Date	Name	Domain	Focus	Keyword	Task Types	Metrics	Models	Citation
2020-09-07	MMLU (Massive Multitask Language Understanding)	Multidomain	Academic knowledge and reasoning across 57 subjects	[multitask, multiple-choice, zero-shot, few-shot, knowledge probing]	[Multiple choice]	[Accuracy]	[GPT-4o, Gemini 1.5 Pro, o1, DeepSeek-R1]	[1] ⇒
2023-11-20	GPQA Diamond	Science	Graduate-level scientific reasoning	[Google-proof, graduate-level, science QA, chemistry, physics]	[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	[grade-school, science QA, challenge set, reasoning]	[Multiple choice]	[Accuracy]	[GPT-4, Claude]	[3] ⇒
2025-01-24	Humanity's Last Exam	Multidomain	Broad cross-domain academic reasoning	[cross-domain, academic exam, multiple-choice, multidisciplinary]	[Multiple choice]	[Accuracy]	[]	[4] ⇒
2024-11-07	FrontierMath	Mathematics	Challenging advanced mathematical reasoning	[symbolic reasoning, number theory, algebraic geometry, category theory]	[Problem solving]	[Accuracy]	[]	[5] ⇒
2024-07-18	SciCode	Scientific Programming	Scientific code generation and problem solving	[code synthesis, scientific computing, programming benchmark]	[Coding]	[Solve rate (percent)]	[Claude3.5-Sonnet]	[6] ⇒
2025-03-13	AIME (American Invitational Mathematics Examination)	Mathematics	Pre-college advanced problem solving	[algebra, combinatorics, number theory, geometry]	[Problem solving]	[Accuracy]	[]	[7] ⇒
2025-02-15	MATH-500	Mathematics	Math reasoning generalization	[calculus, algebra, number theory, geometry]	[Problem solving]	[Accuracy]	[]	[8] ⇒
2024-04-02	CURIE (Scientific Long-Context Understanding, Reasoning and Information Extraction)	Multidomain Science	Long-context scientific reasoning	[long-context, information extraction, multimodal]	[Information extraction, Reasoning, Concept tracking, Aggregation, Algebraic manipulation, Multimodal comprehension]	[Accuracy]	[]	[9] ⇒
2023-01-26	FEABench (Finite Element Analysis Benchmark)	Computational Engineering	FEA simulation accuracy and performance	[finite element, simulation, PDE]	[Simulation, Performance evaluation]	[Solve time, Error norm]	[FEniCS, deal.II]	[10] ⇒

Date	Name	Domain	Focus	Keyword	Task Types	Metrics	Models	Citation
2024-07-12	SPIQA (Scientific Paper Image Question Answering)	Computer Science	Multimodal QA on scientific figures	[multimodal QA, figure understanding, table comprehension, chain-of-thought]	[Question answering, Multimodal QA, Chain-of-Thought evaluation]	[Accuracy, F1 score]	[Chain-of-Thought models, Multimodal QA systems]	[11] ⇒
2020-09-28	MedQA	Medical Question Answering	Medical board exam QA	[USMLE, diagnostic QA, medical knowledge, multilingual]	[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	[single-cell annotation, biological QA, autonomous discovery]	[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	[SELFIES, GAN, property 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 prediction, link prediction, graph classification]	[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	[DFT, materials genome, high-throughput]	[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	[DFT relaxations, adsorption energy, graph neural networks]	[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	[leaderboards, materials methods, simulation]	[Method benchmarking, Leaderboard ranking]	[MAE, RMSE, Accuracy]	[]	[21] ⇒
2022-02-22	Quantum Computing Benchmarks (QML)	Quantum Computing	Quantum algorithm performance evaluation	[quantum circuits, state preparation, error correction]	[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	[neural operators, CFD, FNO, DeepONet]	[Surrogate modeling]	[L2 error, MAE]	[FNO, DeepONet, U-Net]	[23] ⇒
None	SatImgNet	Remote Sensing	Satellite imagery classification	[land-use, zero-shot, multi-task]	[Image classification]	[Accuracy]	[]	[24] ⇒
2023-07-19	ClimateLearn	Climate Science; Forecasting	ML for weather and climate modeling	[medium-range forecasting, ERA5, data-driven]	[Forecasting]	[RMSE, Anomaly correlation]	[CNN baselines, ResNet variants]	[25] ⇒

Continued on next page

Date	Name	Domain	Focus	Keyword	Task Types	Metrics	Models	Citation
2022-06-09	BIG-Bench (Beyond the Imitation Game Benchmark)	NLP; AI Evaluation	Diverse reasoning and generalization tasks	[few-shot, multitask, bias analysis]	[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	[ConceptNet, multiple-choice, adversarial]	[Multiple choice]	[Accuracy]	[BERT-large, RoBERTa, GPT-3]	[27] ⇒
2019-07-24	Winogrande	NLP; Commonsense	Winograd Schema-style pronoun resolution	[adversarial, pronoun resolution]	[Pronoun resolution]	[Accuracy, AUC]	[RoBERTa, BERT, GPT-2]	[28] ⇒

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. r. .-. c. n. 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`.