Date	**Name**	**Domain**	**Focus**	**Task Types**	**Metrics**	**Models**	**Citation**
2020-09-07	MMLU (Massive Multitask Language Understanding)	Multidomain	Academic knowledge and reasoning across 57 subjects	Multiple choice	Accuracy	GPT-4o, Gemini 1.5 Pro, o1, DeepSeek-R1	[1] ⇒
2023-11-20	GPQA Diamond	Science	Graduate-level scientific reasoning	Multiple choice, Multi-step QA	Accuracy	o1, DeepSeek-R1	[2] ⇒
2018-03-14	ARC-Challenge (Advanced Reasoning Challenge)	Science	Grade-school science with reasoning emphasis	Multiple choice	Accuracy	GPT-4, Claude	[3] ⇒
2025-01-24	Humanity's Last Exam	Multidomain	Broad cross-domain academic reasoning	Multiple choice	Accuracy		[4] ⇒
2024-11-07	FrontierMath	Mathematics	Challenging advanced mathematical reasoning	Problem solving	Accuracy		[5] ⇒
2024-07-18	SciCode	Scientific Programming	Scientific code generation and problem solving	Coding	Solve rate (percent)	Claude3.5-Sonnet	[6] ⇒
2025-03-13	AIME (American Invitational Mathematics Examination)	Mathematics	Pre-college advanced problem solving	Problem solving	Accuracy		[7] ⇒
2025-02-15	MATH-500	Mathematics	Math reasoning generalization	Problem solving	Accuracy		[8] ⇒
2024-04-02	CURIE (Scientific Long-Context Understanding, Reasoning and Information Extraction)	Multidomain Science	Long-context scientific reasoning	Information extraction, Reasoning, Concept tracking, Aggregation, Algebraic manipulation, Multimodal comprehension	Accuracy		[9] ⇒
2023-01-26	FEABench (Finite Element Analysis Benchmark)	Computational Engineering	FEA simulation accuracy and performance	Simulation, Performance evaluation	Solve time, Error norm	FEniCS, deal.II	[10] ⇒
2024-07-12	SPIQA (Scientific Paper Image Question Answering)	Computer Science	Multimodal QA on scientific figures	Question answering, Multimodal QA, Chain-of-Thought evaluation	Accuracy, F1 score	Chain-of-Thought models, Multimodal QA systems	[11] ⇒

Continued on next page

Date	Name	Domain	Focus	Task Types	Metrics	Models	Citation
2020-09-28	MedQA	Medical Question Answering	Medical board exam QA	Multiple choice	Accuracy	Neural reader, Retrieval-based QA systems	[12] ⇒
2025-05-13	BaisBench (Biological AI Scientist Benchmark)	Computational Biology	Omics-driven AI research tasks	Cell type annotation, Multiple choice	Annotation accuracy, QA accuracy	LLM-based AI scientist agents	[13] ⇒
2023-01-26	MOLGEN	Computational Chemistry	Molecular generation and optimization	Distribution learning, Goal-oriented generation	Validity percent, Novelty percent, QED, Docking score	MolGen	[14] ⇒
2020-05-02	Open Graph Benchmark (OGB) - Biology	Graph ML	Biological graph property prediction	Node property prediction, Link property prediction, Graph property prediction	Accuracy, ROC-AUC	GCN, GraphSAGE, GAT	[15] ⇒
2011-10-01	Materials Project	Materials Science	DFT-based property prediction	Property prediction	MAE, R^2	Automatminer, Crystal Graph Neural Networks	[16] ⇒
2020-10-20	OCP (Open Catalyst Project)	Chemistry; Materials Science	Catalyst adsorption energy prediction	Energy prediction, Force prediction	MAE (energy), MAE (force)	CGCNN, SchNet, DimeNet++, GemNet-OC	[17]–[20] ⇒
2023-06-20	JARVIS-Leaderboard	Materials Science; Benchmarking	Comparative evaluation of materials design methods	Method benchmarking, Leaderboard ranking	MAE, RMSE, Accuracy		[21] ⇒
2022-02-22	Quantum Computing Benchmarks (QML)	Quantum Computing	Quantum algorithm performance evaluation	Circuit benchmarking, State classification	Fidelity, Success probability	IBM Q, IonQ, AQT@LBNL	[22] ⇒
2024-10-01	CFDBench (Fluid Dynamics)	Fluid Dynamics; Scientific ML	Neural operator surrogate modeling	Surrogate modeling	L2 error, MAE	FNO, DeepONet, U-Net	[23] ⇒
None	SatImgNet	Remote Sensing	Satellite imagery classification	Image classification	Accuracy		[24] ⇒
2023-07-19	ClimateLearn	Climate Science; Forecasting	ML for weather and climate modeling	Forecasting	RMSE, Anomaly correlation	CNN baselines, ResNet variants	[25] ⇒
2022-06-09	BIG-Bench (Beyond the Imitation Game Benchmark)	NLP; AI Evaluation	Diverse reasoning and generalization tasks	Few-shot evaluation, Multi-task evaluation	Accuracy, Task-specific metrics	GPT-3, Dense Transformers, Sparse Transformers	[26] ⇒
2019-11-20	CommonSenseQA	NLP; Commonsense	Commonsense question answering	Multiple choice	Accuracy	BERT-large, RoBERTa, GPT-3	[27] ⇒

Continued on next page

Date	Name	Domain	Focus	Task Types	Metrics	Models	Citation
2019-07-24	Winogrande	NLP; Commonsense	Winograd Schema-style pronoun resolution	Pronoun resolution	Accuracy, AUC	RoBERTa, BERT, GPT-2	[28] ⇒
2024-05-01	Jet Classification	Particle Physics	Real-time classification of particle jets using HL-LHC simulation features	Classification	Accuracy, AUC	Keras DNN, QKeras quantized DNN	[29] ⇒
2024-05-01	Irregular Sensor Data Compression	Particle Physics	Real-time compression of sparse sensor data with autoencoders	Compression	MSE, Compression ratio	Autoencoder, Quantized autoencoder	[30] ⇒
2024-05-01	Beam Control	Accelerators and Magnets	Reinforcement learning control of accelerator beam position	Control	Stability, Control loss	DDPG, PPO (planned)	[31], [32] ⇒
2024-07-08	Ultrafast jet classification at the HL-LHC	Particle Physics	FPGA-optimized real-time jet origin classification at the HL-LHC	Classification	Accuracy, Latency, Resource utilization	MLP, Deep Sets, Interaction Network	⇒
2024-10-15	Quench detection	Accelerators and Magnets	Real-time detection of superconducting magnet quenches using ML	Anomaly detection, Quench localization	ROC-AUC, Detection latency	Autoencoder, RL agents (in development)	⇒
2024-10-15	DUNE	Particle Physics	Real-time ML for DUNE DAQ time-series data	Trigger selection, Time-series anomaly detection	Detection efficiency, Latency	CNN, LSTM (planned)	⇒
2025-01-08	Intelligent experiments through real-time AI	Instrumentation & Detectors; Nuclear Physics; Particle Physics	Real-time FPGA-based triggering and detector control for sPHENIX and future EIC	Trigger classification, Detector control, Real-time inference	Accuracy (charm & beauty detection), Latency (µs), Resource utilization (LUT/FF/BRAM/DSP)	Bipartite Graph Network with Set Transformers (BGN-ST), GarNet (edge-classifier)	[33] ⇒

Continued on next page

Date	Name	Domain	Focus	Task Types	Metrics	Models	Citation
2025-01-09	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	Classification, Peak finding	Accuracy, Latency, Resource utilization	NAC-based BraggNN, NAC-optimized Deep Sets (jet)	[34] ⇒
2024-06-24	Smart Pixels for LHC	Particle Physics; Instrumentation & Detectors	On-sensor, in-pixel ML filtering for high-rate LHC pixel detectors	Image Classification, Data filtering	Data rejection rate, Power per pixel	2-layer pixel NN	[35] ⇒
2023-10-03	HEDM (BraggNN)	Material Science	Fast Bragg peak analysis using deep learning in diffraction microscopy	Peak detection	Localization accuracy, Inference time	BraggNN	[36] ⇒
2023-12-03	4D-STEM	Material Science	Real-time ML for scanning transmission electron microscopy	Image Classification, Streamed data inference	Classification accuracy, Throughput	CNN models (prototype)	[37] ⇒
2023-12-05	In-Situ High-Speed Computer Vision	Fusion/Plasma	Real-time image classification for in-situ plasma diagnostics	Image Classification	Accuracy, FPS	CNN	[38] ⇒
2020-01-01	BenchCouncil AIBench	General	End-to-end AI benchmarking across micro, component, and application levels	Training, Inference, End-to-end AI workloads	Throughput, Latency, Accuracy	ResNet, BERT, GANs, Recommendation systems	[39] ⇒
2020-01-01	BenchCouncil BigDataBench	General	Big data and AI benchmarking across structured, semi-structured, and unstructured data workloads	Data preprocessing, Inference, End-to-end data pipelines	Data throughput, Latency, Accuracy	CNN, LSTM, SVM, XGBoost	[40] ⇒
2021-10-20	MLPerf HPC	Cosmology, Climate, Protein Structure, Catalysis	Scientific ML training and inference on HPC systems	Training, Inference	Training time, Accuracy, GPU utilization	CosmoFlow, DeepCAM, OpenCatalyst	[41] ⇒

Continued on next page

Date	Name	Domain	Focus	Task Types	Metrics	Models	Citation
2023-06-01	MLCommons Science	Earthquake, Satellite Image, Drug Discovery, Electron Microscope, CFD	AI benchmarks for scientific applications including time-series, imaging, and simulation	Time-series analysis, Image classification, Simulation surrogate modeling	MAE, Accuracy, Speedup vs simulation	CNN, GNN, Transformer	[42] ⇒
2021-07-05	LHC New Physics Dataset	Particle Physics; Real-time Triggering	Real-time LHC event filtering for anomaly detection using proton collision data	Anomaly detection, Event classification	ROC-AUC, Detection efficiency	Autoencoder, Variational autoencoder, Isolation forest	[43] ⇒
2023-07-17	MLCommons Medical AI	Healthcare; Medical AI	Federated benchmarking and evaluation of medical AI models across diverse real-world clinical data	Federated evaluation, Model validation	ROC AUC, Accuracy, Fairness metrics	MedPerf-validated CNNs, GaNDLF workflows	[44] ⇒
2024-10-28	CaloChallenge 2022	LHC Calorimeter; Particle Physics	Fast generative-model-based calorimeter shower simulation evaluation	Surrogate modeling	Histogram similarity, Classifier AUC, Generation latency	VAE variants, GAN variants, Normalizing flows, Diffusion models	[45] ⇒
ongoing	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	Multiple (Classification, Detection, NLP, etc.)	Task-specific (Accuracy, F1, BLEU, etc.)	All published models with code	[46] ⇒
2022-01-01	Codabench	General ML; Multiple	Open-source platform for organizing reproducible AI benchmarks and competitions	Multiple	Submission count, Leaderboard ranking, Task-specific metrics	Arbitrary code submissions	[47]

Continued on next page

Date	Name	Domain	Focus	Task Types	Metrics	Models	Citation
2021-09-27	Sabath (SBI-FAIR)	Systems; Metadata	FAIR metadata framework for ML-driven surrogate workflows in HPC systems	Systems benchmarking	Metadata completeness, FAIR compliance	N/A	[48]
2022-10-13	PDEBench	CFD; Weather Modeling	Benchmark suite for ML-based surrogates solving time-dependent PDEs	Supervised Learning	RMSE, boundary RMSE, Fourier RMSE	FNO, U-Net, PINN, Gradient-Based inverse methods	[49] ⇒
2024-12-03	The Well	biological systems, fluid dynamics, acoustic scattering, astrophysical MHD	Foundation model + surrogate dataset spanning 16 physical simulation domains	Supervised Learning	Dataset size, Domain breadth	FNO baselines, U-Net baselines	[50]
2024-10-31	LLM-Inference-Bench	LLM; HPC/inference	Hardware performance benchmarking of LLMs on AI accelerators	Inference Benchmarking	Token throughput (tok/s), Latency, Framework-hardware mix performance	LLaMA-2-7B, LLaMA-2-70B, Mistral-7B, Qwen-7B	[51]
2023-12-12	SGLang Framework	LLM Vision	Fast serving framework for LLMs and vision-language models	Model serving framework	Tokens/sec, Time-to-first-token, Throughput gain vs baseline	LLaVA, DeepSeek, Llama	[52] ⇒
2023-09-12	vLLM Inference & Serving Engine	LLM; HPC/inference	High-throughput, memory-efficient inference and serving engine for LLMs	Inference Benchmarking	Tokens/sec, Time to First Token (TTFT), Memory footprint	LLaMA, Mixtral, FlashAttention-based models	[53]
2022-06-22	vLLM Performance Dashboard	LLM; HPC/inference	Interactive dashboard showing inference performance of vLLM	Performance visualization	Tokens/sec, TTFT, Memory usage	LLaMA-2, Mistral, Qwen	[54] ⇒

Continued on next page

Date	Name	Domain	Focus	Task Types	Metrics	Models	Citation
2022-04-01	Nixtla NeuralForecast	Time-series forecasting; General ML	High-performance neural forecasting library with ¿30 models	Time-series forecasting	RMSE, MAPE, CRPS	NBEATS, NHITS, TFT, DeepAR	[55] ⇒
2023-06-01	Nixtla Neural Forecast NHITS	Time-series; General ML	Official NHITS implementation for long-horizon time series forecasting	Time-series forecasting	RMSE, MAPE	NHITS	[56]
2023-10-03	Nixtla Neural Forecast TimeLLM	Time-series; General ML	Reprogramming LLMs for time series forecasting	Time-series forecasting	RMSE, MAPE	Time-LLM	[57]
2023-10-05	Nixtla Neural Forecast TimeGPT	Time-series; General ML	Time-series foundation model ”TimeGPT” for forecasting and anomaly detection	Time-series forecasting, Anomaly detection	RMSE, Anomaly detection metrics	TimeGPT	[58] ⇒
2025-03-03	HDR ML Anomaly Challenge (Gravitational Waves)	Astrophysics; Time-series	Detecting anomalous gravitational-wave signals from LIGO/Virgo datasets	Anomaly detection	ROC-AUC, Precision/Recall	Deep latent CNNs, Autoencoders	[59] ⇒
2025-03-03	HDR ML Anomaly Challenge (Butterfly)	Genomics; Image/CV	Detecting hybrid butterflies via image anomaly detection in genomic-informed dataset	Anomaly detection	Classification accuracy, F1 score	CNN-based detectors	[60] ⇒
2025-03-03	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	ROC-AUC, Precision/Recall	CNNs, RNNs, Transformers	[61] ⇒

Continued on next page

Date	Name	Domain	Focus	Task Types	Metrics	Models	Citation
2025-01-24	Single Qubit Readout on QICK System	Quantum Computing	Real-time single-qubit state classification using FPGA firmware	Classification	Accuracy, Latency	hls4ml quantized NN	[62] ⇒
2023-11-20	GPQA: A Graduate-Level Google-Proof Q&A Benchmark	Science (Biology, Physics, Chemistry)	Graduate-level, expert-validated multiple-choice questions hard even with web access	Multiple choice	Accuracy	GPT-4 baseline	[63] ⇒
2024-12-13	SeafloorAI	Marine Science; Vision-Language	Large-scale vision-language dataset for seafloor mapping and geological classification	Image segmentation, Vision-language QA	Segmentation pixel accuracy, QA accuracy	SegFormer, ViLT-style multimodal models	[64] ⇒
2024-12-13	SuperCon3D	Materials Science; Superconductivity	Dataset and models for predicting and generating high-Tc superconductors using 3D crystal structures	Regression (Tc prediction), Generative modeling	MAE (Tc), Validity of generated structures	SODNet, DiffCSP-SC	[65]
2024-12-13	GeSS	Scientific ML; Geometric Deep Learning	Benchmark suite evaluating geometric deep learning models under real-world distribution shifts	Classification, Regression	Accuracy, RMSE, OOD robustness delta	GCN, EGNN, DimeNet++	[66]
2024-12-13	Vocal Call Locator (VCL)	Neuroscience; Bioacoustics	Benchmarking sound-source localization of rodent vocalizations from multi-channel audio	Sound source localization	Localization error (cm), Recall/Precision	CNN-based SSL models	[67] ⇒

Continued on next page

Date	Name	Domain	Focus	Task Types	Metrics	Models	Citation
2024-12-13	MassSpecGym	Cheminformatics; Molecular Discovery	Benchmark suite for discovery and identification of molecules via MS/MS	De novo generation, Retrieval, Simulation	Structure accuracy, Retrieval precision, Simulation MSE	Graph-based generative models, Retrieval baselines	[68] ⇒
2024-12-13	Urban Data Layer (UDL)	Urban Computing; Data Engineering	Unified data pipeline for multi-modal urban science research	Prediction, Classification	Task-specific accuracy or RMSE	Baseline regression/classification pipelines	[69] ⇒
2024-12-13	Delta Squared-DFT	Computational Chemistry; Materials Science	Benchmarking machine-learning corrections to DFT using Delta Squared-trained models for reaction energies	Regression	Mean Absolute Error (eV), Energy ranking accuracy	Delta Squared-ML correction networks, Kernel ridge regression	[70] ⇒
2024-12-13	LLMs for Crop Science	Agricultural Science; NLP	Evaluating LLMs on crop trait QA and textual inference tasks with domain-specific prompts	Question Answering, Inference	Accuracy, F1 score	GPT-4, LLaMA-2-13B, T5-XXL	[71] ⇒
2024-12-13	SPIQA (LLM)	Multimodal Scientific QA; Computer Vision	Evaluating LLMs on image-based scientific paper figure QA tasks (LLM Adapter performance)	Multimodal QA	Accuracy, F1 score	LLaVA, MiniGPT-4, Owl-LLM adapter variants	[72] ⇒

References

[1] D. Hendrycks, C. Burns, S. Kadavath, *et al.*, "Measuring massive multitask language understanding," *arXiv preprint arXiv:2009.03300*, 2021. [Online]. Available: `https://arxiv.org/abs/2009.03300`.

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

[3] P. Clark, I. Cowhey, O. Etzioni, *et al.*, "Think you have solved question answering? try arc, the ai2 reasoning challenge," in *EMNLP 2018*, 2018, pp. 237–248. [Online]. Available: `https://allenai.org/data/arc`.

[4] L. Phan, A. Gatti, Z. Han, *et al.*, *Humanity's last exam*, 2025. [Online]. Available: `https://arxiv.org/abs/2501.14249`.

[5] E. Glazer, E. Erdil, T. Besiroglu, *et al.*, *Frontiermath: A benchmark for evaluating advanced mathematical reasoning in ai*, 2024. [Online]. Available: `https://arxiv.org/abs/2411.04872`.

[6] M. Tian, L. Gao, S. Zhang, *et al.*, *Scicode: A research coding benchmark curated by scientists*, 2024. [Online]. Available: `https://arxiv.org/abs/2407.13168`.

[7] TBD, *Aime*, [Online accessed 2025-06-24], Mar. 2025. [Online]. Available: `https://www.vals.ai/benchmarks/aime-2025-03-13`.

[8] HuggingFaceH4, *Math-500*, 2025. [Online]. Available: `https://huggingface.co/datasets/HuggingFaceH4/MATH-500`.

[9] T. A. authors, *Scientific reasoning benchmarks from the curie dataset*, 2024. [Online]. Available: `https://arxiv.org/abs/2404.02029`.

[10] A. Institute, *Feabench: A finite element analysis benchmark*, 2023. [Online]. Available: `https://github.com/alleninstitute/feabench`.

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

[12] D. Jin, Y. Li, Y. Zhang, *et al.*, "What disease does this patient have? a large-scale open-domain question answering dataset from medical exams," 2020. [Online]. Available: `https://arxiv.org/abs/2009.13081`.

[13] E. Luo, J. Jia, Y. Xiong, *et al.*, *Benchmarking ai scientists in omics data-driven biological research*, 2025. [Online]. Available: `https://arxiv.org/abs/2505.08341`.

[14] Y. Fang, N. Zhang, Z. Chen, *et al.*, "Domain-agnostic molecular generation with chemical feedback," 2023. [Online]. Available: `https://arxiv.org/abs/2301.11259`.

[15] W. Hu, M. Fey, M. Zitnik, *et al.*, *Open graph benchmark: Datasets for machine learning on graphs*, 2020. [Online]. Available: `https://arxiv.org/abs/2005.00687`.

[16] A. Jain, S. P. Ong, G. Hautier, *et al.*, "The materials project: A materials genome approach," *APL Materials*, vol. 1, no. 1, 2013. DOI: `10.1063/1.4812323`. [Online]. Available: `https://materialsproject.org/`.

[17] L. Chanussot, A. Das, S. Goyal, *et al.*, "The open catalyst 2020 (oc20) dataset and community challenges," *ACS Catalysis*, vol. 11, no. 10, pp. 6059–6072, 2021. DOI: `10.1021/acscatal.0c04525`. [Online]. Available: `https://pubs.acs.org/doi/10.1021/acscatal.0c04525`.

[18] R. Tran, J. Lan, M. Shuaibi, *et al.*, "The open catalyst 2022 (oc22) dataset and challenges for oxide electrocatalysts," *ACS Catalysis*, vol. 13, no. 5, pp. 3066–3084, 2023. DOI: `10.1021/acscatal.2c05426`. [Online]. Available: `https://pubs.acs.org/doi/10.1021/acscatal.2c05426`.

[19] L. Chanussot, A. Das, S. Goyal, *et al.*, "Open catalyst 2020 (oc20) dataset and community challenges," *ACS Catalysis*, vol. 11, no. 10, pp. 6059–6072, 2021. DOI: `10.1021/acscatal.0c04525`. eprint: `https://doi.org/10.1021/acscatal.0c04525`. [Online]. Available: `https://doi.org/10.1021/acscatal.0c04525`.

[20] R. Tran, J. Lan, M. Shuaibi, *et al.*, "The open catalyst 2022 (oc22) dataset and challenges for oxide electrocatalysts," *ACS Catalysis*, vol. 13, no. 5, pp. 3066–3084, Feb. 2023, ISSN: 2155-5435. DOI: `10.1021/acscatal.2c05426`. [Online]. Available: `http://dx.doi.org/10.1021/acscatal.2c05426`.

[21] K. Choudhary, D. Wines, K. Li, *et al.*, "JARVIS-Leaderboard: A large scale benchmark of materials design methods," *npj Computational Materials*, vol. 10, no. 1, p. 93, 2024. DOI: `10.1038/s41524-024-01259-w`. [Online]. Available: `https://doi.org/10.1038/s41524-024-01259-w`.

[22] F. J. Kiwit, M. Marso, P. Ross, C. A. Riofrío, J. Klepsch, and A. Luckow, "Application-oriented benchmarking of quantum generative learning using quark," in *2023 IEEE International Conference on Quantum Computing and Engineering (QCE)*, IEEE, Sep. 2023, pp. 475–484. DOI: `10.1109/qce57702.2023.00061`. [Online]. Available: `http://dx.doi.org/10.1109/QCE57702.2023.00061`.

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

[24] J. Roberts, K. Han, and S. Albanie, *Satin: A multi-task metadataset for classifying satellite imagery using vision-language models*, 2023. arXiv: `2304.11619 [cs.CV]`. [Online]. Available: `https://arxiv.org/abs/2304.11619`.

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

[26] A. Srivastava, A. Rastogi, A. Rao, *et al.*, *Beyond the imitation game: Quantifying and extrapolating the capabilities of language models*, 2023. arXiv: `2206.04615 [cs.CL]`. [Online]. Available: `https://arxiv.org/abs/2206.04615`.

[27] A. Talmor, J. Herzig, N. Lourie, and J. Berant, *Commonsenseqa: A question answering challenge targeting commonsense knowledge*, 2019. arXiv: `1811.00937 [cs.CL]`. [Online]. Available: `https://arxiv.org/abs/1811.00937`.

[28] K. Sakaguchi, R. L. Bras, C. Bhagavatula, and Y. Choi, *Winogrande: An adversarial winograd schema challenge at scale*, 2019. arXiv: `1907.10641 [cs.CL]`. [Online]. Available: `https://arxiv.org/abs/1907.10641`.

[29] B. Hawks, N. Tran, *et al.*, "Fast machine learning for science: Benchmarks and dataset," 2022. [Online]. Available: `https://arxiv.org/abs/2207.07958`.

[30] B. Hawks, N. Tran, *et al.*, "Fast machine learning for science: Benchmarks and dataset," 2022. [Online]. Available: `https://arxiv.org/abs/2207.07958`.

[31] B. Hawks, N. Tran, *et al.*, "Fast machine learning for science: Benchmarks and dataset," 2022. [Online]. Available: `https://arxiv.org/abs/2207.07958`.

[32] Q. Wang *et al.*, "Boostr: A dataset for accelerator control systems," 2021. [Online]. Available: `https://arxiv.org/abs/2101.08359`.

[33] J. Kvapil, G. Borca-Tasciuc, N. ... Tran, *et al.*, "Intelligent experiments through real-time ai: Fast data processing and autonomous detector control for sphenix and future eic detectors," 2025. [Online]. Available: `https://arxiv.org/abs/2501.04845`.

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

[35] B. Parpillon, ..., and N. Tran, "Smart pixels: In-pixel ai for on-sensor data filtering," 2024. [Online]. Available: `https://arxiv.org/abs/2406.14860`.

[36] Y. Xiao and ..., "Braggnn: Fast x-ray bragg peak analysis using deep learning," 2020. [Online]. Available: `https://arxiv.org/abs/2008.08198`.

[37] Anonymous, "4d-stem: Real-time ml for electron microscopy," 2023. [Online]. Available: `https://openreview.net/pdf?id=7yt3N0o0W9`.

[38] J. Smith and J. Doe, "In-situ high-speed computer vision for plasma diagnostics," 2023. [Online]. Available: `https://arxiv.org/abs/2312.00128`.

[39] W. Gao, J. Zhan, *et al.*, "Aibench: An industry standard internet service ai benchmark suite," 2020. [Online]. Available: `https://arxiv.org/abs/1908.08998`.

[40] W. Gao, J. Zhan, *et al.*, "Bigdatabench: A scalable and unified big data and ai benchmark suite," 2018. [Online]. Available: `https://arxiv.org/abs/1802.08254`.

[41] S. Farrell, M. Emani, *et al.*, "Mlperf hpc: A holistic benchmark suite for scientific machine learning on hpc systems," 2021. [Online]. Available: `https://arxiv.org/abs/2110.11466`.

[42] M. S. W. Group, *Mlcommons science working group benchmarks*, 2023. [Online]. Available: `https://github.com/mlcommons/science`.

[43] E. Govorkova, E. Puljak, M. Pierini, *et al.*, "Lhc physics dataset for unsupervised new physics detection at 40 mhz," *Scientific Data*, 2022. DOI: `10.6084/m9.figshare.5046389`. [Online]. Available: `https://doi.org/10.5281/zenodo.5046389`.

[44] M. J. Karargyris Alex and Sheller *et al.*, "Federated benchmarking of medical artificial intelligence with medperf," *Nature Machine Intelligence*, 2023. [Online]. Available: `https://www.nature.com/articles/s42256-023-00652-2`.

[45] C. Krause, B. Nachman, *et al.*, "Calochallenge 2022: A community challenge for fast calorimeter simulation," 2024. [Online]. Available: `https://arxiv.org/abs/2410.21611`.

[46] P. W. Code, *Papers with code: Open machine learning benchmarks and leaderboards*, 2025. [Online]. Available: `https://paperswithcode.com`.

[47] Z. Xu, S. Escalera, *et al.*, "Codabench: Flexible, easy-to-use, and reproducible meta-benchmark platform," *Patterns*, vol. 3, no. 7, p. 100 543, 2022. DOI: `10.1016/j.patter.2022.100543`.

[48] P. Luszczek *et al.*, "Sabath: Fair metadata technology for surrogate benchmarks," University of Tennessee, Tech. Rep., 2021.

[49] M. Takamoto, T. Praditia, *et al.*, "Pdebench: An extensive benchmark for scientific machine learning," in *NeurIPS Datasets Benchmarks Track*, 2022. [Online]. Available: `https://arxiv.org/abs/2210.07182`.

[50] R. Ohana, M. McCabe, L. Meyer, *et al.*, "The well: A large-scale collection of diverse physics simulations for machine learning," *NeurIPS*, vol. 37, pp. 44 989–45 037, 2024.

[51] K. T. Chitty-Venkata, S. Raskar, *et al.*, "Llm-inference-bench: Inference benchmarking of large language models on ai accelerators," *arXiv preprint arXiv:2411.00136*, 2024.

[52] L. Zheng, L. Yin, *et al.*, "Sglang: Efficient execution of structured language model programs," 2023. [Online]. Available: `https://arxiv.org/abs/2312.07104`.

[53] W. Kwon *et al.*, "Efficient memory management for large language model serving with pagedattention," in *SOSP 2023*, 2023.

[54] S. Mo, *Vllm performance dashboard*, 2024. [Online]. Available: `https://simon-mo-workspace.observablehq.cloud/vllm-dashboard-v0/`.

[55] K. G. Olivares, C. Challú, *et al.*, *Neuralforecast: User friendly state-of-the-art neural forecasting models*, PyCon US, 2022. [Online]. Available: `https://github.com/Nixtla/neuralforecast`.

[56] C. Challu, K. G. Olivares, *et al.*, "Nhits: Neural hierarchical interpolation for time series forecasting," in *AAAI 2023*, 2023.

[57] M. Jin, S. Wang, *et al.*, "Time-llm: Time series forecasting by reprogramming large language models," *arXiv preprint arXiv:2310.01728*, 2023.

[58] A. Garza, C. Challu, *et al.*, "Timegpt-1: A foundation model for time series," 2023. [Online]. Available: `https://arxiv.org/abs/2310.03589`.

[59] E. G. Campolongo *et al.*, "Building machine learning challenges for anomaly detection in science," 2025. [Online]. Available: `https://arxiv.org/abs/2503.02112`.

[60] E. G. Campolongo *et al.*, "Building machine learning challenges for anomaly detection in science," 2025. [Online]. Available: `https://arxiv.org/abs/2503.02112`.

[61] E. G. Campolongo *et al.*, “Building machine learning challenges for anomaly detection in science,” 2025. [Online]. Available: `https://arxiv.org/abs/2503.02112`.

[62] G. Di Guglielmo, J. Campos, *et al.*, “End-to-end workflow for machine learning-based qubit readout with qick and hls4ml,” 2025. [Online]. Available: `https://arxiv.org/abs/2501.14663`.

[63] D. Rein, B. L. Hou, A. C. Stickland, *et al.*, “Gpqa: A graduate-level google-proof q&a benchmark,” 2023. [Online]. Available: `https://arxiv.org/abs/2311.12022`.

[64] K. X. Nguyen, F. Qiao, *et al.*, “Seafloorai: A large-scale vision-language dataset for seafloor geological survey,” 2024. [Online]. Available: `https://arxiv.org/abs/2411.00172`.

[65] Z. Zuo *et al.*, “Supercon3d: Learning superconductivity from ordered and disordered material structures,” 2024, NeurIPS Poster.

[66] D. Zou, S. Liu, *et al.*, “Gess: Benchmarking geometric deep learning under scientific applications with distribution shifts,” 2024, NeurIPS Poster.

[67] R. Peterson, A. Tanelus, *et al.*, “Vocal call locator benchmark for localizing rodent vocalizations,” 2024, NeurIPS Poster. [Online]. Available: `https://neurips.cc/virtual/2024/poster/97470`.

[68] R. Bushuiev, A. Bushuiev, *et al.*, “Massspecgym: A benchmark for the discovery and identification of molecules,” 2024, NeurIPS Spotlight Poster. [Online]. Available: `https://neurips.cc/virtual/2024/poster/97823`.

[69] Y. Wang, T. Wang, *et al.*, “Urbandatalayer: A unified data pipeline for urban science,” 2024, NeurIPS Poster. [Online]. Available: `https://neurips.cc/virtual/2024/poster/97837`.

[70] W. Liu, R. Chen, *et al.*, “Delta squared-dft: Machine-learning corrected density functional theory for reaction energetics,” 2024, NeurIPS Poster. [Online]. Available: `https://neurips.cc/virtual/2024/poster/97788`.

[71] D. Patel, L. Zhao, *et al.*, “Large language models for crop science: Benchmarking domain reasoning and qa,” 2024, NeurIPS Poster. [Online]. Available: `https://neurips.cc/virtual/2024/poster/97570`.

[72] X. Zhong, Y. Gao, *et al.*, “Spiqa-llm: Evaluating llm adapters on scientific figure qa,” 2024, NeurIPS Poster. [Online]. Available: `https://neurips.cc/virtual/2024/poster/97575`.