The Ouroboros of Benchmarking: Reasoning Evaluation in an Era of Saturation

İbrahim Ethem Deveci
Department of Cognitive Science
Graduate School of Informatics
Middle East Technical University
Ankara, Turkey
ethem.deveci@metu.edu.tr

Duygu Ataman
Department of Cognitive Science
Graduate School of Informatics
Middle East Technical University
Ankara, Turkey
dataman@metu.edu.tr

Abstract

The rapid rise of Large Language Models (LLMs) and Large Reasoning Models (LRMs) has been accompanied by an equally rapid increase of benchmarks used to assess them. However, due to both improved model competence resulting from scaling and novel training advances as well as likely many of these datasets being included in pre or post training data, results become saturated, driving a continuous need for new and more challenging replacements. In this paper, we discuss whether surpassing a benchmark truly demonstrates reasoning ability or are we simply tracking numbers divorced from the capabilities we claim to measure? We present an investigation focused on three model families, OpenAI, Anthropic, and Google, and how their reasoning capabilities across different benchmarks evolve over the years. We also analyze performance trends over the years across different reasoning tasks and discuss the current situation of benchmarking and remaining challenges. By offering a comprehensive overview of benchmarks and reasoning tasks, our work aims to serve as a first reference to ground future research in reasoning evaluation and model development.

1 Introduction

Benchmarks have long played a central role in evaluating and comparing machine learning models [1]. As models scale up in size and capability, particularly Large Language Models (LLMs) and the specialized Large Reasoning Models (LRMs), many benchmarks quickly saturate, often reaching or surpassing human-level performance. Whether this saturation is driven primarily by improved model capability or dataset contamination is generally unknown. Nevertheless, this quick saturation forces the development of new and more challenging benchmarks that could be used to further compare new model families. In this paper, we investigate several key research questions: How effective are current benchmarks at measuring model capabilities, and does surpassing a benchmark reliably indicate genuine reasoning?

To examine these questions, we select three model families, OpenAI, Anthropic, and Google, and compile performance data from official sources [2–22]. We gather a comprehensive list of 52 benchmarks used in evaluating these models and classify them according to the types of reasoning they aim to evaluate. Analyzing performance trends over the years, we highlight where models improve, where they struggle, and what these trends reveal about the current state of benchmarking. Finally, we discuss the implications of the saturation cycle and emphasize the need for improved evaluation practices that more accurately capture model capabilities.

Our contributions are threefold: (1) we provide a curated list of reasoning benchmarks, classified by the types of reasoning they aim to assess (2) we analyze performance trends over the years to assess

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Evaluating the Evolving LLM Lifecycle: Benchmarks, Emergent Abilities, and Scaling.

benchmarking effectiveness; (3) we examine current landscape of existing benchmarks, identifying which benchmarks have reached high performance thresholds and which seem to remain unsolved.

By situating our analysis within the broader evaluation landscape, our work collects evidence to emphasize the need for reasoning tasks that are more representative of the nature of reasoning process and target evaluation beyond downstream accuracy.

2 Benchmark Landscape and Categorization

In order to provide a general analysis of how the creation and adoption of reasoning benchmarks have evolved over time, we examine three model families and compile the set of benchmarks employed to evaluate them. Our aim is to provide a comprehensive overview of current benchmarking practices and to trace how the creation and adoption of benchmarks have evolved over time. The complete list of benchmarks, their assigned reasoning types, and short summaries can be found in Appendix A. To facilitate analysis, we categorize benchmarks into seven reasoning types: commonsense and logical reasoning, mathematical reasoning, multimodal reasoning, programming and coding, reading comprehension and question answering, reasoning with general knowledge, and LLM-specific capabilities such as safety, tool use, and instruction following. Figure 1 illustrates a marked increase in benchmark adoption for multimodal reasoning, mathematical reasoning, programming, reasoning with general knowledge, and LLM-specific benchmarks after 2023. In contrast, no new benchmarks in reading comprehension or commonsense reasoning were adopted by these model families during this period. While the literature contains several other benchmarks in these areas [23–29], our analysis shows they have not been utilized by any of the prominent model families. This likely reflects the evolving understanding of what constitutes reasoning in computational models, in accordance with their current capabilities and what the community deems important to evaluate. Since most models now have direct commercial applications, their performance in more applicable domains, such as coding and tool-use benchmarks, may also motivate the evaluation in certain categories of reasoning tasks.

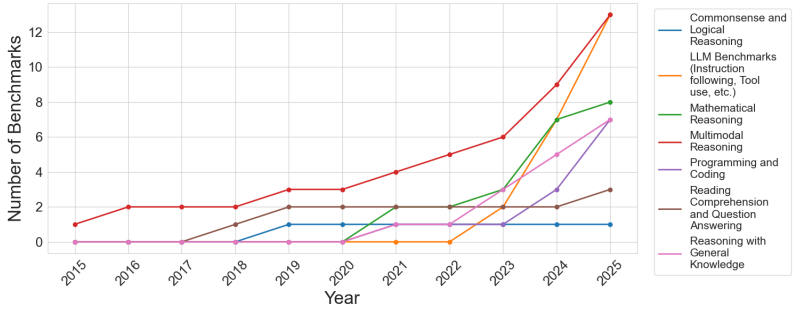


Figure 1: Number of benchmarks in different reasoning types over time.

3 Performance Trends Across Models

Across all three model families there is a consistent effort to develop newer models or architectural improvements to achieve higher benchmark performance. However, comparing performance across families is challenging, as each family often employs different benchmarks, and even within a single family, benchmarks used can vary between model iterations. This variation appears to stem from two main factors: first, certain benchmarks reach saturation due to high performance; second, benchmark updates or more challenging subsets are introduced, such as the transition from MATH to MATH-500 [30].

We observe a recurring pattern: once a model family achieves a high performance on a particular benchmark, subsequent models tend to use that benchmark less frequently or may discontinue its use entirely. This reflects both practical and conceptual considerations: benchmarks that no longer discriminate between models provide limited evaluative value, and benchmark selection increasingly reflects the evolving understanding of which reasoning tasks remain challenging for current architectures.

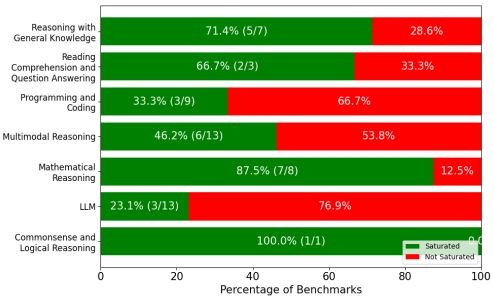
Interestingly, performance trends reveal consistent directional correlations across benchmarks within the same reasoning type. For example, when a model demonstrates improved performance on a benchmark, it generally shows corresponding improvements on other benchmarks of the same type, while lower performance on one benchmark tends to coincide with lower performance on others. Nevertheless, the extent of performance differs across benchmarks, potentially due to variations in problem complexity and the scaling limitations evident in smaller models, as seen within the OpenAI family. This pattern suggests that benchmarks within a reasoning type often capture overlapping aspects of reasoning, so that advances in a models' capabilities tend to propagate across related tasks. At the same time, variations in the magnitude of performance gains provide insight into the relative difficulty of different benchmarks within the same reasoning type. Detailed plots illustrating performance changes within model families for different reasoning types are provided in Appendix B.

Finally, we note that newer models generally achieve higher performance on previously low-scoring benchmarks. However, the limited overlap of common benchmarks across model families complicates cross-family comparisons. This raises a critical question: if benchmarks are intended to evaluate and compare model capabilities, why are they not consistently adopted or reported across families? If benchmarks are intended to provide a shared measure of capability, their fragmented and selective use undermines that goal and exemplifies the need for more standardized, representative, and domain-informed evaluation frameworks.

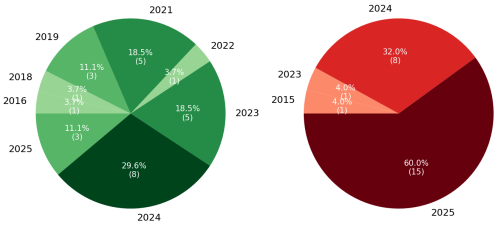
4 Performance of Models within Benchmarks

We collect all reported model performances across benchmarks and analyze saturation by defining it as whether a model has achieved at least 80% accuracy on the given benchmark. Out of the full set of benchmarks, we find that 27 benchmarks surpass this threshold in at least one model family, while 25 benchmarks never reach it. The majority of "solved" benchmarks belong to commonsense and logical reasoning, mathematical reasoning, reasoning with general knowledge, and reading comprehension and question answering. By contrast, benchmarks targeting LLM-specific capabilities and programming and coding remain comparatively difficult, with few instances of performance above 80%.

We then examine the release years of benchmarks that never surpass the 80% threshold. The distribution is striking: 60% of unsolved benchmarks were introduced in 2025, 32% in 2024, and only two pre-2023 benchmarks remain unsolved, which are ActivityNet [31] and EgoSchema [32], both multimodal reasoning benchmarks. This distribution suggests a clear trend. Nearly all benchmarks released prior to 2025 have already been surpassed by at least one model family, indicating rapid saturation. By contrast, the benchmarks still below the threshold overwhelmingly correspond to the most recently introduced evaluation tasks.



(a) Distribution of benchmarks that models surpassed 80% threshold and those not yet surpassed, grouped by reasoning type.



(b) Release years of benchmarks relative to the 80% threshold: left pie shows surpassed benchmarks, right pie shows unsolved benchmarks.

Figure 2: Benchmark saturation dynamics.

This temporal pattern highlights the central dynamic of the saturation cycle: older benchmarks are rapidly mastered and lose discriminative power, while newly introduced benchmarks become the standards for demonstrating progress. Nearly all unsolved benchmarks are recent, highlighting both the accelerating pace of benchmark creation and the difficulty of maintaining evaluations

that remain challenging over time. Yet this difficulty seems only temporary. It is highly plausible that within one or two years many of these currently unsolved benchmarks will also be surpassed, at which point model families will shift to alternative or newly designed evaluations to preserve differentiation. Crucially, this pattern reflects the fact that performance gains are often specific to individual benchmarks rather than to the broader reasoning type they are intended to assess. As the analyses indicate, while models often perform consistently and even strongly on benchmarks within a domain, the introduction of a more challenging, novel benchmark frequently leads to a drop in performance. This pattern may arise from the increased difficulty of the new benchmark, or from contamination that inflated performance on earlier benchmarks without truly reflecting generalizable reasoning ability. This situation raises the question of whether what appears as "reasoning ability" is often tied more to benchmark design and prior exposure than to robust mastery of the reasoning type itself. This saturation cycle casts doubt on the long-term evaluation value of benchmarks.

5 Discussion: Limitations of Current Benchmarking

Our analysis of three model families demonstrates that benchmark performance has generally increased over time, with newer models achieving higher scores across most reasoning types and benchmarks. However, given that many benchmarks have already been surpassed with high accuracy, we would like to highlight a question originally posed in [25] regarding commonsense reasoning, reframed here for reasoning in general: *Have neural language models successfully acquired reasoning, or are we overestimating the true capabilities of machine reasoning?* Several studies in the literature show that these models still perform poorly when required to generalize to longer contexts or handle tasks requiring inductive and compositional reasoning [33–38]. This discrepancy suggests a limitation of current benchmarking practices: improvements in benchmark scores do not necessarily reflect generalizable reasoning ability.

We believe this discrepancy can be reduced by developing more sophisticated, task-specific evaluation metrics that capture intermediate reasoning steps or different modes of error. Additionally, formalizing reasoning for different task types can support these efforts, enabling more structured analyses and clearer assessment of models' reasoning abilities. Such a formalization enables structured representations of diverse reasoning types and their interrelationships [39–41], and facilitates the design of layered, targeted evaluation procedures that assess specific reasoning capabilities rather than merely reporting overall accuracy. Furthermore, formal reasoning frameworks can support the development of algorithms that deliver structured feedback to models, guiding the refinement of their reasoning abilities. By integrating formalized reasoning with task-specific evaluations, benchmarking can be conducted in a more targeted and informative manner.

6 Limitations

The analysis in our study focuses on 52 benchmarks used by the three model families. Other model families and reasoning-focused models are not fully explored because including them, along with more than two hundred benchmarks identified from other model families and several studies evaluating different types of reasoning in large models, would create a combinatorial explosion of comparisons. This restriction was necessary to maintain the scope of our work on a qualitative evaluation of benchmark design and adoption rather than an exhaustive quantitative analysis of all models and benchmarks. A comprehensive comparison across a wider range of models and benchmarks is left for future work.

7 Conclusion

In this work, we analyze 52 benchmarks across three model families, covering multiple reasoning types. Our study reveals the rapid saturation of older benchmarks, selective adoption of new ones, and temporal dynamics that govern the utility of benchmarks in evaluating model performance. While model performance generally improves over time and correlations within reasoning types indicate overlapping evaluation properties, the introduction of more challenging benchmarks generally resets performance, suggesting that apparent reasoning ability is influenced more by extrinsic factors than by mastering the reasoning itself, as supported by other studies. This saturation cycle highlights the limitations of current practices: benchmarks provide only a partial view of model reasoning.

Meaningful progress requires formalized reasoning tasks, layered evaluation procedures, and task-specific metrics that go beyond accuracy scores.

References

[1] Thomas Liao, Rohan Taori, Deborah Raji, and Ludwig Schmidt. Are we learning yet? a meta review of evaluation failures across machine learning. In J. Vanschoren and S. Yeung, editors, *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021.

[2] Anthropic. Introducing the next generation of claude, March 2024. Accessed: 2025-08-28.

[3] Anthropic. Claude 3.5 sonnet, June 2024. Accessed: 2025-08-28.

[4] Anthropic. Introducing claude 4, May 2025. Accessed: 2025-08-28.

[5] Anthropic. Introducing claude 3.5 haiku, October 2024. Accessed: 2025-08-28.

[6] Anthropic. Claude 3.7 sonnet and claude code, February 2025. Accessed: 2025-08-28.

[7] Anthropic. Claude opus 4.1, August 2025. Accessed: 2025-08-28.

[8] Google DeepMind. Gemini 2.5 flash-lite, June 2025. Accessed: 2025-08-28.

[9] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities, 2025.

[10] Google DeepMind. Gemini 2.5: Our most intelligent ai model, March 2025. Accessed: 2025-08-28.

[11] Gemini Team, Petko Georgiev, Ving Ian Lei, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context, 2024.

[12] Gemini Team, Rohan Anil, Sebastian Borgeaud, et al. Gemini: A family of highly capable multimodal models, 2025.

[13] OpenAI. Openai o1-mini: Advancing cost-efficient reasoning, September 2024. Accessed: 2025-08-28.

[14] OpenAI. Introducing gpt-4.1 in the api, April 2025. Accessed: 2025-08-28.

[15] OpenAI. Introducing gpt-4.5, February 2025. Accessed: 2025-08-28.

[16] OpenAI. gpt-oss-120b & gpt-oss-20b model card, August 2025. Accessed: 2025-08-28.

[17] OpenAI. Introducing gpt-5, August 2025. Accessed: 2025-08-28.

[18] OpenAI. Model release notes. Accessed: 2025-08-28.

[19] OpenAI. Introducing openai o3 and o4-mini, April 2025. Accessed: 2025-08-28.

[20] OpenAI. Gpt-4o mini: Advancing cost-efficient intelligence, July 2024. Accessed: 2025-08-28.

[21] OpenAI. Hello gpt-4o, May 2024. Accessed: 2025-08-28.

[22] OpenAI. Learning to reason with llms, September 2024. Accessed: 2025-08-28.

[23] Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jiasen Gao, and Yejin Choi. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7432–7439, 2020.

[24] Bill Yuchen Lin, Wangchunshu Zhou, Ming Shen, Pei Zhou, Chandra Bhagavatula, Yejin Choi, and Xiang Ren. CommonGen: A constrained text generation challenge for generative commonsense reasoning. In Trevor Cohn, Yulan He, and Yang Liu, editors, *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1823–1840, Online, November 2020. Association for Computational Linguistics.

[25] Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: an adversarial winograd schema challenge at scale. *Commun. ACM*, 64(9):99–106, August 2021.

[26] Alon Talmor, Ori Yoran, Ronan Le Bras, Chandra Bhagavatula, Yoav Goldberg, Yejin Choi, and Jonathan Berant. Commonsenseqa 2.0: Exposing the limits of ai through gamification, 2022.

[27] Andong Wang, Bo Wu, Sunli Chen, Zhenfang Chen, Haotian Guan, Wei-Ning Lee, Li Erran Li, and Chuang Gan. Sok-bench: A situated video reasoning benchmark with aligned open-world knowledge, 2024.

[28] Jian Liu, Leyang Cui, Hanmeng Liu, Dandan Huang, Yile Wang, and Yue Zhang. Logiqa: a challenge dataset for machine reading comprehension with logical reasoning. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, IJCAI'20, 2021.

[29] Weihao Yu, Zihang Jiang, Yanfei Dong, and Jiashi Feng. Reclor: A reading comprehension dataset requiring logical reasoning. In *International Conference on Learning Representations*, 2020.

[30] Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. In J. Vanschoren and S. Yeung, editors, *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021.

[31] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 961–970, 2015.

[32] Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding, 2023.

[33] Nouha Dziri, Ximing Lu, Melanie Sclar, Xiang Lorraine Li, Liwei Jiang, Bill Yuchen Lin, Peter West, Chandra Bhagavatula, Ronan Le Bras, Jena D. Hwang, Soumya Sanyal, Sean Welleck, Xiang Ren, Allyson Ettinger, Zaid Harchaoui, and Yejin Choi. Faith and fate: limits of transformers on compositionality. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

[34] Iman Mirzadeh, Keivan Alizadeh, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio, and Mehrdad Farajtabar. Gsm-symbolic: Understanding the limitations of mathematical reasoning in large language models, 2025.

[35] Parshin Shojaee, Iman Mirzadeh, Keivan Alizadeh, Maxwell Horton, Samy Bengio, and Mehrdad Farajtabar. The illusion of thinking: Understanding the strengths and limitations of reasoning models via the lens of problem complexity, 2025.

[36] Jackson Petty, Michael Y. Hu, Wentao Wang, Shauli Ravfogel, William Merrill, and Tal Linzen. Relic: Evaluating compositional instruction following via language recognition, 2025.

[37] S. Bedi, Y. Jiang, P. Chung, S. Koyejo, and N. Shah. Fidelity of medical reasoning in large language models. *JAMA Network Open*, 8(8):e2526021, 2025.

[38] Karthik Valmeekam, Kaya Stechly, Atharva Gundawar, and Subbarao Kambhampati. A systematic evaluation of the planning and scheduling abilities of the reasoning model o1. *Transactions on Machine Learning Research*, 2025.

[39] P. N. Johnson-Laird. *Mental models: towards a cognitive science of language, inference, and consciousness*. Harvard University Press, USA, 1986.

[40] Patrick Blackburn and Johannes Bos. *Representation and Inference for Natural Language: A First Course in Computational Semantics*. Center for the Study of Language and Information, Stanford, Calif., 2005.

[41] Brenden M. Lake, Tomer D. Ullman, Joshua B. Tenenbaum, and Samuel J. Gershman. Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40:e253, 2017.

[42] Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: Can a machine really finish your sentence? In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy, July 2019. Association for Computational Linguistics.

[43] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding, 2021.

[44] Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc Le, Ed Chi, Denny Zhou, and Jason Wei. Challenging BIG-bench tasks and whether chain-of-thought can solve them. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13003–13051, Toronto, Canada, July 2023. Association for Computational Linguistics.

[45] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding, 2021.

[46] Long Phan, Alice Gatti, Ziwen Han, et al. Humanity's last exam, 2025.

[47] Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David Ifeoluwa Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, Raymond Ng, Shayne Longpre, Sebastian Ruder, Wei-Yin Ko, Antoine Bosselut, Alice Oh, Andre Martins, Leshem Choshen, Daphne Ippolito, Enzo Ferrante, Marzieh Fadaee, Beyza Ermis, and Sara Hooker. Global MMLU: Understanding and addressing cultural and linguistic biases in multilingual evaluation. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18761–18799, Vienna, Austria, July 2025. Association for Computational Linguistics.

[48] David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. Gpqa: A graduate-level google-proof q&a benchmark, 2023.

[49] Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark, 2024.

[50] Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge, 2018.

[51] Omer Goldman, Uri Shaham, Dan Malkin, Sivan Eiger, Avinatan Hassidim, Yossi Matias, Joshua Maynez, Adi Mayrav Gilady, Jason Riesa, Shruti Rijhwani, Laura Rimell, Idan Szpektor, Reut Tsarfaty, and Matan Eyal. Eclektic: a novel challenge set for evaluation of cross-lingual knowledge transfer, 2025.

[52] Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2368–2378, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[53] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems, 2021.

[54] Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. Language models are multilingual chain-of-thought reasoners, 2022.

[55] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts, 2024.

[56] Elliot Glazer, Ege Erdil, Tamay Besiroglu, Diego Chicharro, Evan Chen, Alex Gunning, Caroline Falkman Olsson, Jean-Stanislas Denain, Anson Ho, Emily de Oliveira Santos, Olli Järviniemi, Matthew Barnett, Robert Sandler, Matej Vrzala, Jaime Sevilla, Qiuyu Ren, Elizabeth Pratt, Lionel Levine, Grant Barkley, Natalie Stewart, Bogdan Grechuk, Tetiana Grechuk, Shreepranav Varma Enugandla, and Mark Wildon. Frontiermath: A benchmark for evaluating advanced mathematical reasoning in ai, 2024.

[57] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi, 2024.

[58] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images, 2016.

[59] Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2263–2279, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[60] Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. Docvqa: A dataset for vqa on document images, 2021.

[61] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read, 2019.

[62] Kairui Hu, Penghao Wu, Fanyi Pu, Wang Xiao, Yuanhan Zhang, Xiang Yue, Bo Li, and Ziwei Liu. Video-mmmu: Evaluating knowledge acquisition from multi-discipline professional videos, 2025.

[63] Piotr Padlewski, Max Bain, Matthew Henderson, Zhongkai Zhu, Nishant Relan, Hai Pham, Donovan Ong, Kaloyan Aleksiev, Aitor Ormazabal, Samuel Phua, Ethan Yeo, Eugenie Lamprecht, Qi Liu, Yuqi Wang, Eric Chen, Deyu Fu, Lei Li, Che Zheng, Cyprien de Masson d'Autume, Dani Yogatama, Mikel Artetxe, and Yi Tay. Vibe-eval: A hard evaluation suite for measuring progress of multimodal language models, 2024.

[64] Jonathan Roberts, Mohammad Reza Taesiri, Ansh Sharma, Akash Gupta, Samuel Roberts, Ioana Croitoru, Simion-Vlad Bogolin, Jialu Tang, Florian Langer, Vyas Raina, Vatsal Raina, Hanyi Xiong, Vishaal Udandarao, Jingyi Lu, Shiyang Chen, Sam Purkis, Tianshuo Yan, Wenye Lin, Gyungin Shin, Qiaochu Yang, Anh Totti Nguyen, David I. Atkinson, Aaditya Baranwal, Alexandru Coca, Mikah Dang, Sebastian Dziadzio, Jakob D. Kunz, Kaiqu Liang, Alexander Lo, Brian Pulfer, Steven Walton, Charig Yang, Kai Han, and Samuel Albanie. Zerobench: An impossible visual benchmark for contemporary large multimodal models, 2025.

[65] Zirui Wang, Mengzhou Xia, Luxi He, Howard Chen, Yitao Liu, Richard Zhu, Kaiqu Liang, Xindi Wu, Haotian Liu, Sadhika Malladi, Alexis Chevalier, Sanjeev Arora, and Danqi Chen. Charxiv: Charting gaps in realistic chart understanding in multimodal llms. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 113569–113697. Curran Associates, Inc., 2024.

[66] Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, Yu Su, Wenhu Chen, and Graham Neubig. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark, 2025.

[67] Google DeepMind. Gemini robotics: Bringing ai into the physical world, 2025. Accessed: 2025-08-29.

[68] Carlos E. Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. Swe-bench: Can language models resolve real-world github issues?, 2024.

[69] Stanford University and Laude Institute. Terminal-bench: A benchmark for ai agents in terminal environments, 2025. Accessed: 2025-08-29.

[70] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. Evaluating large language models trained on code, 2021.

[71] Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. Livecodebench: Holistic and contamination free evaluation of large language models for code, 2024.

[72] Aider. o1 tops aider's new polyglot leaderboard, 2024. Accessed: 2025-08-29.

[73] Samuel Miserendino, Michele Wang, Tejal Patwardhan, and Johannes Heidecke. Swe-lancer: Can frontier llms earn $1 million from real-world freelance software engineering?, 2025.

[74] Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. τ-bench: A benchmark for tool-agent-user interaction in real-world domains, 2024.

[75] Victor Barres, Honghua Dong, Soham Ray, Xujie Si, and Karthik Narasimhan. τ^2-bench: Evaluating conversational agents in a dual-control environment, 2025.

[76] Shunyu Yao, Howard Chen, Austin W. Hanjie, Runzhe Yang, and Karthik Narasimhan. Collie: Systematic construction of constrained text generation tasks, 2023.

[77] Jason Wei, Nguyen Karina, Hyung Won Chung, Yunxin Joy Jiao, Spencer Papay, Amelia Glaese, John Schulman, and William Fedus. Measuring short-form factuality in large language models, 2024.

[78] Alon Jacovi, Andrew Wang, Chris Alberti, Connie Tao, Jon Lipovetz, Kate Olszewska, Lukas Haas, Michelle Liu, Nate Keating, Adam Bloniarz, Carl Saroufim, Corey Fry, Dror Marcus, Doron Kukliansky, Gaurav Singh Tomar, James Swirhun, Jinwei Xing, Lily Wang, Madhu Gurumurthy, Michael Aaron, Moran Ambar, Rachana Fellinger, Rui Wang, Zizhao Zhang, Sasha Goldshtein, and Dipanjan Das. The facts grounding leaderboard: Benchmarking llms' ability to ground responses to long-form input, 2025.

[79] Jason Wei, Zhiqing Sun, Spencer Papay, Scott McKinney, Jeffrey Han, Isa Fulford, Hyung Won Chung, Alex Tachard Passos, William Fedus, and Amelia Glaese. Browsecomp: A simple yet challenging benchmark for browsing agents, 2025.

[80] Lucen Zhong, Zhengxiao Du, Xiaohan Zhang, Haiyi Hu, and Jie Tang. Complexfuncbench: Exploring multi-step and constrained function calling under long-context scenario, 2025.

[81] Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models, 2023.

[82] Yun He, Di Jin, Chaoqi Wang, Chloe Bi, Karishma Mandyam, Hejia Zhang, Chen Zhu, Ning Li, Tengyu Xu, Hongjiang Lv, Shruti Bhosale, Chenguang Zhu, Karthik Abinav Sankararaman, Eryk Helenowski, Melanie Kambadur, Aditya Tayade, Hao Ma, Han Fang, and Sinong Wang. Multi-if: Benchmarking llms on multi-turn and multilingual instructions following, 2024.

[83] Jinhyuk Lee, Anthony Chen, Zhuyun Dai, Dheeru Dua, Devendra Singh Sachan, Michael Boratko, Yi Luan, Sébastien M. R. Arnold, Vincent Perot, Siddharth Dalmia, Hexiang Hu, Xudong Lin, Panupong Pasupat, Aida Amini, Jeremy R. Cole, Sebastian Riedel, Iftekhar Naim, Ming-Wei Chang, and Kelvin Guu. Can long-context language models subsume retrieval, rag, sql, and more?, 2024.

[84] Kaustubh Deshpande, Ved Sirdeshmukh, Johannes Baptist Mols, Lifeng Jin, Ed-Yeremai Hernandez-Cardona, Dean Lee, Jeremy Kritz, Willow E. Primack, Summer Yue, and Chen Xing. MultiChallenge: A realistic multi-turn conversation evaluation benchmark challenging to frontier LLMs. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Findings of the Association for Computational Linguistics: ACL 2025*, pages 18632–18702, Vienna, Austria, July 2025. Association for Computational Linguistics.

[85] Rahul K. Arora, Jason Wei, Rebecca Soskin Hicks, Preston Bowman, Joaquin Quiñonero-Candela, Foivos Tsimpourlas, Michael Sharman, Meghan Shah, Andrea Vallone, Alex Beutel, Johannes Heidecke, and Karan Singhal. Healthbench: Evaluating large language models towards improved human health, 2025.

A Reasoning Benchmarks

Table 1: Taxonomy of benchmarks used in this study.

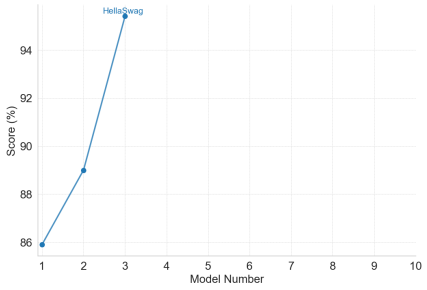
Benchmark	Reasoning Type	Year	Explanation
HellaSwag [42]	Commonsense and Logical Reasoning	2019	Multiple-choice task: choose the most plausible sentence continuation.
MMLU [43]	Reasoning with General Knowledge	2021	Multiple-choice task: answer questions across 57 domains to test knowledge and problem-solving.
Big-Bench-Hard [44]	Reasoning with General Knowledge	2023	Open-generation task: solve difficult BIG-Bench problems testing multi-step reasoning and problem-solving.
MMMLU [45]	Reasoning with General Knowledge	2024	Multiple-choice task: answer 57 domain questions translated into 14 languages to test multilingual knowledge and problem-solving.
Humanity's Last Exam [46]	Reasoning with General Knowledge	2025	Multi-modal task: answer closed-ended questions across many subjects to test verifiable knowledge.
Global MMLU (Lite) [47]	Reasoning with General Knowledge	2025	Multiple-choice task: answer 42-language questions with culturally sensitive labeling to test equitable multilingual knowledge.
GPQA Diamond [48]	Reasoning with General Knowledge	2023	Multiple-choice task: answer 448 expert-level science questions in biology, physics, and chemistry that are Google-proof and highly challenging.
MMLU Pro [49]	Reasoning with General Knowledge	2024	Multiple-choice task: extended from MMLU, answer more challenging reasoning questions with 10 options across diverse domains.
ARC (AI2 Reasoning Challenge) [50]	Reading Comprehension and Question Answering	2018	Multiple-choice task: answer grade-school science questions requiring advanced knowledge and reasoning beyond simple retrieval.
ECLeKTic [51]	Reading Comprehension and Question Answering	2025	Closed-book QA task: answer 12-language questions to test cross-lingual knowledge transfer.
DROP [52]	Reading Comprehension and Question Answering	2019	Open-ended QA task: answer 96k English questions requiring discrete reasoning over paragraph content.
GSM8K [53]	Mathematical Reasoning	2021	Open-ended QA task: solve grade-school problems requiring multi-step mathematical reasoning.
MATH [30]	Mathematical Reasoning	2021	Open-ended QA: solve 12,500 challenging competition problems with step-by-step solutions to test advanced mathematical reasoning.
MATH 500 [30]	Mathematical Reasoning	2024	Open-ended QA: Challenging subset of MATH benchmark.

Benchmark	Reasoning Type	Year	Explanation
MGSM [54]	Mathematical Reasoning	2023	Open-ended QA: solve 250 GSM8K problems translated into 10 languages.
MathVista [55]	Mathematical Reasoning	2024	Open-ended multimodal QA: solve 6,141 math problems requiring visual and compositional reasoning.
AIME 2024	Mathematical Reasoning	2024	Open-ended QA: solve challenging competition-level mathematics problems.
AIME 2025	Mathematical Reasoning	2025	Open-ended QA: solve challenging competition-level mathematics problems.
FrontierMath [56]	Mathematical Reasoning	2024	Open-ended QA: tests advanced mathematical reasoning across diverse and expert-level domains, requiring multi-step problem solving and deep mathematical knowledge.
MMMU [57]	Multimodal Reasoning	2024	Question answering task: multimodal multiple-choice and open-ended questions across 30 subjects requiring advanced reasoning and domain-specific knowledge.
AI2D [58]	Multimodal Reasoning	2016	Open-ended QA: multimodal questions with 5,000 diagrams and 15,000 Q&A pairs requiring diagram structure understanding and reasoning.
ChartQA [59]	Multimodal Reasoning	2022	Open-ended QA: multimodal questions with 32.7K chart-based problems requiring visual and logical reasoning.
EgoSchema [32]	Multimodal Reasoning	2023	Multiple-choice QA: multimodal questions with 5,000 long-form video clips requiring understanding of human activity and temporal reasoning.
DocVQA [60]	Multimodal Reasoning	2021	Open-ended QA: multimodal questions with 50,000 document images requiring reading and interpreting document layout and structure.
TextVQA [61]	Multimodal Reasoning	2019	Open-ended QA: multimodal questions with 45,336 images requiring reading and reasoning about embedded text.
VideoMMMU [62]	Multimodal Reasoning	2025	Open-ended QA: multimodal questions with 300 expert-level videos and 900 Q&A pairs assessing knowledge acquisition through perception, comprehension, and adaptation.
Vibe-Eval [63]	Multimodal Reasoning	2024	Open-ended QA: multimodal questions, testing visual understanding and multimodal chat capabilities.
ZeroBench [64]	Multimodal Reasoning	2025	Open-ended QA: multimodal questions with 434 visual reasoning problems designed to be impossible for current LMMs.
CharXiv [65]	Multimodal Reasoning	2024	Open-ended QA: multimodal questions with 2,323 charts requiring descriptive analysis and complex reasoning.

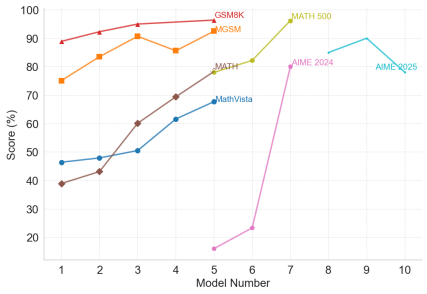
Benchmark	Reasoning Type	Year	Explanation
MMMU Pro [66]	Multimodal Reasoning	2025	QA task: multimodal multiple-choice and open-ended questions, extended from MMMU, testing integrated visual and textual reasoning.
ActivityNet [31]	Multimodal Reasoning	2015	Multiple-choice and open-ended QA: evaluates recognition and understanding of complex human activities in untrimmed videos, testing visual perception and temporal reasoning.
ERQA [67]	Multimodal Reasoning	2025	Multiple-choice QA: evaluates embodied reasoning and spatial understanding in real-world scenarios, requiring models to integrate text and visual inputs to select the correct answer.
SWE-bench Verified [68]	Programming and Coding	2024	Open-ended QA: answer 2,294 software engineering problems requiring multi-file code edits and complex reasoning.
Terminal-bench [69]	Programming and Coding	2025	Open-ended QA: answer complex tasks in terminal environments using text-based commands and reasoning.
HumanEval [70]	Programming and Coding	2021	Open-ended QA: answer Python programming problems from docstrings requiring functional code synthesis.
LiveCode Bench [71]	Programming and Coding	2025	Open-ended QA: answer 600+ coding problems from contests, testing generation, self-repair, execution, and test prediction.
Aider Polygot [72]	Programming and Coding	2024	Open-ended QA: answer 225 difficult coding problems in C++, Go, Java, JavaScript, Python, and Rust.
SWE-Lancer [73]	Programming and Coding	2025	Open-ended QA: answer 1,400 freelance software engineering tasks, including implementation and managerial decisions, with real-world evaluation.
SWE-Lancer Diamond [73]	Programming and Coding	2025	Open-ended QA: answer tasks from the public SWE-Lancer Diamond split, including implementation and managerial software engineering problems.
TAU-bench [74]	Tool Use – LLM	2024	Open-ended QA: tests reasoning, consistency, and rule-following in dynamic, tool-assisted human-agent interactions.
TAU2-bench [75]	Tool Use – LLM	2025	Open-ended QA: tests multi-turn reasoning, coordination, and communication in dual-control environments where both agent and user act with tools.
COLLIE [76]	Constrained Text Generation – LLM	2023	Open-ended QA: answer 2,080 prompts requiring constrained text generation with compositional, grammar-based, and reasoning challenges.
SimpleQA [77]	Factuality – LLM	2024	Factual QA benchmark designed to test factual accuracy and knowledge calibration.

Benchmark	Reasoning Type	Year	Explanation
FACTS Grounding [78]	Factuality – LLM	2024	Open-ended QA: answer questions requiring LLMs to generate factually accurate and well-grounded responses from provided source material.
BrowseComp [79]	Factuality – LLM	2025	Open-ended QA: answer 1,266 questions by persistently navigating the internet to find hard-to-locate information.
ComplexFunc Bench [80]	Tool Use – LLM	2025	Open-ended QA: answer complex function-calling tasks in five real-world scenarios requiring multi-step reasoning, parameter management, and long-context handling.
IFEval [81]	Instruction Following – LLM	2023	Open-ended QA: answer 500 prompts requiring LLMs to follow verifiable natural language instructions.
Multi-IF [82]	Instruction Following – LLM	2024	Open-ended QA: answer 4,501 multilingual multi-turn prompts requiring accurate instruction-following across languages and conversation turns.
LOFT [83]	Long-Context – LLM	2024	Open-ended QA: answer real-world tasks requiring reasoning and in-context retrieval over millions of tokens.
Graphwalks [14]	Long-Context – LLM	2025	Open-ended QA: perform multi-hop reasoning across a graph of millions of tokens to answer questions requiring breadth-first traversal.
Multi Challenge [84]	Multi-turn Conversation – LLM	2025	Open-ended QA: answer multi-turn conversation prompts requiring instruction-following, context management, and in-context reasoning.
HealthBench [85]	Safety – LLM	2025	Open-ended QA: evaluates LLMs on multi-turn healthcare conversations, requiring factual reasoning, safety awareness, and context-sensitive decision-making across diverse medical contexts.

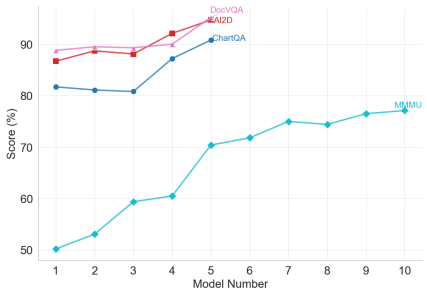
B Performance of Models



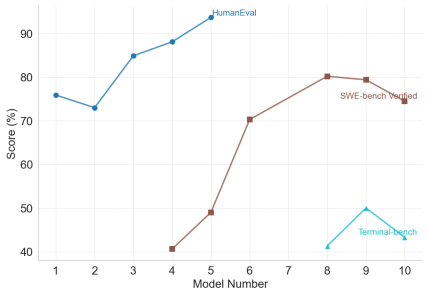
(a) Commonsense and Logical Reasoning



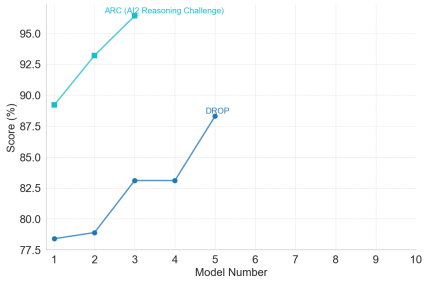
(b) Mathematical Reasoning



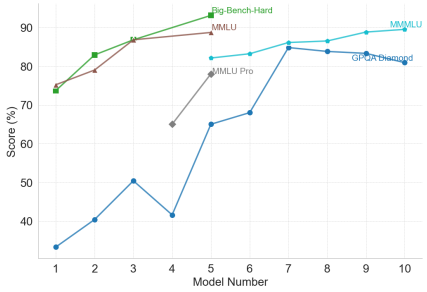
(c) Multimodal Reasoning



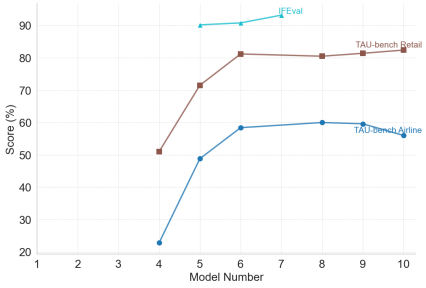
(d) Programming and Coding



(e) Reading Comprehension and QA

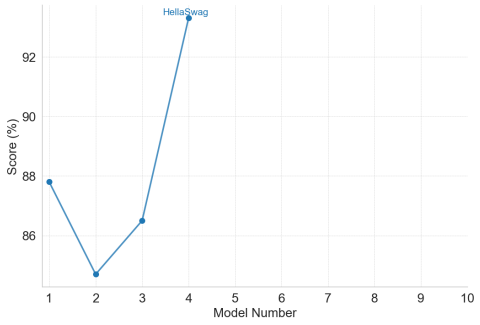


(f) Reasoning with General Knowledge

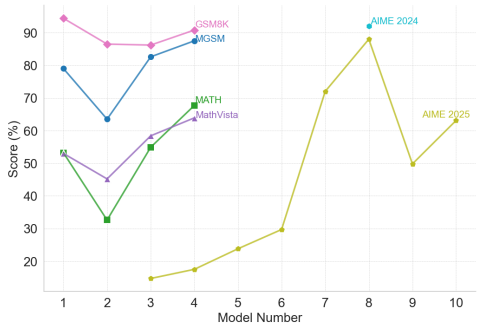


(g) LLM Benchmarks

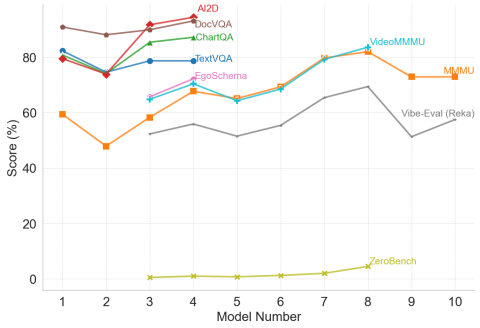
Figure 3: Performance of the Claude family on reasoning benchmarks by category. Model numbers and corresponding names are as follows: 1 – Claude 3 Haiku; 2 – Claude 3 Sonnet; 3 – Claude 3 Opus; 4 – Claude 3.5 Haiku; 5 – Claude 3.5 Sonnet; 6 – Claude 3.7 Sonnet; 7 – Claude 3.7 Sonnet (64K Extended Thinking); 8 – Claude Sonnet 4; 9 – Claude Opus 4; 10 – Claude Opus 4.1.



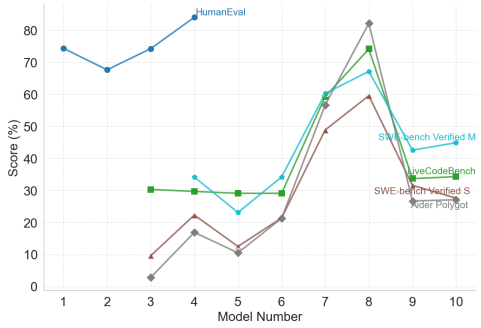
(a) Commonsense and Logical Reasoning



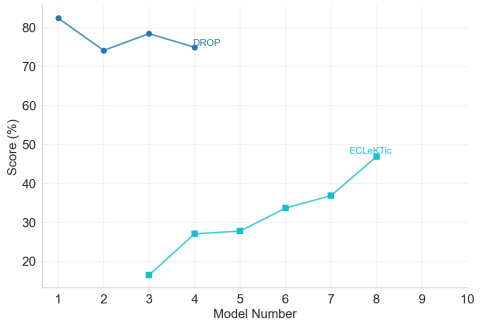
(b) Mathematical Reasoning



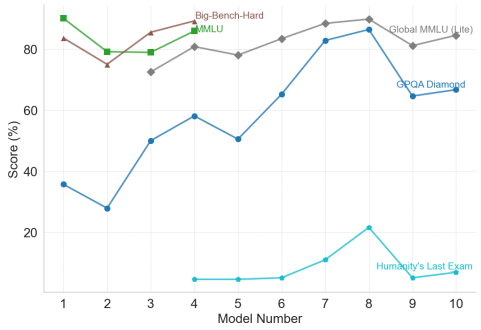
(c) Multimodal Reasoning



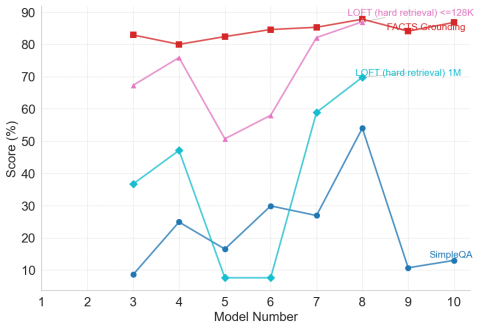
(d) Programming and Coding



(e) Reading Comprehension and QA

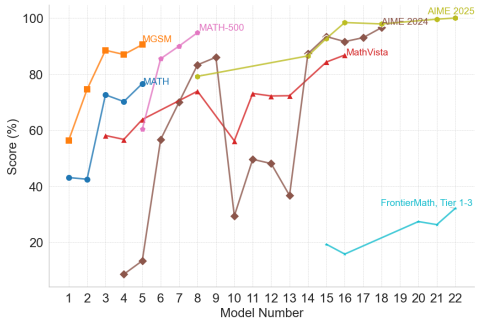


(f) Reasoning with General Knowledge

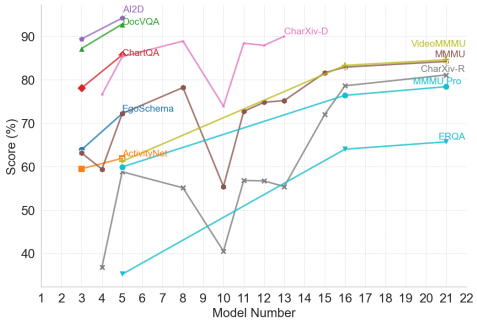


(g) LLM Benchmarks

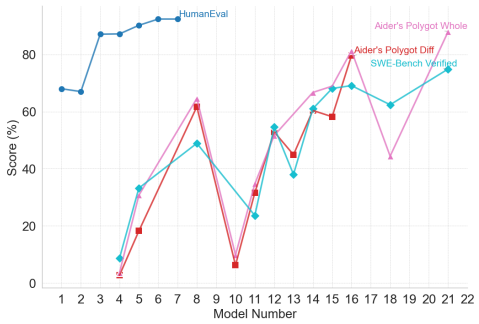
Figure 4: Performance of the Gemini family on reasoning benchmarks by category. Model numbers and corresponding names are as follows: 1 – Gemini Ultra; 2 – Gemini Pro; 3 – Gemini 1.5 Flash; 4 – Gemini 1.5 Pro; 5 – Gemini 2.0 Flash-Lite; 6 – Gemini 2.0 Flash; 7 – Gemini 2.5 Flash; 8 – Gemini 2.5 Pro; 9 – Gemini 2.5 Flash Lite (no thinking); 10 – Gemini 2.5 Flash Lite (thinking).



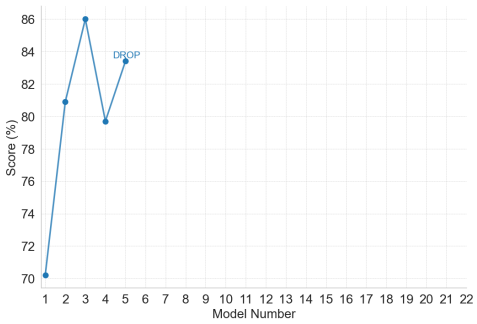
(a) Mathematical Reasoning



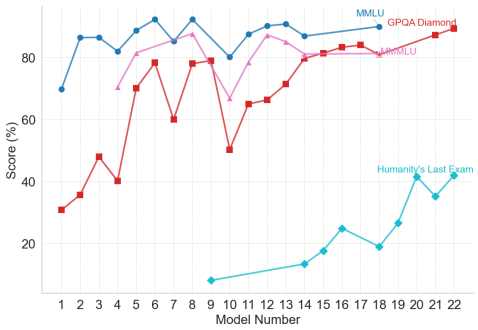
(b) Multimodal Reasoning



(c) Programming and Coding

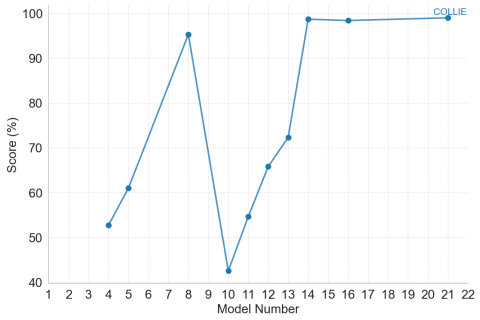


(d) Reading Comprehension and QA

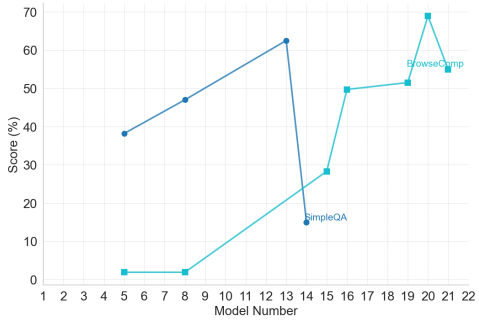


(e) Reasoning with General Knowledge

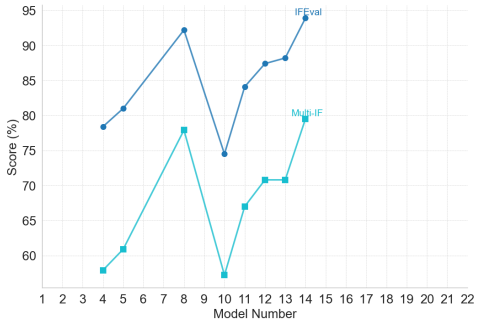
Figure 5: Performance of the GPT family on general reasoning benchmarks. Model numbers and corresponding names are as follows: 1 – GPT-3.5; 2 – GPT-4; 3 – GPT-4 Turbo; 4 – GPT-4o mini; 5 – GPT-4o; 6 – o1-preview; 7 – o1-mini; 8 – o1; 9 – o1-pro; 10 – GPT-4.1 nano; 11 – GPT-4.1 mini; 12 – GPT-4.1; 13 – GPT-4.5; 14 – o3-mini; 15 – o4-mini; 16 – o3; 17 – o3-pro; 18 – gpt-oss-120b; 19 – GPT-5 with Deep Research; 20 – ChatGPT Agent; 21 – GPT-5; 22 – GPT-5 Pro.



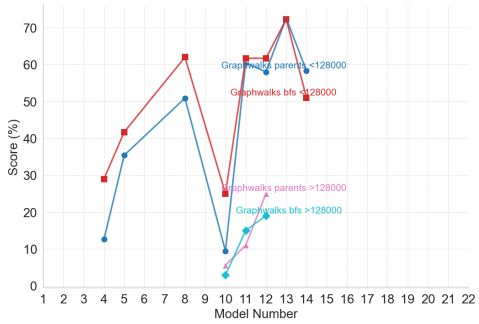
(a) Constrained Text Generation



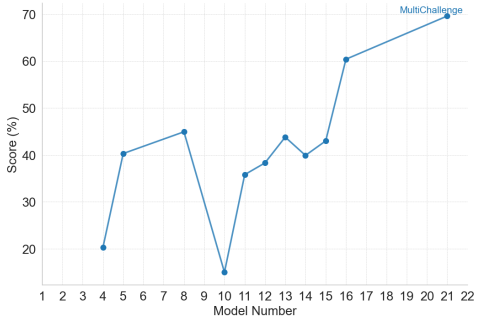
(b) Factuality



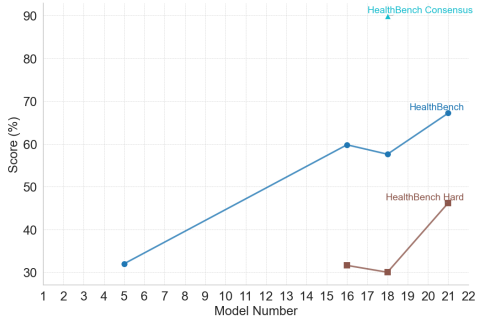
(c) Instruction Following



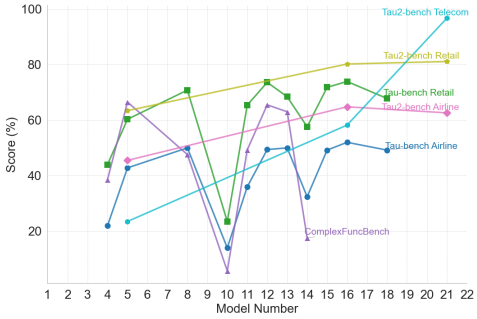
(d) Long Context



(e) Multi-turn Conversation



(f) Safety



(g) Tool Use

Figure 6: Performance of the GPT family on LLM-specific benchmarks. Model numbers and corresponding names are as follows: 1 – GPT-3.5; 2 – GPT-4; 3 – GPT-4 Turbo; 4 – GPT-4o mini; 5 – GPT-4o; 6 – o1-preview; 7 – o1-mini; 8 – o1; 9 – o1-pro; 10 – GPT-4.1 nano; 11 – GPT-4.1 mini; 12 – GPT-4.1; 13 – GPT-4.5; 14 – o3-mini; 15 – o4-mini; 16 – o3; 17 – o3-pro; 18 – gpt-oss-120b; 19 – GPT-5 with Deep Research; 20 – ChatGPT Agent; 21 – GPT-5; 22 – GPT-5 Pro.