: A universal quantum chemistry dataset of drug-like molecules and a benchmark for neural network potentials*, 2024. arXiv: 2406.14347 [physics.chem-ph]. [Online]. Available: https://arxiv.org/abs/2406.14347.

[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, *Spiqa: A dataset for multimodal question answering on scientific papers*, 2025. arXiv: 2407.09413 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2407.09413.