Centre for International Governance Innovation

Digital Policy Hub

Digital Policy Hub – Working Paper

Are Large Language Models Actually Getting Safer?

Ashley Ferreira

Fall 2024 cohort

About the Hub

The Digital Policy Hub at CIGI is a collaborative space for emerging scholars and innovative thinkers from the social, natural and applied sciences. It provides opportunities for undergraduate and graduate students and post-doctoral and visiting fellows to share and develop research on the rapid evolution and governance of transformative technologies. The Hub is founded on transdisciplinary approaches that seek to increase understanding of the socio-economic and technological impacts of digitalization and improve the quality and relevance of related research. Core research areas include data, economy and society; artificial intelligence; outer space; digitalization, security and democracy; and the environment and natural resources.

The Digital Policy Hub working papers are the product of research related to the Hub's identified themes prepared by participants during their fellowship.

About CIGI

The Centre for International Governance Innovation (CIGI) is an independent, non-partisan think tank whose peer-reviewed research and trusted analysis influence policy makers to innovate. Our global network of multidisciplinary researchers and strategic partnerships provide policy solutions for the digital era with one goal: to improve people's lives everywhere. Headquartered in Waterloo, Canada, CIGI has received support from the Government of Canada, the Government of Ontario and founder Jim Balsillie.

Partners

Thank you to Mitacs for its partnership and support of Digital Policy Hub fellows through the Accelerate program. We would also like to acknowledge the many universities, governments and private sector partners for their involvement allowing CIGI to offer this holistic research environment.

67 Erb Street West
Waterloo, ON, Canada N2L 6C2
www.cigionline.org

Key Points

- Each new release of large language models (LLMs) often comes with claims of both improved performance and enhanced safety. However, there is a lack of standardized safety assessments and a gap in studying these metrics over time.
- This working paper aims to address this gap by analyzing performance on various standardized safety benchmarks across various LLMs released in the last three years to gauge if they are becoming safer.
- Under this method of evaluation, newer models are overall scoring higher on these benchmarks; however, these improvements are not dramatic, and when the newer models do fail, these failures are far more consequential as more current models are more capable of causing harm.
- Going forward, these safety benchmarks should consider this added dimension of quantifying how harmful LLM failures can be.
- It is recommended to devise a system in which the vulnerabilities of LLMs can be studied, shared and addressed, but the specifics on how to exploit them are guarded by bad actors.
- Finally, since improvements in safety do not seem to be naturally keeping pace with improvements in overall artificial intelligence (AI), more external pressure is required to ensure we sufficiently guard against the release of dangerous models.

Introduction

It is widely recognized that LLMs have risen to remarkable prominence in recent years, fundamentally transforming many aspects of our world. Their rise can be traced to the 2017 invention of a specific deep learning model architecture called the transformer (Vaswani et al. 2017) and the subsequent development of powerful systems such as OpenAI's GPT-3 model (Brown et al. 2020) and the initial release of ChatGPT (OpenAI 2022).

However, as LLMs have become more powerful, concerns over their safety have escalated (Yudkowsky 2023; Hendrycks, Mazeika and Woodside 2023; Future of Life Institute 2023).[1] These concerns have led to the emergence of AI safety as a growing field, in which researchers focus on identifying, mitigating and regulating risks associated with AI systems. In particular, the field of AI safety benchmarking aims to systematically evaluate the safety and reliability of LLMs using structured tests. AI safety benchmarks are designed to measure various aspects of undesirable and harmful model behaviour (Ren et al. 2024).

With each new model release, companies often claim that their LLMs are becoming safer; for example, OpenAI has made repeated assertions of this in its successive

1 See www.safe.ai/work/statement-on-ai-risk.

LLM versions (OpenAI et al. 2024).[2] But despite these claims, there is a notable lack of independent verification on standardized safety tests (Kaiyom et al. 2024). Without consistent, widely adopted benchmarks, it remains difficult to objectively measure whether LLMs are genuinely becoming safer across releases. This difficulty is not necessarily malicious; it is, by and large, a product of the fact that the field of AI safety is actively under development and it matures simultaneously as LLMs mature. As such, the most holistic and widely respected safety benchmarks have only begun to be released recently (Ghosh et al. 2025).

This working paper assesses the safety improvements of LLMs over the last three years, focusing on the most widely recognized frontier model providers: OpenAI, Anthropic and Meta. This assessment has been made by analyzing the performance of different generations of LLMs on consistent evaluation frameworks and comparing their performance over time. The goal of this assessment is to determine whether claims of enhanced safety align with actual observed improvements from an impartial third party.

Measuring LLM Safety

In recent years, there has been a concerted effort to quantitatively measure the safety of LLMs, known as safety benchmarking, which is done both by the companies that release the models themselves and by third-party evaluators (OpenAI 2024a; Vidgen et al. 2024). AI safety benchmarking works by defining a set of safety metrics, such as robustness against adversarial inputs, bias detection, ability to reject unsafe prompts and adherence to ethical guidelines. These metrics are tested using standardized data sets and scenarios specifically designed to challenge the LLM. For instance, benchmarks such as RealToxicityPrompts (Gehman et al. 2020) evaluate whether a model produces toxic outputs in response to certain defined prompts. The goal of AI safety benchmarking is to ensure that as LLMs continue to evolve, they not only get more powerful but also safer and more trustworthy. However, there has been recent criticism as to whether current AI benchmarks accurately measure safety separately from general model improvement (Ren et al. 2024), and there remain numerous significant challenges to reliably measure AI systems (Ganguli et al. 2023).

Mechanically, there are various ways to evaluate the LLM responses, which can vary from a human annotator or even another LLM scoring the response, to requiring that the LLM output a multiple-choice response for which a scorecard already exists (Vidgen et al. 2024). Several collaborations are currently under way to produce holistic AI safety benchmarks (Ghosh et al. 2025).

Instead of picking one benchmark, this assessment uses various benchmarks that have already been used to evaluate frontier LLMs. All of the benchmarks in Table 1 aim to measure safety broadly by combining scores from different dimensions such as bias or susceptibility to jailbreaking. Most of these benchmarks pull from previous benchmarks that focused on specific aspects of safety. For example, the Holistic Evaluation of Language Models (HELM) Safety Leaderboard (Liang et al. 2023) aggregates five benchmarks that each emphasize a distinct safety risk vector to provide a holistic

2 See OpenAI (2024b) or OpenAI (n.d.).

measure of AI safety, primarily through LLM automated evaluations: Bias Benchmark for QA for risks of social discrimination; SimpleSafetyTest for obviously harmful requests; HarmBench for susceptibility to jailbreaking techniques; XSTest for evaluating edge cases and strict refusals; and AnthropicRedTeam to evaluate elicitation of harmful model behaviour by human and model-assisted red-teaming (ibid.).

Table 1: List of AI Safety Benchmarks and Leaderboards Used in Analysis

Name	Overview	Paper Citation
Libra-Leaderboard	LibrAI initiative to measure the performance and safety of LLMs that combines saferty benchmarks from 57 datasets that cover different evaluation techniques.	(H. Li et al. 2025)
SALAD-Bench	Led by the Shanghai Artificial Intelligence Laboratory, this benchmark provides a hierarchical taxonomy of safety dimensions. Models are evaluated by LLMs for simple question answering, as well as attack-ehanced and multiple choice question answering that can be evaluated without an LLM.	(L. Li et al. 2024)
DecodingTrust	This highly awarded research endeavor aims to assess the trustworthiness of LLMs using various evaluation techniques.	(Wang et al. 2023)
SafetyBench	Using 11,435 multiple choice questions, this benchmark evaluates safety in both English and Chinese.	(Zhang et al. 2024)
HELM Safety Leaderboard	HELM was created by Stanford University's Center for Research on Foundation Models and one of the leaderboards they provided focuses on safety.	(Liang et al. 2023)

Source: Author.

Monitoring Safety Over Time

Recently, more independent research labs have been working to evaluate frontier models and even creating LLM safety leaderboards (Li, Tang and Fourrier 2024; Wang et al. 2023). However, these leaderboards focus on trying to determine with the most accuracy which of the frontier models is currently the safest and not what the trend of these metrics is over time. In addition, most of these leaderboards focus only on one or two models from each organization, with GPT-3.5 and GPT-4 being the most popular OpenAI models to evaluate. The DecodingTrust benchmark (Wang et al. 2023), in particular, did a deep-dive comparing these two OpenAI models and found that both can be easily misled to generate toxic and biased outputs, as well as leak private information. But, as expected, GPT-4 is overall more trustworthy when benchmarked; however, this work also noted GPT-4 is more vulnerable to jailbreaking than GPT-3.5.

The big outstanding question of whether we are on the right track when it comes to improving AI safety can be partly answered by comparing the safety of various generations of frontier LLMs. The goal of this comparison is to determine whether claims

of enhanced safety align with actual significant observed improvements from a third party.[3]

Figure 1 below compares the performance of LLMs released over the past few years based on four safety evaluations. While there are some cases in which it is clear that some newer generations of LLMs score higher than older ones, this is also not the only factor in determining these scores. The openly available frontier models, such as Meta's Llama series, score systematically lower on these benchmarks than the frontier models only accessible through a limited chat interface or application programming interface (API), such as OpenAI and Anthropic's families of models.

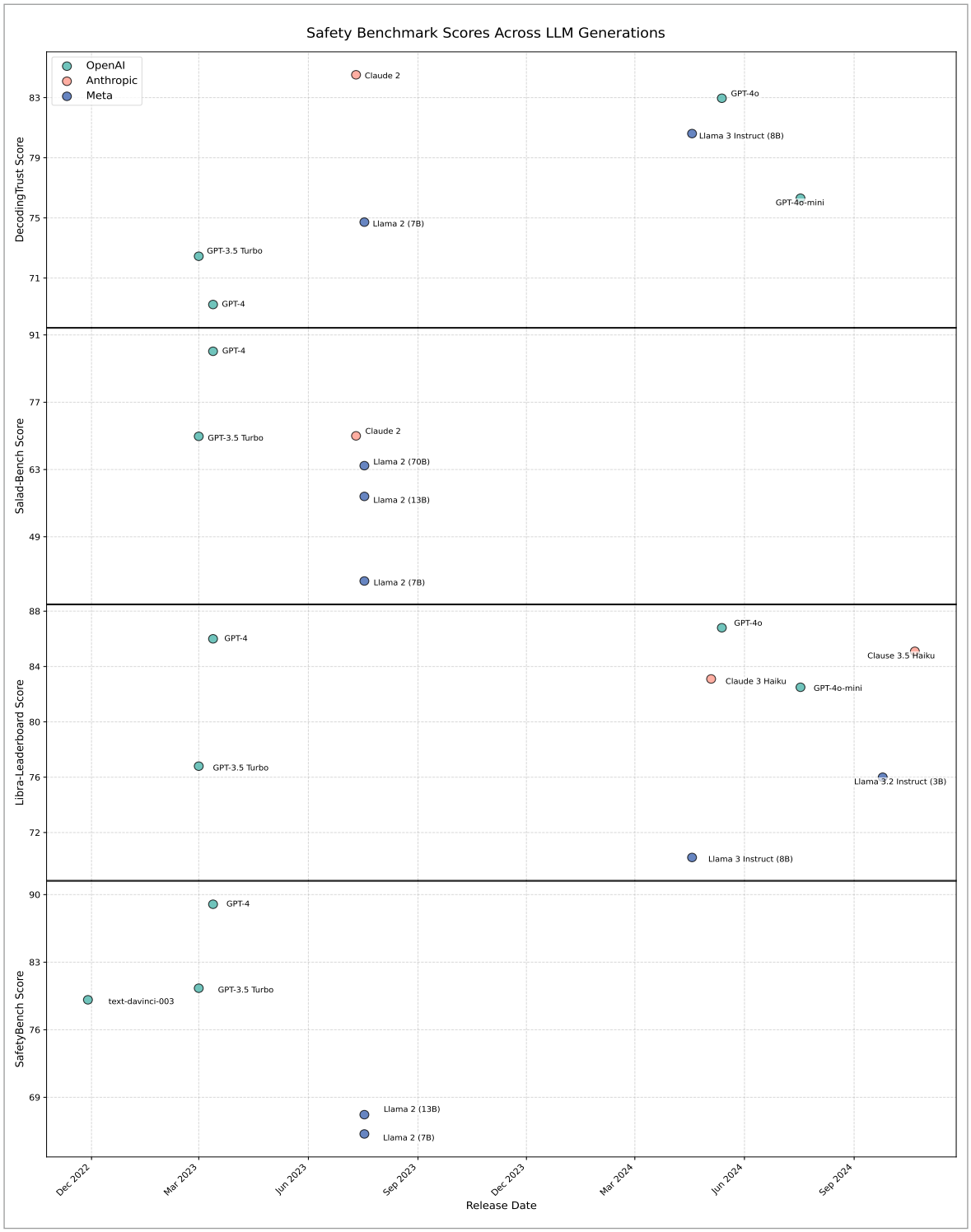
While users may not notice a difference when using ChatGPT, OpenAI has been iterating on the base model behind the scenes to go from GPT-3 to GPT-3.5 to GPT-3.5 Turbo, all the way to GPT-4 and GPT-4o in limited availability and for paid users (OpenAI 2024).[4] Models as far back as GPT-3.5 remain available through their API, while GPT-1 and GPT-2 are available through a model-hosting service called Hugging Face;[5] however, they are far less capable and controllable (Radford et al. 2018, 2019) than their successors and therefore extremely challenging to evaluate, which is why they are not included in these benchmarks. For certain named model releases (such as GPT 3.5 Turbo), there are also updates to these models; therefore, it is also important to consider which snapshot of the model is being evaluated.

3 For the raw data and further details from this analysis, please see the Technical Appendix: https://github.com/ashley-ferreira/LLM-safety-tracker.

4 See https://huggingface.co/docs/transformers/en/model_doc/openai-gpt.

5 See https://huggingface.co/.

Figure 1: LLM Scores on Various Safety Benchmarks



Source: Figure generated by author.

Figure 2 dives deeper into this analysis by showing model scores on the HELM Safety Leaderboard compared to their corresponding snapshot date for a consistent set of 22 models. The top panel shows the average scores, where it seems once again that, by and large, there are improvements in the scores with time; however, they are nowhere near the improvements seen with general performance. The panels below show the breakdown of the various benchmarks that contribute to this average score. The scores themselves are very near saturation, which provides the false illusion that these models are very safe. This is a misleading narrative (which the HELM team itself actively works against) for four specific reasons:

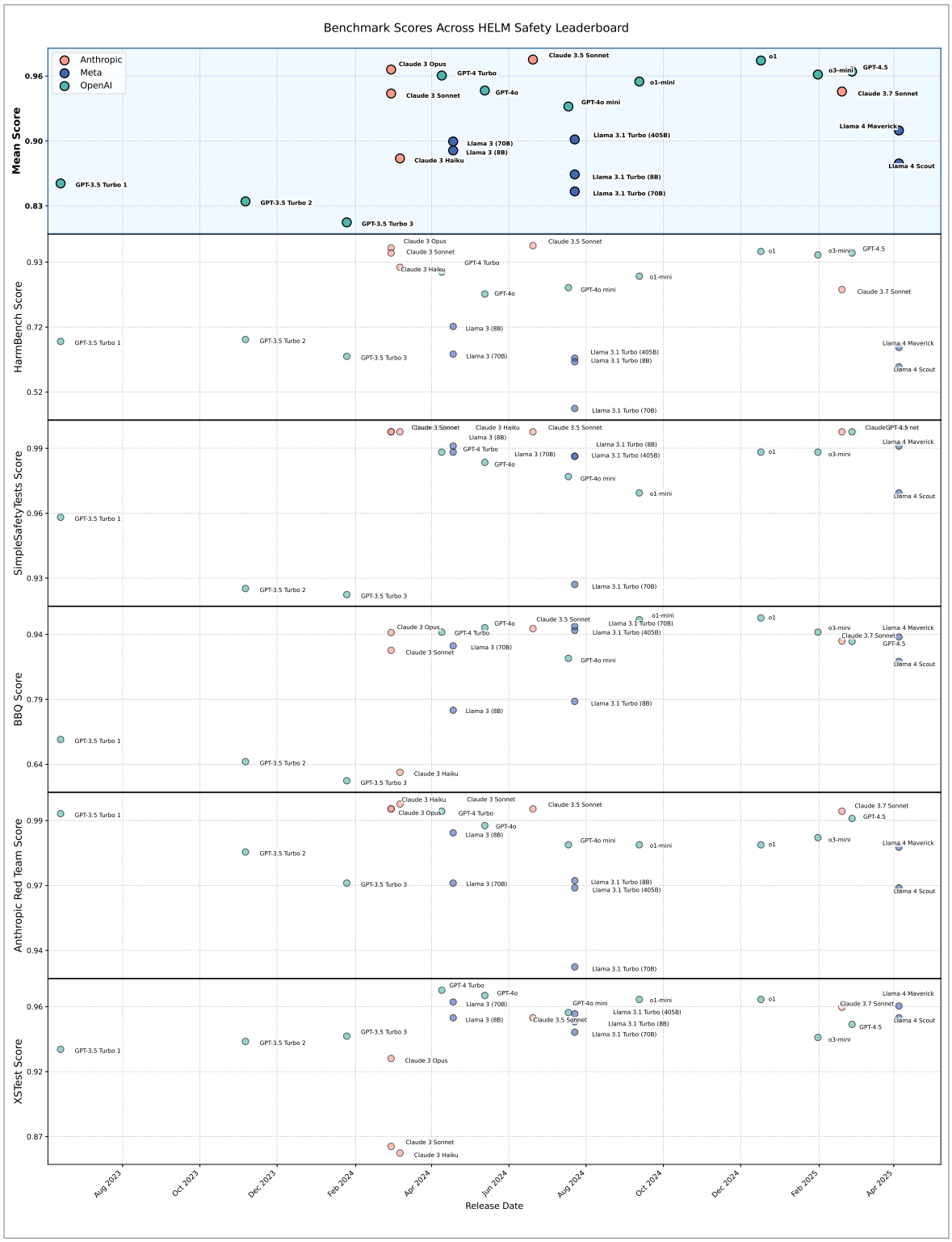
- even an LLM with just one failure mode can be harmful to release publicly;
- newer jailbreaking methods can dramatically lower these scores;
- there are major limitations with most safety benchmarks; and
- it is challenging for these evaluations to adequately capture the amount of harm a model can cause.

When exploring responses to these benchmarks, one very strong quantitative trend is how much more risk resides in a very advanced model that is able to be jailbroken compared to a much less capable one. For example, if earlier models are asked to provide steps for a dangerous task, such as creating a weapon, they may comply and provide instructions, but the instructions will likely be wrong, so the jailbroken model is still not very harmful. This seems to be a blind spot in current AI safety benchmarks that is very challenging to patch as it would require evaluating the effectiveness of the replies.

To build on this idea further, there is also an inherent risk in deploying models that are more intelligent than humans and not aligend with our best interests, which is not captured in these evaluations and is a large factor that some LLM providers such as Anthropic use to decide whether models can be responsibly released.[6] Currently, this finding contradicts the results from many benchmarks, which show improved safety over time. For example, when the Libra-Leaderboard team studied the relationship between safety scores and general capability scores, their results showed a strong correlation between these metrics (H. Li et al. 2025).

6 See www.anthropic.com/rsp-updates.

Figure 2: Safety Benchmark Scores for a Consistent Set of 22 LLMs



Source: Figure generated by author; see https://github.com/ashley-ferreira/LLM-safety-tracker for more details.

Conclusions

The results shown above lead to the conclusion that, at least for the evaluations and models considered, LLMs are indeed becoming better at obeying safety guardrails. This trend may not be as strong as expected, and there still exist many points of failure for these LLMs (Mazeika et al. 2024). The main risk will likely be the effects of increasingly more intelligent models, for which safety is an inherent concern, as advanced intelligence itself could be threatening to humanity (Hendrycks, Mazeika and Woodside 2023).

More work is needed to solidify these results, in particular, to further isolate the signal that comes from the progression over time. These evaluations should also be more thorough and include more detailed breakdowns of performance across various domains of safety concerns. However, due to the ongoing challenges in measuring AI models (Ganguli et al. 2023), it remains difficult to fully answer the question of whether models are truly getting safer.

Recommendations

There are three broad recommendations resulting from this working paper. The first is a reminder that even though these models are improving on certain aspects of their safety evaluations, they remain vulnerable to jailbreaking (Chao et al. 2024; Chen and Lu 2024), which becomes an increasingly big risk as models continue to improve in general performance. While work into identifying and exploiting the vulnerabilities of LLMs is extremely valuable, the public sharing of tricks that allow people to bypass the guardrails poses a significant risk, and a framework should be put in place by AI safety institutions to minimize the chance that these strategies make it into the hands of bad actors. This approach could be modelled on existing cybersecurity frameworks such as a coordinated flaw-reporting mechanism (Longpre et al. 2025).

The second recommendation is that future work on AI safety benchmarking should incorporate an additional dimension of examining not only whether the model is susceptible to being compromised but also how harmful it could be, should this occur. This safeguard would allow for the better identification of areas where compromised models could be most harmful. Following this recommendation would likely mean moving away from prioritizing issues such as toxicity, which seem to be naturally improving and where compromised models have harmful but limited risk, and instead focusing on more catastrophic risks that AI companies are themselves unequipped to handle alone, such as how to collectively govern models that are extremely powerful (Cass-Beggs et al. 2024).

Finally, there seems to be immense pressure for frontier AI labs to develop increasingly more capable models that perform dramatically better on capability-specific benchmarks than their predecessors and even surpass human performance in some areas (Maslej et al. 2024). However, the same is not true for safety benchmarks. Therefore, if we want to see significant improvements in model safety, it will likely need to be encouraged through regulation.

Author's Note

The views and content presented in this paper are solely those of the author and do not represent any affiliated organizations or individuals.

Acknowledgments

The author would like to thank Karim Sallaudin Karim, Matthew da Mota and Halyna Padalko for their feedback on this work. The author is also very grateful for administrative support from Reanne Cayenne and Dianna English, as well as the many insightful presentations put on throughout the Digital Policy Hub program. Finally, the author would like to acknowledge the insightful presentations and discussion during the 2024 Conference on Neural Information Processing Systems workshops from which this work has greatly benefited (particularly those on Socially Responsible Language Modelling Research, Regulatable ML and Red Teaming GenAI).

Works Cited

Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan et al. 2020. "Language Models are Few-Shot Learners." *Advances in Neural Information Processing Systems* 33: 1877–901. https://arxiv.org/abs/2005.14165.

Cass-Beggs, Duncan, Stephen Clare, Dawn Dimowo and Zaheed Kara. 2024. *Framework Convention on Global AI Challenges.* CIGI Discussion Paper. Waterloo, ON: CIGI. www.cigionline.org/publications/framework-convention-on-global-ai-challenges/.

Chao, Patrick, Edoardo Debenedetti, Alexander Robey, Maksym Andriushchenko, Francesco Croce, Vikash Sehwag, Edgar Dobriban et al. 2024. "JailbreakBench: An Open Robustness Benchmark for Jailbreaking Large Language Models." *Advances in Neural Information Processing Systems* 37: 55005–29. https://proceedings.neurips.cc/paper_files/paper/2024/file/63092d79154adebd7305dfd498cbff70-Paper-Datasets_and_Benchmarks_Track.pdf.

Chen, Jay and Royce Lu. 2024. "Deceptive Delight: Jailbreak LLMs Through Camouflage and Distraction." *Unit 42* (blog), October 23. https://unit42.paloaltonetworks.com/jailbreak-llms-through-camouflage-distraction/.

Future of Life Institute. 2023. "Pause Giant AI Experiments: An Open Letter." Open letter, March 22. https://futureoflife.org/open-letter/pause-giant-ai-experiments/.

Ganguli, Deep, Nicholas Schiefer, Marina Favaro and Jack Clark. 2023. "Challenges in evaluating AI systems." Anthropic, October 4. www.anthropic.com/research/evaluating-ai-systems.

Gehman, Samuel, Suchin Gururangan, Maarten Sap, Yejin Choi and Noah A. Smith. 2020. "RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models." In *Findings of the Association for Computational Linguistics* 3356–69. https://doi.org/10.18653/v1/2020.findings-emnlp.301.

Ghosh, Shaona, Heather Frase, Adina Williams, Sarah Luger, Paul Röttger, Fazl Barez, Sean McGregor et al. 2025. "AILuminate: Introducing v1.0 of the AI Risk and Reliability Benchmark from MLCommons." Preprint, *arXiv*, April 18. https://arxiv.org/abs/2503.05731.

Hendrycks, Dan, Mantas Mazeika and Thomas Woodside. 2023. "An Overview of Catastrophic AI Risks." Preprint, *arXiv*, October 9. http://arxiv.org/abs/2306.12001.

Kaiyom, Farzaan, Ahmed Ahmed, Yifan Mai, Kevin Klyman, Rishi Bommasani and Percy Liang. 2024. "HELM Safety: Towards Standardized Safety Evaluations of Language Models." Stanford, CA: Stanford Center for Research on Foundation Models. https://crfm.stanford.edu/2024/11/08/helm-safety.html.

Li, Haonan, Xudong Han, Zenan Zhai, Honglin Mu, Hao Wang, Zhenxuan Zhang, Yilin Geng et al. 2025. “Libra-Leaderboard: Towards Responsible AI through a Balanced Leaderboard of Safety and Capability.” In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (System Demonstrations),* 268–86.

Li, Lijun, Bowen Dong, Ruohui Wang, Xuhao Hu, Wangmeng Zuo, Dahua Lin, Yu Qiao et al. 2024. “SALAD-Bench: A Hierarchical and Comprehensive Safety Benchmark for Large Language Models.” *Findings of the Association for Computational Linguistics*, 3923–54. https://arxiv.org/abs/2402.05044.

Li, Steve, Leonard Tang and Clémentine Fourrier. 2024. “Introducing the Red-Teaming Resistance Leaderboard.” *Hugging Face* (blog), February 23. https://huggingface.co/blog/leaderboard-haizelab.

Liang, Percy, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang et al. 2023. “Holistic Evaluation of Language Models.” *Transactions on Machine Learning Research*, 2835–56. https://arxiv.org/abs/2211.09110.

Longpre, Shayne, Kevin Klyman, Ruth E. Appel, Sayash Kapoor, Rishi Bommasani, Michelle Sahar, Sean McGregor et al. 2025. “In-House Evaluation Is Not Enough: Towards Robust Third-Party Flaw Disclosure for General-Purpose AI.” Preprint, *arXiv*, March 25. https://arxiv.org/abs/2503.16861.

Maslej, Nestor, Loredana Fattorini, Raymond Perrault, Vanessa Parli, Anka Reuel, Erik Brynjolfsson, John Etchemendy et al. 2024. *Artificial Intelligence Index Report 2024*. Stanford University Human-Centered Artificial Intellignece. https://hai.stanford.edu/ai-index/2024-ai-index-report.

Mazeika, Mantas, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee et al. 2024. “HarmBench: A Standardized Evaluation Framework for Automated Red Teaming and Robust Refusal.” *Proceedings of the 41st International Conference on Machine Learning* 235: 35181–224. https://arxiv.org/abs/2402.04249.

OpenAI. n.d. “GPT-4 System Card.” https://cdn.openai.com/papers/gpt-4-system-card.pdf.

– – –. 2022. “Introducing ChatGPT.” November 30. https://openai.com/index/chatgpt/.

– – –. 2024a. “Hello GPT-4o.” May 13. https://openai.com/index/hello-gpt-4o/.

– – –. 2024b. *OpenAI o1 System Card*. December 5. https://openai.com/index/openai-o1-system-card/.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman et al. 2024. “GPT-4 Technical Report.” Preprint, *arXiv*, March 4. http://arxiv.org/abs/2303.08774.

Radford, Alec, Karthik Narasimhan, Tim Salimans and Ilya Sutskever. 2018. “Improving Language Understanding by Generative Pre-Training.” Preprint, OpenAI. https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf.

Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei and Ilya Sutskever. 2019. “Language Models are Unsupervised Multitask Learners.” Preprint, OpenAI. https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

Ren, Richard, Steven Basart, Adam Khoja, Alice Gatti, Long Phan, Xuwang Yin et al. 2024. "Safetywashing: Do AI Safety Benchmarks Actually Measure Safety Progress?" *Advances in Neural Information Processing Systems* 37: 68559–94. https://proceedings.neurips.cc/paper_files/paper/2024/file/7ebcdd0de471c027e67a11959c666d74-Paper-Datasets_and_Benchmarks_Track.pdf.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser et al. 2017. "Attention Is All You Need." In *Advances in Neural Information Processing Systems* 30: 1–11. https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Vidgen, Bertie, Adarsh Agrawal, Ahmed M. Ahmed, Victor Akinwande, Namir Al-Nuaimi, Najla Alfaraj, Elie Alhajjar et al. 2024. "Introducing v0.5 of the AI Safety Benchmark from MLCommons." Preprint, *arXiv*, May 13. http://arxiv.org/abs/2404.12241.

Wang, Boxin, Weixin Chen, Hengzhi Pei, Chulin Xie, Mintong Kang, Chenhui Zhang, Chejian Xu et al. 2023. "DecodingTrust: A Comprehensive Assessment of Trustworthiness in GPT Models." *Advances in Neural Information Processing Systems* 36: 31232–339. https://proceedings.neurips.cc/paper_files/paper/2023/file/63cb9921eecf51bfad27a99b2c53dd6d-Paper-Datasets_and_Benchmarks.pdf.

Yudkowsky, Eliezer. 2023. "Pausing AI Developments Isn't Enough. We Need to Shut it All Down." *TIME*, March 29. https://time.com/6266923/ai-eliezer-yudkowsky-open-letter-not-enough/.

Zhang, Zhexin, Leqi Lei, Lindong Wu, Rui Sun, Yongkang Huang, Chong Long, Xiao Liu et al. 2024. "SafetyBench: Evaluating the Safety of Large Language Models." *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics,* 15537–53.