3 Security

We protect the security of the environment of our models to help ensure their integrity using a variety of connection authentication and authorization techniques; people are required to use multi-factor authentication at all times. Our advanced models are protected by two-party controls. Access to AI model infrastructure is granted explicitly per user and validated per access attempt. All accounts with access to the serving infrastructure hosting our services are protected via rigorous password requirements and multi-factor authentication. Each account is provisioned with the minimum privilege levels needed by its owner. Additional layers of defense include continuous systems' monitoring, 24/7 alert response, endpoint hardening, data storage and sharing controls, personnel vetting, and physical security hardening. We take significant care in testing any code changes prior to deployment to production environments including code review. Finally, we engage with penetration testers to exercise our detection systems and improve our defense posture.

4 Social Responsibility

As a PBC, Anthropic is committed to developing safe and responsible AI systems throughout each stage of the development process. Claude 3 models show a more nuanced understanding of requests, recognize real harm, and refuse to answer harmless prompts less often than prior models. That said, they can still make mistakes and our work to make Claude more helpful, harmless, and honest is ongoing. Ethical considerations also shape both our AUP, which delineates permissible and impermissible uses of Claude, and the Trust and Safety processes that enforce it.

4.1 Constitutional AI

Our core research focus has been training Claude models to be helpful, honest, and harmless. Currently, we do this by giving models a Constitution – a set of ethical and behavioral principles that the model uses to guide its outputs. The majority of the principles in Claude's constitution are the same as those we published in May 2023 [6]. Using this Constitution, models are trained to avoid sexist, racist, and toxic outputs, as well as to avoid helping a human engage in illegal or unethical activities. In response to our work on Collective Constitutional AI [17], we added an additional principle informed by our public input process, which instructs Claude to be understanding of and accessible to individuals with disabilities, resulting in lower model stereotype bias.

4.2 Labor

Anthropic works with several data work platforms which are responsible for engaging and managing data workers who work on Anthropic's projects.

Data work tasks include selecting preferred model outputs in order to train AI models to align with those preferences; evaluating model outputs according to a broad range of criteria (e.g., accuracy, helpfulness, harmlessness, etc.); and adversarially testing (i.e., red teaming) our models to identify potential safety vulnerabilities. This data work is primarily used in our technical safety research, and select aspects of it are also used in our model training.

4.3 Sustainability

We offset our emissions (including from our cloud computing usage) and work with cloud providers that prioritize renewable energy and carbon neutrality. Anthropic works to fully offset our operational carbon emissions each year, partnering with external experts to conduct a rigorous analysis of our company-wide carbon footprint. Once measured, we invest in verified carbon credits to fully offset our annual footprint. Our credits directly fund emissions reduction projects. Our goal is to maintain net zero climate impact on an annual basis through such initiatives and offsets.

5 Core Capabilities Evaluations

We conducted a comprehensive evaluation of the Claude 3 family to analyze trends in their capabilities across various domains. Our assessment included several broad categories:

- **Reasoning:** Benchmarks in this category require mathematical, scientific, and commonsense reasoning, testing the models' ability to draw logical conclusions and apply knowledge to real-world scenarios.
- **Multilingual:** This category comprises tasks for translation, summarization, and reasoning in multiple languages, evaluating the models' linguistic versatility and cross-lingual understanding.
- **Long Context:** These evaluations are focused on question answering and retrieval, assessing the models' performance in handling extended texts and extracting relevant information.
- **Honesty / Factuality:** Questions in this category assess the models' ability to provide accurate and reliable responses, either in terms of factual accuracy or fidelity to provided source materials. When unsure, the models are expected to be honest about their limitations, expressing uncertainty or admitting that they do not have sufficient information to provide a definitive answer.
- **Multimodal:** Evaluations include questions on science diagrams, visual question answering, and quantitative reasoning based on images.

These capabilities evaluations helped measure the models' skills, strengths, and weaknesses across a range of tasks. Many of these evaluations are industry standard, and we have invested in additional evaluation techniques and topics described below. We also present internal benchmarks we've developed over the course of training to address issues with harmless refusals.

5.1 Reasoning, Coding, and Question Answering

We evaluated the Claude 3 family on a series of industry-standard benchmarks covering reasoning, reading comprehension, math, science, and coding. The Claude 3 models demonstrate superior capabilities in these areas, surpassing previous Claude models, and in many cases achieving state-of-the-art results. These improvements are highlighted in our results presented in Table 1.

We tested our models on challenging domain-specific questions in GPQA [1], MMLU [2], ARC-Challenge [22], and PubMedQA [23]; math problem solving in both English (GSM8K, MATH) [24, 25] and multilingual settings (MGSM) [26]; common-sense reasoning in HellaSwag [27], WinoGrande [28]; reasoning over text in DROP [29]; reading comprehension in RACE-H [30] and QuALITY [31] (see Table 6); coding in HumanEval [32], APPS [33], and MBPP [34]; and a variety of tasks in BIG-Bench-Hard [35, 36].

GPQA (A Graduate-Level Google-Proof Q&A Benchmark) is of particular interest because it is a new evaluation released in November 2023 with difficult questions focused on graduate level expertise and reasoning. We focus mainly on the Diamond set as it was selected by identifying questions where domain experts agreed on the solution, but experts from other domains could not successfully answer the questions despite spending more than 30 minutes per problem, with full internet access. We found the GPQA evaluation to have very high variance when sampling with chain-of-thought at $T = 1$. In order to reliably evaluate scores on the Diamond set 0-shot CoT (50.4%) and 5-shot CoT (53.3%), we compute *the mean over 10 different evaluation rollouts*. In each rollout, we randomize the order of the multiple choice options. We see that Claude 3 Opus typically scores around 50% accuracy. This improves greatly on prior models but falls somewhat short of graduate-level domain experts, who achieve accuracy scores in the 60-80% range [1] on these questions.

We leverage majority voting [37] at test time to evaluate the performance by asking models to solve each problem using chain-of-thought reasoning (CoT) [38] N different times, sampling at $T = 1$, and then we report the answer that occurs most often. When we evaluate in this way in a few-shot setting Maj@32 Opus achieves a score of **73.7%** for MATH and **59.5%** for GPQA. For the latter, we averaged over 10 iterations of Maj@32 as even with this evaluation methodology, there was significant variance (with some rollouts scoring in the low 60s, and others in the mid-to-high 50s).

		Claude 3 Opus	Claude 3 Sonnet	Claude 3 Haiku	GPT-4[3]	GPT-3.5[3]	Gemini 1.0 Ultra[4]	Gemini 1.5 Pro[4]	Gemini 1.0 Pro[4]
MMLU General reasoning	5-shot	**86.8%**	79.0%	75.2%	86.4%	70.0%	83.7%	81.9%	71.8%
	5-shot CoT	**88.2%**	81.5%	76.7%	—	—	—	—	—
MATH[5] Mathematical problem solving	4-shot	**61%**	40.5%	40.9%	52.9% [6,7]	34.1%	53.2%	58.5%	32.6%
	0-shot	**60.1%**	43.1%	38.9%	42.5% (from [39])	—	—	—	—
	Maj@32 4-shot	**73.7%**	55.1%	50.3%	—	—	—	—	—
GSM8K Grade school math		**95.0%** 0-shot CoT	92.3% 0-shot CoT	88.9% 0-shot CoT	92.0% SFT, 5-shot CoT	57.1% 5-shot	94.4% Maj1@32	91.7% 11-shot	86.5% Maj1@32
HumanEval Python coding tasks	0-shot	**84.9%**	73.0%	75.9%	67.0%[6]	48.1%	74.4%	71.9%	67.7%
GPQA (Diamond) Graduate level Q&A	0-shot CoT	**50.4%**	40.4%	33.3%	35.7% (from [1])	28.1% (from [1])	—	—	—
	Maj@32 5-shot CoT	**59.5%**	46.3%	40.1%	—	—	—	—	—
MGSM Multilingual math		**90.7%** 0-shot	83.5% 0-shot	75.1% 0-shot	74.5%[7] 8-shot	—	79.0% 8-shot	88.7% 8-shot	63.5% 8-shot
DROP Reading comprehension, arithmetic	F1 Score	**83.1** 3-shot	78.9 3-shot	78.4 3-shot	80.9 3-shot	64.1 3-shot	82.4 Variable shots	78.9 Variable shots	74.1 Variable shots
BIG-Bench-Hard Mixed evaluations	3-shot CoT	**86.8%**	82.9%	73.7%	83.1%[7]	66.6%	83.6%	84.0%	75.0%
ARC-Challenge Common-sense reasoning	25-shot	**96.4%**	93.2%	89.2%	96.3%	85.2%	—	—	—
HellaSwag Common-sense reasoning	10-shot	**95.4%**	89.0%	85.9%	95.3%	85.5%	87.8%	92.5%	84.7%
PubMedQA[8] Biomedical questions	5-shot	75.8%	**78.3%**	76.0%	74.4%	60.2%	—	—	—
	0-shot	74.9%	**79.7%**	78.5%	75.2%	71.6%	—	—	—
WinoGrande Common-sense reasoning	5-shot	**88.5%**	75.1%	74.2%	87.5%	—	—	—	—
RACE-H Reading comprehension	5-shot	92.9%	88.8%	87.0%	—	—	—	—	—
APPS Python coding tasks	0-shot	70.2%	55.9%	54.8%	—	—	—	—	—
MBPP Code generation	Pass@1	86.4%	79.4%	80.4%	—	—	—	—	—

Table 1 We show evaluation results for reasoning, math, coding, reading comprehension, and question answering. More results on GPQA are given in Table 8.

[3] All GPT scores reported in the GPT-4 Technical Report [40], unless otherwise stated.

[4] All Gemini scores reported in the Gemini Technical Report [41] or the Gemini 1.5 Technical Report [42], unless otherwise stated.

[5] Claude 3 models were evaluated using chain-of-thought prompting.

[6] Researchers have reported higher scores [43] for a newer version of GPT-4T.

[7] GPT-4 scores on MATH (4-shot CoT), MGSM, and Big Bench Hard were reported in the Gemini Technical Report [41].

[8] PubMedQA scores for GPT-4 and GPT-3.5 were reported in [44].

		Claude 3 Opus	Claude 3 Sonnet	Claude 3 Haiku	GPT-4[3]	GPT-3.5[3]
LSAT	5-shot CoT	161	158.3	156.3	**163**	149
MBE	0-shot CoT	**85%**	71%	64%	75.7% (from [51])	45.1% (from [51])
AMC 12[9]	5-shot CoT	**63** / 150	27 / 150	48 / 150	60 / 150	30 / 150
AMC 10[9]	5-shot CoT	**72** / 150	24 / 150	54 / 150	36 / 150[10]	36 / 150
AMC 8[9]	5-shot CoT	84 / 150	54 / 150	36 / 150	–	–
GRE (Quantitative)	5-shot CoT	159	–	–	**163**	147
GRE (Verbal)	5-shot CoT	166	–	–	**169**	154
GRE (Writing)	k-shot CoT	**5.0** (2-shot)	–	–	4.0 (1-shot)	4.0 (1-shot)

Table 2 This table shows evaluation results for the LSAT, the MBE (multistate bar exam), high school math contests (AMC), and the GRE General test. The number of shots used for GPT evaluations is inferred from Appendix A.3 and A.8 of [40].

5.2 Standardized Tests

We evaluated the Claude 3 family of models on the Law School Admission Test (LSAT) [45], the Multistate Bar Exam (MBE) [46], the American Mathematics Competition [47] 2023 math contests, and the Graduate Record Exam (GRE) General Test [48]. See Table 2 for a summary of results.

We obtained LSAT scores for Claude 3 family models by averaging the scaled score of 3 Official LSAT Practice tests: PT89 from Nov 2019, PT90 and PT91 from May 2020. We generated few-shot examples using PT92 and PT93 from June 2020. For the MBE or bar exam, we used NCBE's official 2021 MBE practice exam [49].

We tested our models on all 150 official AMC 2023 problems (50 each from AMC 8, 10, and 12) [47]. Because of high variance, we sampled answers to each question five times at $T = 1$, and report the overall percent answered correctly for each exam multiplied by 150. Official AMC exams have 25 questions, and contestants earn 6 points for correct answers, 1.5 points for skipped questions, and 0 points for incorrect answers, for a maximum possible score of 150.

Our score for Claude Opus was obtained on the Educational Testing Service's official GRE Practice Test 2, with few-shot examples from the official GRE Practice Test 1 [50].

5.3 Vision Capabilities

The Claude 3 family of models are multimodal (image and video-frame input) and have demonstrated significant progress in tackling complex multimodal reasoning challenges that go beyond simple text comprehension.

A prime example is the models' performance on the AI2D science diagram benchmark [52], a visual question answering evaluation that involves diagram parsing and answering corresponding questions in a multiple-choice format. Claude 3 Sonnet reaches the state of the art with 89.2% in 0-shot setting, followed by Claude 3 Opus (88.3%) and Claude 3 Haiku (80.6%) (see Table 3).

All the results in Table 3 have been obtained by sampling at temperature $T = 0$. For AI2D, some images were upsampled such that their longer edges span 800 pixels while preserving their aspect ratios. This upsampling method yielded a 3-4% improvement in performance. For MMMU, we also report Claude 3 models' performance per discipline in Table 3.

Figure 1 shows Claude 3 Opus reading and analyzing a chart, and Appendix B includes some additional vision examples.

[9] For AMC 10 and 12, we evaluated our models on Set A and B for the 2023 exam. For AMC 8, we evaluated our models on the 25-question 2023 exam. GPT scores are for the 2022 exams.

[10] GPT-4 outperforms GPT-4V on AMC 10 [40]; we report the higher score here.

	Claude 3 Opus	Claude 3 Sonnet	Claude 3 Haiku	GPT-4V[11]	Gemini 1.0 Ultra[4]	Gemini 1.5 Pro[4]	Gemini 1.0 Pro[4]
MMMU [3] (val)							
→ Art & Design	67.5%	61.7%	60.8%	65.8%	**70.0%**	—	—
→ Business	**67.2%**	58.2%	52.5%	59.3%	56.7%	—	—
→ Science	48.9%	37.1%	37.1%	**54.7%**	48.0%	—	—
→ Health & Medicine	61.1%	57.1%	52.3%	64.7%	**67.3%**	—	—
→ Humanities & Social Science	70.0%	68.7%	66.0%	72.5%	**78.3%**	—	—
→ Technology & Engineering	**50.6%**	45.0%	41.5%	36.7%	47.1%	—	—
Overall	**59.4%**	53.1%	50.2%	56.8% (from [3])	**59.4%**	58.5%	47.9%
DocVQA [53] (test, ANLS score) Document understanding	89.3%	89.5%	88.8%	88.4%	**90.9%**	86.5%	88.1%
MathVista [54] (testmini) Math	50.5%†	47.9%†	46.4%†	49.9% (from [54])	**53%**	52.1%	45.2%
AI2D [52] (test) Science diagrams	88.1%	**88.7%**	86.7%	78.2%	79.5%	80.3%	73.9%
ChartQA [55] (test, relaxed accuracy) Chart understanding	80.8%†	81.1%†	**81.7%**†	78.5%† 4-shot	80.8%	81.3%	74.1%

Table 3 This table shows evaluation results on multimodal tasks including visual question answering, chart and document understanding. † indicates Chain-of-Thought prompting. All evaluations are 0-shot unless otherwise stated.

[11] All GPT scores reported in the GPT-4V(ision) system card [56], unless otherwise stated.

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] 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.

[4] Anthropic, "Model Card and Evaluations for Claude Models." July, 2023. https://www-cdn.anthropic.com/files/4zrzovbb/website/bd2a28d2535bfb0494cc8e2a3bf135d2e7523226.pdf.

[5] Anthropic, "Anthropic's Responsible Scaling Policy." September, 2023. https://www.anthropic.com/index/anthropics-responsible-scaling-policy.

[6] Anthropic, "Claude's Constitution." May, 2023. https://www.anthropic.com/index/claudes-constitution.

[7] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, eds., pp. 8024–8035. Curran Associates, Inc., 2019. http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

[8] J. Bradbury, R. Frostig, P. Hawkins, M. J. Johnson, C. Leary, D. Maclaurin, G. Necula, A. Paszke, J. VanderPlas, S. Wanderman-Milne, and Q. Zhang, "JAX: composable transformations of Python+NumPy programs." 2018. http://github.com/google/jax.

[9] P. Tillet, H. T. Kung, and D. Cox, *Triton: An Intermediate Language and Compiler for Tiled Neural Network Computations*, pp. 10–19. Association for Computing Machinery, New York, NY, USA, 2019. https://doi.org/10.1145/3315508.3329973.

[10] Anthropic, "Challenges in evaluating AI systems." October, 2023. https://www.anthropic.com/index/evaluating-ai-systems.

[11] Anthropic, "Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviors, and Lessons Learned." August, 2022. https://www.anthropic.com/index/red-teaming-language-models-to-reduce-harms-methods-scaling-behaviors-and-lessons-learned.

[12] Anthropic, "The Capacity for Moral Self-Correction in Large Language Models." February, 2023. https://www.anthropic.com/index/the-capacity-for-moral-self-correction-in-large-language-models.

[13] E. Durmus, K. Nyugen, T. I. Liao, N. Schiefer, A. Askell, A. Bakhtin, C. Chen, *et al.*, "Towards measuring the representation of subjective global opinions in language models." 2023.

[14] Anthropic, "Frontier Threats Red Teaming for AI Safety." July, 2023. https://www.anthropic.com/index/frontier-threats-red-teaming-for-ai-safety.

[15] Anthropic, "Acceptable Use Policy," https://console.anthropic.com/legal/aup.

[16] Y. Bai, S. Kadavath, S. Kundu, A. Askell, J. Kernion, A. Jones, A. Chen, *et al.*, "Constitutional AI: Harmlessness from AI Feedback." 2022. https://arxiv.org/abs/2212.08073.

[17] Anthropic, "Collective Constitutional AI: Aligning a Language Model with Public Input." October, 2023. https://www.anthropic.com/index/collective-constitutional-ai-aligning-a-language-model-with-public-input.

[18] "Dataset Card for HH-RLHF," https://huggingface.co/datasets/Anthropic/hh-rlhf.

[19] Y. Bai, A. Jones, K. Ndousse, A. Askell, A. Chen, N. DasSarma, D. Drain, *et al.*, "Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback," *arXiv preprint arXiv:2204.05862* (April, 2022) . https://arxiv.org/abs/2204.05862.

[20] National Institute of Standards and Technology, "Artificial Intelligence Risk Management Framework." January, 2023. https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf.

[21] "Anthropic Privacy Policy." July, 2023. https://console.anthropic.com/legal/privacy.

[22] P. Clark, I. Cowhey, O. Etzioni, T. Khot, A. Sabharwal, C. Schoenick, and O. Tafjord, "Think you have Solved Question Answering? Try ARC, the AI2 Reasoning Challenge." March, 2018.

[23] Q. Jin, B. Dhingra, Z. Liu, W. W. Cohen, and X. Lu, "PubMedQA: A Dataset for Biomedical Research Question Answering." September, 2019.

[24] 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) .

[25] 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) .

[26] 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.

[27] R. Zellers, A. Holtzman, Y. Bisk, A. Farhadi, and Y. Choi, "HellaSwag: Can a Machine Really Finish Your Sentence?" May, 2019.

[28] K. Sakaguchi, R. L. Bras, C. Bhagavatula, and Y. Choi, "WinoGrande: An Adversarial Winograd Schema Challenge at Scale." November, 2019.

[29] 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.

[30] G. Lai, Q. Xie, H. Liu, Y. Yang, and E. Hovy, "RACE: Large-scale ReAding Comprehension Dataset From Examinations," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 785–794. Association for Computational Linguistics, Copenhagen, Denmark, Sept., 2017. https://aclanthology.org/D17-1082.

[31] R. Y. Pang, A. Parrish, N. Joshi, N. Nangia, J. Phang, A. Chen, V. Padmakumar, J. Ma, J. Thompson, H. He, *et al.*, "QuALITY: Question Answering with Long Input Texts, Yes!," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 5336–5358. 2022.

[32] 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) .

[33] D. Hendrycks, S. Basart, S. Kadavath, M. Mazeika, A. Arora, E. Guo, C. Burns, S. Puranik, H. He, D. Song, and J. Steinhardt, "Measuring Coding Challenge Competence With APPS," *NeurIPS* (November, 2021) .

[34] J. Austin, A. Odena, M. Nye, M. Bosma, H. Michalewski, D. Dohan, E. Jiang, C. Cai, M. Terry, Q. Le, and C. Sutton, "Program Synthesis with Large Language Models." August, 2021.

[35] 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.

[36] 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.

[37] X. Wang, J. Wei, D. Schuurmans, Q. Le, E. Chi, S. Narang, A. Chowdhery, and D. Zhou, "Self-Consistency Improves Chain of Thought Reasoning in Language Models." March, 2023. https://arxiv.org/abs/2203.11171.

[38] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, and D. Zhou, "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models." January, 2023. https://arxiv.org/abs/2201.11903.

[39] S. Bubeck, V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, *et al.*, "Sparks of Artificial General Intelligence: Early experiments with GPT-4." April, 2023.

[40] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, *et al.*, "GPT-4 Technical Report." 2023.

[41] Gemini Team, R. Anil, S. Borgeaud, Y. Wu, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, *et al.*, "Gemini: A Family of Highly Capable Multimodal Models." December, 2023.

[42] Gemini Team, Google, "Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context." December, 2023. https://storage.googleapis.com/deepmind-media/gemini/gemini_v1_5_report.pdf.

[43] Microsoft, "promptbase." December, 2023. https://github.com/microsoft/promptbase.

[44] H. Nori, N. King, S. M. McKinney, D. Carignan, and E. Horvitz, "Capabilities of GPT-4 on Medical Challenge Problems." April, 2023.

[45] Law School Admission Council, "The LSAT." February, 2024. https://www.lsac.org/lsat.

[46] N. C. of Bar Examiners, "Multistate Bar Examination," https://www.ncbex.org/exams/mbe. Accessed: 2023-07-03.

[47] Mathematical Association of America, "About AMC | Mathematical Association of America." February, 2024. https://maa.org/math-competitions/about-amc.

[48] Educational Testing Services, "The GRE Tests." February, 2024. https://www.ets.org/gre.html.

[49] N. C. of Bar Examiners, "NCBE Releases First Full-Length Simulated MBE Study Aid," https://www.ncbex.org/news-resources/ncbe-releases-first-full-length-simulated-mbe-study-aid, 2021. Accessed: 2023-07-03.

[50] ETS, "POWERPREP Practice Tests: Prepare for the GRE General Test," https://www.ets.org/gre/test-takers/general-test/prepare/powerprep.html. Accessed: 2024-02-24.

[51] D. M. Katz, M. J. Bommarito, S. Gao, and P. Arredondo, "GPT-4 Passes the Bar Exam," *SSRN preprint* (April, 2023) . https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4389233.

[52] 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.

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

[54] 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.

[55] 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.

[56] OpenAI, "GPT-4V(ision) System Card." September, 2023. https://cdn.openai.com/papers/GPTV_System_Card.pdf.

[57] P. R. Center, "Americans' Social Media Use," https://www.pewresearch.org/internet/2024/01/31/americans-social-media-use, January, 2024. Accessed: 2024-02-24.

[58] 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.

[59] 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.

[60] "Supported Countries and Regions," https://www.anthropic.com/claude-ai-locations.

[61] V. Dac Lai, C. Van Nguyen, N. T. Ngo, T. Nguyen, F. Dernoncourt, R. A. Rossi, and T. H. Nguyen, "Okapi: Instruction-tuned Large Language Models in Multiple Languages with Reinforcement Learning from Human Feedback," *arXiv e-prints* (August, 2023) arXiv–2307.

[62] Anthropic, "Introducing 100K Context Windows." May, 2023. https://www.anthropic.com/news/100k-context-windows.

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

[64] N. F. Liu, K. Lin, J. Hewitt, A. Paranjape, M. Bevilacqua, F. Petroni, and P. Liang, "Lost in the Middle: How Language Models Use Long Contexts," *Transactions of the Association for Computational Linguistics* **12** (November, 2023) 157–173.

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

[66] The White House, "FACT SHEET: Biden-Harris Administration Secures Voluntary Commitments from Leading Artificial Intelligence Companies to Manage the Risks Posed by AI." July, 2023. https://www.whitehouse.gov/briefing-room/statements-releases/2023/07/21/fact-sheet-biden-harris-administration-secures-voluntary-commitments-from-leading-artificial-intelligence-companies-to-ma

[67] The White House, "FACT SHEET: President Biden Issues Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence." October, 2023. https://www.whitehouse.gov/briefing-room/statements-releases/2023/10/30/fact-sheet-president-biden-issues-executive-order-on-safe-secure-and-trustworthy-artificial-intelligence/.

[68] UK Dept. for Science, Innovation & Technology, "Emerging processes for frontier AI safety." October, 2023. https://www.gov.uk/government/publications/emerging-processes-for-frontier-ai-safety/emerging-processes-for-frontier-ai-safety#executive-summary.

[69] A. Krithara, A. Nentidis, B. Konstantinos, and G. Paliouras, "BioASQ-QA: A manually curated corpus for Biomedical Question Answering," *Scientific Data* **10** (2023) .

[70] USMLE, "About the USMLE and Why It's Important," https://www.usmle.org/bulletin-information/about-usmle. Accessed: 2023-07-08.

[71] A. Pal, L. K. Umapathi, and M. Sankarasubbu, "MedMCQA: A Large-scale Multi-Subject Multi-Choice Dataset for Medical domain Question Answering," in *Proceedings of the Conference on Health, Inference, and Learning*, G. Flores, G. H. Chen, T. Pollard, J. C. Ho, and T. Naumann, eds., vol. 174 of *Proceedings of Machine Learning Research*, pp. 248–260. PMLR, 07–08 apr, 2022. https://proceedings.mlr.press/v174/pal22a.html.

[72] A. Tamkin, A. Askell, L. Lovitt, E. Durmus, N. Joseph, S. Kravec, K. Nguyen, J. Kaplan, and D. Ganguli, "Evaluating and Mitigating Discrimination in Language Model Decisions," *arXiv preprint arXiv:2312.03689* (December, 2023) .

[73] A. Parrish, A. Chen, N. Nangia, V. Padmakumar, J. Phang, J. Thompson, P. M. Htut, and S. R. Bowman, "BBQ: A Hand-Built Bias Benchmark for Question Answering," in *CoRR*. March, 2022.