Amazon Nova Premier: Technical Report and Model Card

Amazon Artificial General Intelligence

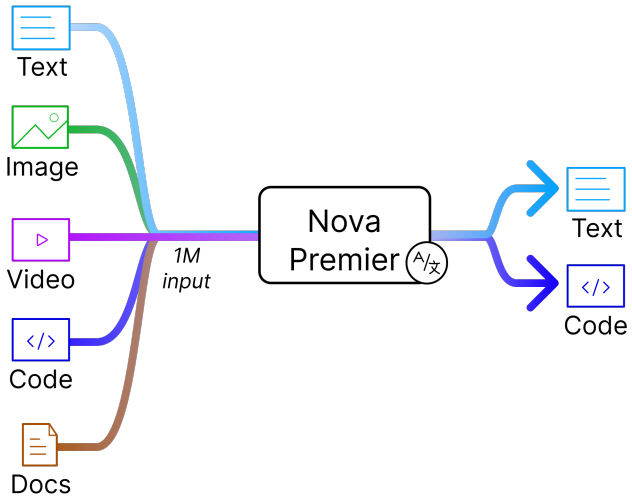

Figure 1: Amazon Nova Premier

Abstract

We present Amazon Nova Premier, our most capable multimodal foundation model and teacher for model distillation. Nova Premier processes text, images, and videos with a one-million token context window enabling analysis of large codebases, long documents, and long videos in a single prompt. It also enables customers to use Amazon Bedrock to create customized variants of Amazon Nova Pro, Nova Lite, and Nova Micro that maintain high accuracy while offering improved speed and cost efficiency. Like all Nova models, Nova Premier is built with integrated safety measures and responsible AI practices, maintaining our commitment to customer trust, security, and reliability. With Nova Premier, we further extend the capabilities and price-performance advantages of the Amazon Nova model family.

This report was published on April 30, 2025 as an addendum to "The Amazon Nova Family of Models: Technical Report and Model Card" [5].

1 Introduction

This technical report builds upon our earlier technical report of the Amazon Nova Family of Models [5] where we introduced Nova Pro, Nova Lite, Nova Micro, Nova Canvas, and Nova Reel. In this document, we introduce Nova Premier, our most capable multimodal understanding model for complex tasks and teacher for model distillation. Nova Premier has a context length of one million tokens, which enables the analysis of large codebases, documents longer than 400 pages, or 90-minute-long videos.

Nova Premier has also been optimized for high performance on a number of tasks important to Amazon customers, such as:

- Code generation;
- Retrieval-Augmented Generation (RAG), in which the model retrieves information from reliable, up-to-date sources, to ensure the accuracy of its responses;
- Video understanding, or interpreting the content of video scenes;
- Document understanding;
- Function calling, or producing outputs that elicit the correct responses from other application APIs;
- Agentic interactions, in which the model engages in multiturn interactions on the customer's behalf.

2 Performance of Nova Premier

In this section, we report benchmarking results for Nova Premier. We also report results for select publicly-available models by citing existing public results, as well as by measuring their performance when public results are not available.[1] We evaluate Nova Premier on a suite of automated public benchmarks to assess core capabilities for text, multimodal, and agentic capabilities.

2.1 Performance on Public Benchmarks

Table 1 summarizes the quantitative results of Nova Premier compared to Nova Pro, as well as select public models in its intelligence tier on text, multimodal, and agentic capabilities. When available, we reference the highest publicly-reported numbers for each benchmark from the official technical reports and websites for Claude Sonnet and GPT-4.5 models. Nova Premier demonstrates strong performance across all benchmarks, showcasing its advanced core intelligence, reasoning, and multimodal capabilities. In cases for which the result is a simple average of binary scores, we assume a Gaussian distribution for the sample and approximate the 95% confidence interval as $1.96 \times \sqrt{\frac{s \times (1-s)}{n}}$ where s is the measured score for the benchmark, and n is the sample size [17, 16]. Below we discuss additional details about these evaluations. We provide the prompt templates for all benchmarks at:

`https://huggingface.co/datasets/amazon-agi/Amazon-Nova-1.0-Premier-evals`

Text-Only Evaluations: We evaluate select core capabilities of Nova Premier on a variety of public text-only benchmarks, spanning general knowledge, reasoning, math, language understanding, coding, and instruction following. For general knowledge, we evaluate the model on MMLU [10], with 0-shot Chain-of-Thought (CoT) [26] prompting and report the macro average of exact match accuracy across all subjects. For reasoning, we evaluate Nova Premier's performance on GPQA Diamond [21], AIME 2025 [19], and MATH-500 [11]. For GPQA Diamond and MATH-500 we use 0-shot CoT for prompting and report the exact match accuracy. For AIME 2025, for each problem we generate 5 responses through sampling and report the average of the exact match accuracy. For coding, we evaluate the model on MBXP (5 languages[2]) [6] and BigCodeBench Hard [33], and we report Pass@1 where correctness of generated code is verified against test cases. For instruction following, we measure the model's performance on IFEval [32] and report the instruction-level accuracy under loose constraints.

[1] Results measured internally by Amazon for evaluation purposes after Nova Premier completed training using (i) the Bedrock API for Claude or (ii) the OpenAI API, as applicable.

[2] Python, Javascript, Typescript, Ruby, and PhP

Modality	Capability	Benchmarks	Nova Premier	Nova Pro	Claude-3.5 Sonnet-v2	Claude-3.7 Sonnet (No thinking)	GPT-4.5
Text	Knowledge	MMLU	87.4	85.9	88.3	90.1*	91.3*
	Reasoning	GPQA Diamond	57.1 ±6.9	50.0 ±7.0	65.0 ±6.6	68.0 ±6.5	71.4 ±6.3
		AIME 2025	16.0	5.3	4.7*	18.0*	30.0*
		MATH-500	82.0 ±3.4	76.6 ±3.7	78.0 ±3.6	82.2 ±3.4	88.0* ±2.8
	Coding	BigCodeBench Hard	28.1	22.3	30.4	32.8	33.1*
		MBXP (5 Languages)	78.4	65.9	80.0*	79.6*	83.3*
	Instruction Following	IFEval	91.5 ±2.4	92.1 ±2.3	90.2 ±2.5	90.8 ±2.4	93.3* ±2.1
Multimodal	Visual Reasoning	MMMU	68.0	62.0	70.4	71.8	74.4
	Document Understanding	OCRBench-v2	56.9	53.7	46.2*	47.9*	53.4*
	Chart Understanding	CharXiv (Descriptive/Reasoning)	84.6/48.8	70.5/40.6	84.3/60.2	88.7*/64.2	90.0/55.4
	Long-Form Video Understanding	EgoSchema	73.8 ±3.9	72.1 ±3.9	–	–	68.8* ±4.1
	Visual Counting	TallyQA	61.5 ±9.1	54.0 ±9.4	40.4* ±9.2	53.2* ±9.4	52.3* ±9.4
Agents	Retrieval-Augmented Generation	SimpleQA (with SerpApi)	86.3 ±1.0	84.6 ±1.1	87.4* ±1.0	91.1* ±0.9	89.1* ±0.9
	Function Calling	BFCL (2025-04-25)	63.7	60.8	56.9	58.6	70.3
	Web Agent	ScreenSpot Web Text	91.7 ±3.8	88.3 ±4.4	–	90.0* ±4.1	14.0* ±4.7
		ScreenSpot Web Icon	84.0 ±4.7	76.7 ±5.5	–	85.4* ±4.6	23.5* ±5.5
	Agentic Coding	SWE-bench Verified**	42.4	–	49.0	62.3	38.0

Table 1: Performance of Nova Premier compared with Nova Pro, as well as Claude-3.5 Sonnet-v2, Claude-3.7 Sonnet, and GPT-4.5. Evaluations are done on text, multimodal, and agentic capabilities for a diverse set of capabilities across different benchmarks. For benchmarks where the result is a simple average of binary scores, confidence intervals are provided. Results marked with * were measured by us for Claude and GPT-4.5 models. ** For SWE-bench, evaluation of Nova Premier is done using an internal agentic scaffold.

Multimodal Evaluations: We evaluate the multimodal capabilities of Nova Premier across knowledge, reasoning, document understanding, chart understanding, and object counting for image inputs. We also evaluate Nova Premier's capabilities in video understanding. For all benchmarks, we follow the suggested metrics and recommended data split for evaluation. To evaluate Nova Premier's general knowledge and reasoning capabilities in multimodal settings, we use MMMU [29] with CoT prompting, and we report exact match accuracy. For general document understanding, we evaluate Nova Premier on OCRBench-v2 that covers understanding of documents, charts, tables, mathematical formula, etc. We use the official prompt of the benchmark for evaluation and we report accuracy across all tasks. For chart understanding, we use CharXiv [24] as the benchmark without CoT prompt, and we report accuracy based on correctness binary score that is assigned by GPT-4o. For object counting in images, we use TallyQA [1] (balanced setting) without CoT prompting and we report exact match accuracy. We also report on EgoSchema [18] for ego-centric video understanding and we report exact match accuracy.

Agentic Evaluations: We evaluate Nova Premier's performance on agentic capabilities for RAG, function calling, web agents, and agentic coding. We evaluate agentic RAG capabilities on the SimpleQA [25] benchmark. We compute accuracy (correctness) using the prompt shared in the SimpleQA paper, and GPT-4o-2024-11-20 as a judge. The model dynamically generates the instructions to call an external web search tool, as predetermined by the user through the prompt. To evaluate function calling capabilities of the model, we use the Berkeley Function Calling Leaderboard (BFCL) V3 [28] that evaluates a model's ability to use a variety of tools in multi-turn settings; and we report overall accuracy. For web agents, we use ScreenSpot [8] as a benchmark for general GUI grounding. For multimodal agentic tasks, we specifically focus on web environments and report click accuracy results on ScreenSpot Web text and Web Icon. To evaluate the model's agentic coding capabilities, we use SWE-bench [13]. Specifically, we use a simple internal agentic scaffold on a 500 instance subset of SWE-bench known as SWE-bench Verified, and we report resolved rate.

2.2 Runtime Performance

We evaluate the runtime performance of Nova Premier using two metrics: Time to First Token (TTFT) and Output Tokens per Second (OTPS). TTFT is measured as the time (in seconds) it takes to receive the first token from the model after an API request is sent. OTPS is measured as the number of tokens generated per second, and is the rate at which a model produces subsequent output tokens after the first token. OTPS reflects overall throughput and efficiency during inference.

In Table 2, we show TTFT and OTPS using 1000 tokens of input and 100 tokens of output. We compare runtime performance of Nova Premier with Nova Pro, as well as select public models in its intelligence tier. All the numbers are reported by Artificial Analysis[3]. Nova Premier is the fastest models in its intelligence tier and demonstrates state-of-the-art runtime performance, ensuring a smooth and responsive user experience in many real world use cases.

	Metrics	Nova Premier	Nova Pro	Claude-3.5 Sonnet-v2	Claude-3.7 Sonnet	GPT-4.5
Latency	Time to First Answer Token (TTFT) (↓)	0.9	0.4	1.0	0.9	1.0
	Output Tokens per Second (OTPS) (↑)	63	150	45	42	54

Table 2: Runtime performance of Nova Premier compared to Nova Pro, as well as other model in its intelligence tier. (↑) indicates that higher numbers are better and (↓) indicates that lower numbers are better. We used Bedrock latency figures for models available on Bedrock.

3 Responsible AI

3.1 Responsible AI development

Our foundational approach to Responsible AI (RAI) for Nova Premier is similar to the one we followed for Nova Pro, Nova Lite, and Nova Micro [5], structured around eight dimensions [3]. We continued to edit and update the details of implementing these design objectives based on feedback from subject matter experts and applied them to the development of Nova Premier. We infused the requirements defined by our RAI objectives throughout various stages of the model development process: data collection, pre-training, post-training, and deployment of runtime mitigations.

In addition to the eight RAI dimensions, we published Amazon's Frontier Model Safety Framework [4] at the Paris AI Action Summit.[4] The framework defines critical thresholds for three risk domains: (i) Chemical, Biological, Radiological, and Nuclear (CBRN) Weapons Proliferation, (ii) Offensive Cyber Operations, and (iii) Automated AI R&D. In addition, the framework presents an evaluation and a risk mitigation strategy. We ensure alignment of Nova Premier with the Frontier Model Safety Framework.

[3] An independent entity that benchmarks AI models and hosting providers. Their benchmarking methodology is outlined at https://artificialanalysis.ai/methodology/performance-benchmarking

[4] https://www.elysee.fr/en/sommet-pour-l-action-sur-l-ia

3.2 Responsible AI Evaluations

This section details the Responsible AI evaluations performed to assess the safety of Nova Premier. It includes two sets of evaluations: (1) standard RAI evaluations, which assess Nova Premier against Amazon RAI objectives using both publicly available and proprietary benchmarks; and (2) Frontier Safety Framework evaluations.

3.2.1 Evaluations on Amazon RAI Objectives

We assess Nova Premier using a two-pronged evaluation strategy that aligns with Amazon's internal RAI objectives and broader public safety commitments. These evaluations span both internal objectives-driven assessments and external benchmark-driven evaluations.

Benchmark Evaluations: Nova Premier was evaluated against internal RAI benchmarks designed to enforce Amazon's RAI objectives across dimensions such as safety, privacy, veracity, and transparency. These evaluations ensure consistent rejection of unsafe or ambiguous prompts, with assessments performed through automated pipelines and reviews by policy experts. To complement internal assessments, Nova Premier was also tested on public benchmarks such as BOLD [9], WILDCHAT non-toxic [31], and StrongReject [22] to ensure consistent strength in performance during each stage of the model build.

Red Teaming: We continue with our three part red teaming strategy (outlined in [5]) consisting of Amazon internal red teaming, automated red teaming and third-party red teaming. For automated red teaming, we enhanced the FLIRT framework [20] to conduct multi-lingual and multimodal automated red teaming against Nova Premier. The findings were used to improve the model for the launch. Additionally, we continued red teaming both internally and with external partners and leveraged the feedback during the model development to improve the RAI adherence of Nova Premier. We leverage red teaming firms, including ActiveFence and Innodata, to conduct testing in areas such as hate speech, political misinformation, extremism, and other RAI dimensions. We also continue collaborations with specialized third-parties to red team our models for CBRN capabilities. Specifically, with the Gomes Group, we expanded upon the framework published in the Nova Technical Report [5] by investigating chemical synonym vulnerabilities. With Nemesys, we explored threats posed to nuclear facilities through different scenarios (such as insider sabotage to release radiation, external cyberattack to release radiation or shut down the facility, and physical assault to steal plutonium). Finally, with Deloitte, we tested an LLM's scientific knowledge and reasoning capabilities that could facilitate the development or use of biological weapons.

3.2.2 Frontier Safety Framework Evaluations

The assessment strategy in the Frontier Safety Framework involves evaluation through automated benchmarks, red teaming, and uplift studies. The objective is to determine whether Nova Premier breaches any critical capability thresholds within specified risk domains that could pose severe public safety risks. We evaluate Nova Premier across each risk domain and conduct further evaluations with third-party assessors (e.g., Nemesys Insights[5], METR[6]). In the sections below, we provide the evaluations for each risk domain.

Chemical, Biological, Radiological, and Nuclear (CBRN) Weapons Proliferation: To assess whether Nova Premier poses risks of contributing to the proliferation of CBRN weapons, we conducted a suite of evaluations aligned with the Frontier Model Safety Framework. These included automated testing using benchmarks such as the Weapons of Mass Destruction Proxy (WMDP) benchmark [15], ProtocolQA[14], and BioLP Bench[12] – each targeting different dimensions of biosafety and laboratory reasoning. In addition, structured red teaming protocols and uplift studies were conducted in collaboration with external assessors (Nemesys Insights) to formalize capability thresholds and assessment metrics. These assessments concluded that Nova Premier remains below the critical threshold for CBRN weapons proliferation, indicating that its deployment does not pose an elevated security risk.

Offensive Cyber Operations: To evaluate risks associated with offensive cybersecurity capabilities, Nova Premier was tested using a comprehensive suite of benchmarks that span both knowledge acquisition and practical execution of cyber attacks. This included SECURE [7], CTIBench [2], and CyberMetric [23] for evaluating foundational cybersecurity knowledge, as well as Cybench [30] for task-based assessments in simulated environments. In parallel, experts from an internal Amazon security team conducted targeted red teaming exercises and uplift studies to identify potential emergent risks. Their assessments concluded that Nova Premier remains below the critical threshold for offensive cyber operations, indicating that its deployment does not pose an elevated security risk.

[5] https://www.nemesysinsights.com
[6] https://metr.org

Automated AI Research and Development (AI R&D): To determine whether Nova Premier could autonomously accelerate AI development in unsafe ways, we conducted comprehensive evaluations using RE-Bench [27] and internal AI research automation simulation tests. These benchmarks test whether the model can plan, replicate, and execute novel AI research independently. Results from these evaluations were reviewed in collaboration with a third-party assessor (METR) to determine alignment with predefined safety criteria. The assessments concluded that Nova Premier does not meet the critical threshold for automated AI R&D and is therefore suitable for public deployment under current safety constraints.

References

[1] M. Acharya, K. Kafle, and C. Kanan. Tallyqa: Answering complex counting questions. In *AAAI*, 2019.

[2] M. T. Alam, D. Bhusal, L. Nguyen, and N. Rastogi. Ctibench: A benchmark for evaluating llms in cyber threat intelligence. *arXiv preprint arXiv:2406.07599*, 2024.

[3] Amazon. Building AI responsibly at AWS. https://aws.amazon.com/ai/responsible-ai/, 2024. Accessed: 2024-11-20.

[4] Amazon. Amazon's frontier model safety framework. *Amazon Technical Reports*, 2025. URL https://www.amazon.science/publications/amazons-frontier-model-safety-framework.

[5] Amazon AGI. The amazon nova family of models: Technical report and model card. *Amazon Technical Reports*, 2024. URL https://www.amazon.science/publications/the-amazon-nova-family-of-models-technical-report-and-model-card.

[6] B. Athiwaratkun, S. K. Gouda, Z. Wang, X. Li, Y. Tian, M. Tan, W. U. Ahmad, S. Wang, Q. Sun, M. Shang, S. K. Gonugondla, H. Ding, V. Kumar, N. Fulton, A. Farahani, S. Jain, R. Giaquinto, H. Qian, M. K. Ramanathan, R. Nallapati, B. Ray, P. Bhatia, S. Sengupta, D. Roth, and B. Xiang. Multi-lingual evaluation of code generation models, 2023. URL https://arxiv.org/abs/2210.14868.

[7] D. Bhusal, M. Tanvirul Alam, L. Nguyen, A. Mahara, Z. Lightcap, R. Frazier, R. Fieblinger, G. L. Torales, and N. Rastogi. Secure: benchmarking generative large language models for cybersecurity advisory. *arXiv e-prints*, pages arXiv–2405, 2024.

[8] K. Cheng, Q. Sun, Y. Chu, F. Xu, Y. Li, J. Zhang, and Z. Wu. Seeclick: Harnessing gui grounding for advanced visual gui agents, 2024.

[9] J. Dhamala, T. Sun, V. Kumar, S. Krishna, Y. Pruksachatkun, K.-W. Chang, and R. Gupta. BOLD: Dataset and metrics for measuring biases in open-ended language generation. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 862–872, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383097. doi: 10.1145/3442188.3445924. URL https://doi.org/10.1145/3442188.3445924.

[10] D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, and J. Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

[11] D. Hendrycks, C. Burns, S. Kadavath, A. Arora, S. Basart, E. Tang, D. Song, and J. Steinhardt. Measuring mathematical problem solving with the MATH dataset. *NeurIPS*, 2021.

[12] I. Ivanov. Biolp-bench: Measuring understanding of biological lab protocols by large language models. *bioRxiv*, pages 2024–08, 2024.

[13] C. E. Jimenez, J. Yang, A. Wettig, S. Yao, K. Pei, O. Press, and K. Narasimhan. Swe-bench: Can language models resolve real-world github issues?, 2024. URL https://arxiv.org/abs/2310.06770.

[14] J. M. Laurent, J. D. Janizek, M. Ruzo, M. M. Hinks, M. J. Hammerling, S. Narayanan, M. Ponnapati, A. D. White, and S. G. Rodriques. Lab-bench: Measuring capabilities of language models for biology research. *arXiv preprint arXiv:2407.10362*, 2024.

[15] N. Li, A. Pan, A. Gopal, S. Yue, D. Berrios, A. Gatti, J. D. Li, A.-K. Dombrowski, S. Goel, L. Phan, et al. The wmdp benchmark: Measuring and reducing malicious use with unlearning. *arXiv preprint arXiv:2403.03218*, 2024.

[16] Llama Team, AI Meta. The Llama 3 herd of models, 2024. URL https://github.com/meta-llama/llama-models/blob/main/models/llama3_1/MODEL_CARD.md.

[17] L. Madaan, A. K. Singh, R. Schaeffer, A. Poulton, S. Koyejo, P. Stenetorp, S. Narang, and D. Hupkes. Quantifying variance in evaluation benchmarks, 2024. URL https://arxiv.org/abs/2406.10229.

[18] K. Mangalam, R. Akshulakov, and J. Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding, 2023. URL https://arxiv.org/abs/2308.09126.

[19] Mathematical Association of America. American invitational mathematics examination (aime) 2025, 2025. URL https://maa.org/math-competitions/aime. Administered by the Mathematical Association of America.

[20] N. Mehrabi, P. Goyal, C. Dupuy, Q. Hu, S. Ghosh, R. Zemel, K.-W. Chang, A. Galstyan, and R. Gupta. Flirt: Feedback loop in-context red teaming. *arXiv preprint arXiv:2308.04265*, 2023.

[21] D. Rein, B. L. Hou, A. C. Stickland, J. Petty, R. Y. Pang, J. Dirani, J. Michael, and S. R. Bowman. GPQA: A graduate-level google-proof Q&A benchmark, 2023. URL https://arxiv.org/abs/2311.12022.

[22] A. Souly, Q. Lu, D. Bowen, T. Trinh, E. Hsieh, S. Pandey, P. Abbeel, J. Svegliato, S. Emmons, O. Watkins, et al. A strongreject for empty jailbreaks. *arXiv preprint arXiv:2402.10260*, 2024.

[23] N. Tihanyi, M. A. Ferrag, R. Jain, T. Bisztray, and M. Debbah. Cybermetric: a benchmark dataset based on retrieval-augmented generation for evaluating llms in cybersecurity knowledge. In *2024 IEEE International Conference on Cyber Security and Resilience (CSR)*, pages 296–302. IEEE, 2024.

[24] Z. Wang, M. Xia, L. He, H. Chen, Y. Liu, R. Zhu, K. Liang, X. Wu, H. Liu, S. Malladi, A. Chevalier, S. Arora, and D. Chen. Charxiv: Charting gaps in realistic chart understanding in multimodal llms. *arXiv preprint arXiv:2406.18521*, 2024.

[25] J. Wei, N. Karina, H. W. Chung, Y. J. Jiao, S. Papay, A. Glaese, J. Schulman, and W. Fedus. Measuring short-form factuality in large language models, 2024. URL https://arxiv.org/abs/2411.04368.

[26] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. H. Chi, Q. V. Le, and D. Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2024. Curran Associates Inc. ISBN 9781713871088.

[27] H. Wijk, T. Lin, J. Becker, S. Jawhar, N. Parikh, T. Broadley, L. Chan, M. Chen, J. Clymer, J. Dhyani, et al. Re-bench: Evaluating frontier ai r&d capabilities of language model agents against human experts. *arXiv preprint arXiv:2411.15114*, 2024.

[28] F. Yan, H. Mao, C. C.-J. Ji, T. Zhang, S. G. Patil, I. Stoica, and J. E. Gonzalez. Berkeley function calling leaderboard. https://gorilla.cs.berkeley.edu/blogs/8_berkeley_function_calling_leaderboard.html, 2024.

[29] X. Yue, Y. Ni, K. Zhang, T. Zheng, R. Liu, G. Zhang, S. Stevens, D. Jiang, W. Ren, Y. Sun, C. Wei, B. Yu, R. Yuan, R. Sun, M. Yin, B. Zheng, Z. Yang, Y. Liu, W. Huang, H. Sun, Y. Su, and W. Chen. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of CVPR*, 2024.

[30] A. K. Zhang, N. Perry, R. Dulepet, J. Ji, C. Menders, J. W. Lin, E. Jones, G. Hussein, S. Liu, D. Jasper, et al. Cybench: A framework for evaluating cybersecurity capabilities and risks of language models. *arXiv preprint arXiv:2408.08926*, 2024.

[31] W. Zhao, X. Ren, J. Hessel, C. Cardie, Y. Choi, and Y. Deng. Wildchat: 1m chatgpt interaction logs in the wild. *arXiv preprint arXiv:2405.01470*, 2024.

[32] J. Zhou, T. Lu, S. Mishra, S. Brahma, S. Basu, Y. Luan, D. Zhou, and L. Hou. Instruction-following evaluation for large language models, 2023. URL https://arxiv.org/abs/2311.07911.

[33] T. Y. Zhuo, V. M. Chien, J. Chim, H. Hu, W. Yu, R. Widyasari, I. N. B. Yusuf, H. Zhan, J. He, I. Paul, S. Brunner, C. GONG, J. Hoang, A. R. Zebaze, X. Hong, W.-D. Li, J. Kaddour, M. Xu, Z. Zhang, P. Yadav, N. Jain, A. Gu, Z. Cheng, J. Liu, Q. Liu, Z. Wang, D. Lo, B. Hui, N. Muennighoff, D. Fried, X. Du, H. de Vries, and L. V. Werra. Bigcodebench: Benchmarking code generation with diverse function calls and complex instructions. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=YrycTjllL0.