Claude 3.5 Sonnet Model Card Addendum

Anthropic

1 Introduction

This addendum to our Claude 3 Model Card describes Claude 3.5 Sonnet, a new model which outperforms our previous most capable model, Claude 3 Opus, while operating faster and at a lower cost. Claude 3.5 Sonnet offers improved capabilities, including better coding and visual processing. Since it is an evolution of the Claude 3 model family, we are providing an addendum rather than a new model card. We provide updated key evaluations and results from our safety testing.

2 Evaluations

Reasoning, Coding, and Question Answering

We evaluated Claude 3.5 Sonnet on a series of industry-standard benchmarks covering reasoning, reading comprehension, math, science, and coding. Across all these benchmarks, Claude 3.5 Sonnet outperforms our previous frontier model, Claude 3 Opus. It also sets new performance standards in evaluations of graduate level science knowledge (GPQA) [1], general reasoning (MMLU) [2], and coding proficiency (HumanEval) [3]. The results are shown in Table 1.

Vision Capabilities

Claude 3.5 Sonnet also outperforms prior Claude 3 models on five standard vision benchmarks, delivering state-of-the-art performance on evaluations of visual math reasoning (MathVista) [4], question answering about charts and graphs (ChartQA) [5], document understanding (DocVQA) [6], and question answering about science diagrams (AI2D) [7]. The results are shown in Table 2.

Agentic Coding

Claude 3.5 Sonnet solves 64% of problems on an internal agentic coding evaluation, compared to 38% for Claude 3 Opus. Our evaluation tests a model's ability to understand an open source codebase and implement a pull request, such as a bug fix or new feature, given a natural language description of the desired improvement. For each problem, the model is evaluated based on whether all the tests of the codebase pass for the completed code submission. The tests are not visible to the model, and include tests of the bug fix or new feature. To ensure the evaluation mimics real world software engineering, we based the problems on real pull requests submitted to open source codebases. The changes involve searching, viewing, and editing multiple files (typically three or four, as many as twenty). The model is allowed to write and run code in an agentic loop and iteratively self-correct during evaluation. We run these tests in a secure sandboxed environment without access to the internet. The results are shown in Table 3.

		Claude 3.5 Sonnet	Claude 3 Opus	Claude 3 Sonnet	GPT-4o [8]	GPT-4T [8]	Gemini 1.5 Pro [9]	Llama 3 400B[1]
GPQA (Diamond) Graduate level Q&A	0-shot CoT	**59.4%**	50.4%	40.4%	53.6%	48.0%	—	—
	Maj@32 5-shot CoT	**67.2%**	59.5%	46.3%	—	—	—	—
MMLU General reasoning	5-shot CoT	**90.4%**	88.2%	81.5%	—	—	—	—
	5-shot	**88.7%**	86.8%	78.3%	—	—	85.9%	86.1%
	0-shot CoT[2]	88.3%	85.7%	77.1%	**88.7%**	86.5%	—	—
MATH [10] Mathematical problem solving		71.1% 0-shot CoT	60.1% 0-shot CoT	43.1% 0-shot CoT	**76.6%** 0-shot CoT	72.6% 0-shot CoT	67.7% 4-shot	57.8% 4-shot CoT
HumanEval Python coding tasks	0-shot	**92.0%**	84.9%	73.0%	90.2%	87.1%	84.1%	84.1%
MGSM [11] Multilingual math		**91.6%** 0-shot CoT	90.7% 0-shot CoT	83.5% 0-shot CoT	90.5% 0-shot CoT	88.5% 0-shot CoT	87.5% 8-shot	—
DROP [12] Reading comprehension, arithmetic	F1 Score	**87.1** 3-shot	83.1 3-shot	78.9 3-shot	83.4 3-shot	86.0 3-shot	74.9 Variable shots	83.5 3-shot
BIG-Bench Hard [13, 14] Mixed evaluations	3-shot CoT	**93.1%**	86.8%	82.9%	—	—	89.2%	85.3%
GSM8K [15] Grade school math		**96.4%** 0-shot CoT	95.0% 0-shot CoT	92.3% 0-shot CoT	—	—	90.8% 11-shot	94.1% 8-shot CoT

Table 1 This table shows evaluation results for reasoning, math, coding, reading comprehension, and question answering evaluations.

	Claude 3.5 Sonnet	Claude 3 Opus	Claude 3 Sonnet	GPT-4o [8]	GPT-4T [8]	Gemini 1.5 Pro [9]
MMMU [18] (validation) Visual question answering	68.3%	59.4%	53.1%	**69.1%**	63.1%	62.2%
MathVista (testmini) Math	**67.7%**	50.5%	47.9%	63.8%	58.1%	63.9%
AI2D (test) Science diagrams	**94.7%**	88.1%	88.7%	94.2%	89.4%	94.4%
ChartQA (test, relaxed accuracy) Chart understanding	**90.8%**	80.8%	81.1%	85.7%	78.1%	87.2%
DocVQA (test, ANLS score) Document understanding	**95.2%**	89.3%	89.5%	92.8%	87.2%	93.1%

Table 2 This table shows evaluation results for multimodal tasks. All of these evaluations are 0-shot. On MMMU, MathVista, and ChartQA, all models use chain-of-thought reasoning before providing a final answer.

[1] Results for Llama 3 400B are from the April 15, 2024 checkpoint [16]. All metrics are for the Instruct version of the model, other than DROP and BIG-Bench Hard, which are evaluated on the pretrained version of the model.

[2] OpenAI's simple-evals suite [17] lists MMLU scores of 88.7% for gpt-4o and 86.5% for gpt-4-turbo-2024-04-09. The simple-evals MMLU implementation uses 0-shot CoT.

	Claude 3.5 Sonnet	Claude 3 Opus	Claude 3 Sonnet	Claude 3 Haiku
% of problems which pass all tests	64%	38%	21%	17%

Table 3 This table shows the results of our internal agentic coding evaluation. For each model, we indicate the percent of problems for which the model's final solution passed all the tests.

Refusals

We assessed Claude 3.5 Sonnet's ability to differentiate between harmful and benign requests. These tests, which use the Wildchat [19] and XSTest [20] datasets, are designed to measure the model's ability to avoid unnecessary refusals with harmless prompts while maintaining appropriate caution with harmful content. Claude 3.5 Sonnet outperformed Opus on both dimensions: It has fewer incorrect refusals and more correct refusals. The results are shown in Table 4.

	Claude 3.5 Sonnet	Claude 3 Opus	Claude 3 Sonnet	Claude 3 Haiku
Wildchat Toxic Correct refusals (higher ↑ is better)	**96.4%**	92.0%	93.6%	90.0%
Wildchat Non-toxic Incorrect refusals (lower ↓ is better)	11.0%	11.9%	8.8%	**6.6%**
XSTest Incorrect refusals (lower ↓ is better)	**1.7%**	8.3%	36.6%	33.1%

Table 4 This table shows refusal rates for the toxic prompts in the Wildchat dataset, and incorrect refusal rates for the non-toxic prompts in the Wildchat and XSTest datasets.

Needle In A Haystack

We evaluated Claude 3.5 Sonnet on our version [21] of the "Needle In a Haystack" task [22] to confirm its retrieval abilities at context lengths up to 200k tokens. The average recall outperformed Claude 3 Opus. The results are shown in Table 5, Figure 1, and Figure 2.

	Claude 3.5 Sonnet	Claude 3 Opus	Claude 3 Sonnet	Claude 3 Haiku	Claude 2.1
All context lengths	99.7%	99.4%	95.4%	95.9%	94.5%
200k context length	99.7%	98.3%	91.4%	91.9%	92.7%

Table 5 Comparison of average recall achieved by our models on the Needle In A Haystack evaluation.

2.1 Human Feedback Evaluations

We evaluated Claude 3.5 Sonnet via direct comparison to prior Claude models. We asked raters to chat with our models and evaluate them on a number of tasks, using task-specific instructions. The charts in Figure 3 show the "win rate" when compared to a baseline of Claude 3 Opus.[3]

We saw large improvements in core capabilities like coding, documents, creative writing, and vision. Domain experts preferred Claude 3.5 Sonnet over Claude 3 Opus, with win rates as high as 82% in Law, 73% in Finance, and 73% in Philosophy.

[3] Section 5.5 of the main model card explains how win rates are calculated.

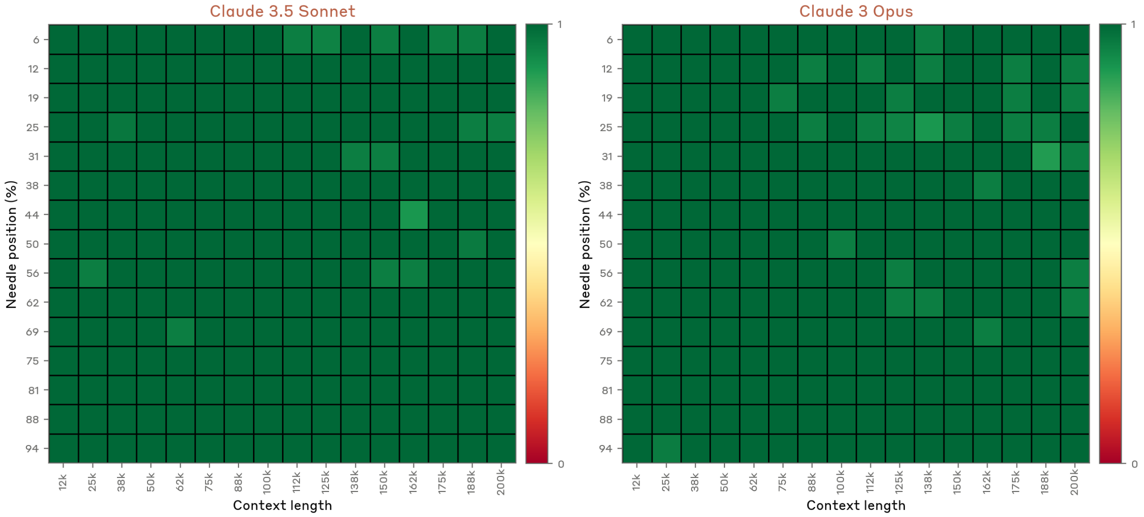

Figure 1 This plot shows recall on the Needle In A Haystack evaluation with a modified prompt. Like Claude 3 Opus, Claude 3.5 Sonnet achieves near perfect recall.

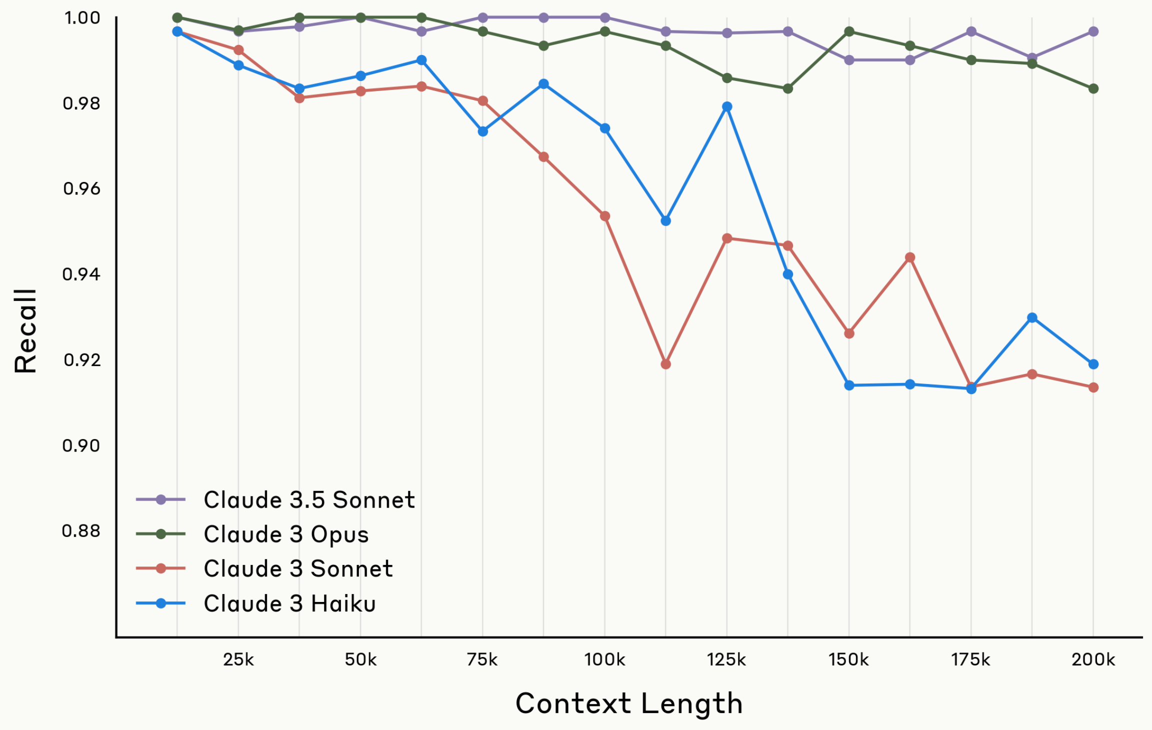

Figure 2 This plot shows the average recall achieved by our models as context length grows in the Needle In A Haystack evaluation.

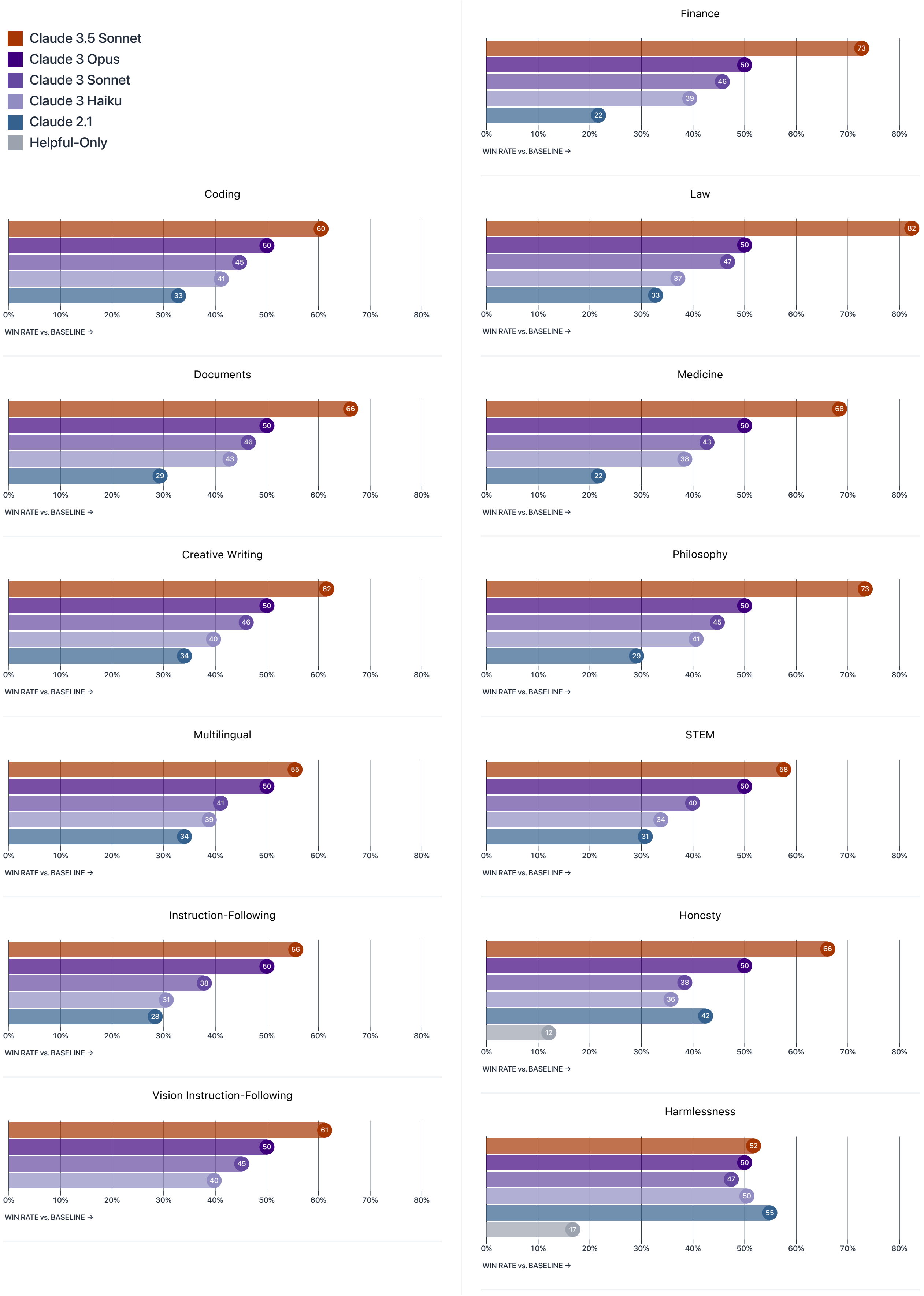

Figure 3 These plots show per-task human preference win rates for: common use cases (left), expert knowledge (top right), and adversarial scenarios (bottom right). Since Claude 3 Opus is the baseline model, it always has a 50% win rate. (It beats itself 50% of the time.)

3 Safety

3.1 Introduction

This section discusses our safety evaluations and commitments, and how we applied them to Claude 3.5 Sonnet.

Anthropic has conducted evaluations of Claude 3.5 Sonnet as part of our ongoing commitment to responsible AI development and deployment. While Claude 3.5 Sonnet represents an improvement in capabilities over our previously released Opus model, it does not trigger the 4x effective compute threshold at which we will run the full evaluation protocol described in our Responsible Scaling Policy (RSP). We previously ran this evaluation protocol on Claude 3 Opus, which you can read about here [23].

Nevertheless, we believe it is important to conduct regular safety testing prior to releasing frontier models, even if not formally required by our RSP. We still have a lot to learn about the science of evaluations and ideal cadences for performing these tests, and regular testing allows us to refine our evaluation methodologies for future, more capable models. This perspective is reflected in our voluntary White House [24], G7 [25], and Seoul Summit Frontier Safety [26] commitments to do targeted testing prior to any major model release.

Our safety teams performed a range of evaluations on Claude 3.5 Sonnet in the areas of Chemical, Biological, Radiological, and Nuclear (CBRN) risks, cybersecurity, and autonomous capabilities. Based on our assessments, we classify Claude 3.5 Sonnet as an AI Safety Level 2 (ASL-2) model, indicating that it does not pose risk of catastrophic harm. Additionally, we worked with external third party evaluation partners such as the UK Artificial Intelligence Safety Institute (UK AISI) to independently assess Claude 3.5 Sonnet.

3.2 Safety Evaluations Overview

We conducted frontier risk evaluations focusing on CBRN, cyber, and autonomous capabilities risks. As part of our efforts to continuously improve our safety testing, we improved on the approach we used for Claude 3 Opus by refining our threat models and designing new and better evaluations for this round of testing. The UK AISI also conducted pre-deployment testing of a near-final model, and shared their results with the US AI Safety Institute as part of a Memorandum of Understanding, made possible by the partnership between the US and UK AISIs announced earlier this year [27]. Additionally, METR [28] did an initial exploration of the model's autonomy-relevant capabilities.

For CBRN, we conducted automated tests of CBRN knowledge and measured the model's ability to improve non-expert performance on CBRN-related tasks. Cybersecurity was assessed through capture-the-flag challenges testing vulnerability discovery and exploit development. Autonomous capabilities were evaluated based on the model's ability to write software engineer-quality code that passes pre-defined code tests, similar to the internal agentic coding evaluation discussed in Section 2. For each evaluation domain, we defined quantitative "thresholds of concern" that, if passed, would be a conservative indication of proximity to our ASL-3 threshold of concern. If the model had exceeded our preset thresholds during testing, we planned to convene a council consisting of our Responsible Scaling Officer, evaluations leads, and external subject matter experts to determine whether the model's capabilities were close enough to a threshold of concern to warrant either more intensive evaluations or an increase in safety and security protections.

Evaluating a Helpful, Honest, and Harmless (HHH)-trained model poses some challenges, because safety guardrails can cause capability evaluations to underestimate a model's underlying capabilities due to refusals caused by the HHH training. Because our goal was to evaluate for capabilities, we accounted for model refusals in several ways. First, we measured the degree of refusals across topics, and found that Claude 3.5 Sonnet refused ASL-3 harmful queries adequately to substantially reduce its usefulness for clearly harmful queries. Second, we employed internal research techniques to acquire non-refusal model responses in order to estimate how the model would have performed if it were trained as a Helpful-only model, instead of as an HHH model. Note that it is possible that the results underestimate the capability of an equivalent Helpful-only model, because alignment training may cause it to hedge concerning answers.

3.3 Safety Evaluations Results

We observed an increase in capabilities in risk-relevant areas compared to Claude 3 Opus. The increases in performance we observed in Claude 3.5 Sonnet in risk-related domains, relative to our prior models, are consistent with the incremental training and elicitation techniques applied to this model. Claude 3.5 Sonnet did not exceed our safety thresholds in these evaluations and we classify it at ASL-2.

References

[1] 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 QA Benchmark," *arXiv preprint arXiv:2311.12022* (2023) .

[2] 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.

[3] M. Chen, J. Tworek, H. Jun, Q. Yuan, H. P. d. O. Pinto, J. Kaplan, H. Edwards, Y. Burda, N. Joseph, G. Brockman, *et al.*, "Evaluating Large Language Models Trained on Code," *arXiv preprint arXiv:2107.03374* (July, 2021) .

[4] P. Lu, H. Bansal, T. Xia, J. Liu, C. Li, H. Hajishirzi, H. Cheng, K.-W. Chang, M. Galley, and J. Gao, "MathVista: Evaluating Mathematical Reasoning of Foundation Models in Visual Contexts." October, 2023.

[5] A. Masry, D. X. Long, J. Q. Tan, S. Joty, and E. Hoque, "ChartQA: A Benchmark for Question Answering about Charts with Visual and Logical Reasoning." 2022.

[6] M. Mathew, D. Karatzas, and C. V. Jawahar, "DocVQA: A Dataset for VQA on Document Images." January, 2021.

[7] A. Kembhavi, M. Salvato, E. Kolve, M. Seo, H. Hajishirzi, and A. Farhadi, "A Diagram is Worth a Dozen Images," *ArXiv* **abs/1603.07396** (2016) . https://api.semanticscholar.org/CorpusID:2682274.

[8] OpenAI, "Hello GPT-4o." June, 2024. https://openai.com/index/hello-gpt-4o/.

[9] Gemini Team, Google, "Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context." May, 2024. https://storage.googleapis.com/deepmind-media/gemini/gemini_v1_5_report.pdf.

[10] 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* (November, 2021) .

[11] F. Shi, M. Suzgun, M. Freitag, X. Wang, S. Srivats, S. Vosoughi, H. W. Chung, Y. Tay, S. Ruder, D. Zhou, *et al.*, "Language Models are Multilingual Chain-of-Thought Reasoners," in *International Conference on Learning Representations*. October, 2022.

[12] D. Dua, Y. Wang, P. Dasigi, G. Stanovsky, S. Singh, and M. Gardner, "DROP: A Reading Comprehension Benchmark Requiring Discrete Reasoning Over Paragraphs," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. April, 2019.

[13] A. Srivastava, A. Rastogi, A. Rao, A. A. M. Shoeb, A. Abid, A. Fisch, A. R. Brown, *et al.*, "Beyond the imitation game: Quantifying and extrapolating the capabilities of language models." June, 2023.

[14] M. Suzgun, N. Scales, N. Schärli, S. Gehrmann, Y. Tay, H. W. Chung, A. Chowdhery, Q. V. Le, E. H. Chi, D. Zhou, and J. Wei, "Challenging BIG-Bench Tasks and Whether Chain-of-Thought Can Solve Them." October, 2022.

[15] K. Cobbe, V. Kosaraju, M. Bavarian, M. Chen, H. Jun, L. Kaiser, M. Plappert, J. Tworek, J. Hilton, R. Nakano, *et al.*, "Training Verifiers to Solve Math Word Problems," *arXiv preprint arXiv:2110.14168* (November, 2021) .

[16] Meta, "Introducing Meta Llama 3: The most capable openly available LLM to date." April, 2024. https://ai.meta.com/blog/meta-llama-3/.

[17] OpenAI, "simple-evals." May, 2024. https://github.com/openai/simple-evals/.

[18] X. Yue, Y. Ni, K. Zhang, T. Zheng, R. Liu, G. Zhang, S. Stevens, *et al.*, "MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI." 2023.

[19] W. Zhao, X. Ren, J. Hessel, C. Cardie, Y. Choi, and Y. Deng, "(InThe)WildChat: 570K ChatGPT Interaction Logs In The Wild," in *International Conference on Learning Representations*. February, 2024.

[20] P. Röttger, H. R. Kirk, B. Vidgen, G. Attanasio, F. Bianchi, and D. Hovy, "XSTest: A Test Suite for Identifying Exaggerated Safety Behaviours in Large Language Models." 2023.

[21] Anthropic, "Long context prompting for Claude 2.1." December, 2023. https://www.anthropic.com/news/claude-2-1-prompting.

[22] G. Kamradt, "Pressure testing Claude-2.1 200K via Needle-in-a-Haystack." November, 2023.

[23] Anthropic, "Responsible Scaling Policy Evaluations Report – Claude 3 Opus." May, 2024. https://cdn.sanity.io/files/4zrzovbb/website/210523b8e11b09c704c5e185fd362fe9e648d457.pdf.

[24] "White House Voluntary AI Commitments." July, 2023. https://www.whitehouse.gov/wp-content/uploads/2023/09/Voluntary-AI-Commitments-September-2023.pdf.

[25] "Hiroshima Process International Code of Conduct for Organizations Developing Advanced AI Systems." October, 2023. https://www.mofa.go.jp/files/100573473.pdf.

[26] "Frontier AI Safety Commitments, AI Seoul Summit 2024." May, 2024. https://www.gov.uk/government/publications/frontier-ai-safety-commitments-ai-seoul-summit-2024/frontier-ai-safety-commitments-ai-seoul-summit-2024.

[27] U.S. Department of Commerce, "U.S. and U.K. Announce Partnership on Science of AI Safety." April, 2024. https://www.commerce.gov/news/press-releases/2024/04/us-and-uk-announce-partnership-science-ai-safety.

[28] https://metr.org/.