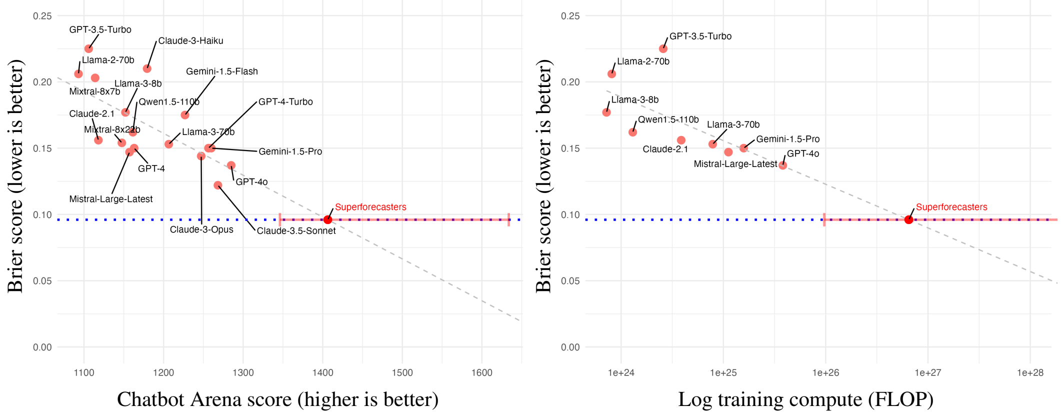

(a) Brier score vs. Chatbot Arena (b) Brier score vs. Log training compute

Figure 1: The graphs show the linear relationship between the Brier scores from Table 2 and **(a)** Chatbot Arena scores and **(b)** estimates of training compute. The dotted blue line represents the Superforecasters' overall Brier score. A red dot with a bootstrapped 95% confidence interval is placed at the intersection of this dotted blue line with the dashed linear fit line to demonstrate the potential intersection of LLM Arena score/training compute and Superforecaster-level forecasting performance. For **(b)**, if estimates from Epoch AI (2024) were not available, we produced estimates following `https://epoch.ai/blog/estimating-training-compute`. The trend-line in **(a)** is $y = 0.506 - 0.000298x$ ($R^2 = 0.47$) and in **(b)** it is $y = 0.844 - 0.01213x$ ($R^2 = 0.41$).

could match superforecaster performance when the Arena score approaches 1406 (bootstrapped 95% CI: 1346–1633).

Figure 1b shows the log-linear relationship between estimated training compute and the overall Brier score from Table 2. Projecting out the log-linear relationship, we find that LLMs could match superforecaster performance when training compute approaches 6.49×10^{26}, though there is a large confidence interval (bootstrapped 95% CI: 9.69×10^{25}–8.65×10^{28}) given the marginally significant relationship ($r = -0.67$, $p = 0.046$).

6 Discussion

We introduced ForecastBench, a dynamic and continuously updated benchmark for evaluating LLM forecasting capabilities. By focusing exclusively on questions that are unresolved at the time of submission, we eliminate the risks of data leakage and ensure a robust evaluation environment. Our initial results demonstrate that while state-of-the-art LLMs exhibit promising potential, they underperform superforecasters. This performance gap highlights the challenges in leveraging current LLMs for accurate, real-time forecasting.

We produce a public leaderboard listing the real-time accuracy of top LLMs and humans as well as a standardized dataset of forecasting questions and rationales. Future work should leverage this auxiliary dataset of predictions and rationales to fine-tune models, explore new architectures, and develop adaptive systems better suited for general reasoning in dynamic, real-world environments. Ultimately, ForecastBench serves as a step toward harnessing the full potential of AI-based systems for forecasting and decision-making.

7 Reproducibility statement

One reason we've open-sourced our code (link in Appendix A) is to allow for independent verification of our results. See Appendix I for reproducing the human forecast sets, Appendix J for reproducing LLM forecast sets, and Appendix K for resolving the forecasts and creating the leaderboard.

8 Ethics statement

Human survey subjects in both the public and superforecaster surveys are made aware prior to their participation in the study via an informed consent form (approved by our IRB, number 855431) that their forecast/rationale data may be publicly released and used to train large language models or other AI systems, with said data carefully reviewed and anonymized.

We have manually reviewed text provided by human participants to ensure that no personally identifiable information is released as part of our human forecast datasets, per IRB requirements. Similar manual reviews of text data will take place as part of every future human forecasting round.

Acknowledgments

We thank Open Philanthropy for providing a grant to fund this work, allowing us to maintain the benchmark at least until mid-2027.

We thank Otto Kuusela for helpful discussions, Sam Glover and Molly Hickman for writing superforecaster prompts, and Harrison Durland and Victoria Schmidt for their research assistance.

References

Mahdi Abolghasemi, Odkhishig Ganbold, and Kristian Rotaru. Humans vs large language models: Judgmental forecasting in an era of advanced AI. *arXiv preprint arXiv:2312.06941*, 2023.

David Adam. Special report: The simulations driving the world's response to COVID-19. *Nature*, 580(7802):316–319, 2020.

Anthropic. Model card and evaluations for Claude models, 2023. https://www-cdn.anthropic.com/files/4zrzovbb/website/5c49cc247484cecf107c699baf29250302e5da70.pdf.

Anthropic. The claude 3 model family: Opus, sonnet, haiku, 2024. https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf.

Jon Scott Armstrong. *Principles of Forecasting: a Handbook for Researchers and Practitioners*. Springer, 2001.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. Qwen technical report, 2023. URL https://arxiv.org/abs/2309.16609.

Simone Balloccu, Patrícia Schmidtová, Mateusz Lango, and Ondřej Dušek. Leak, cheat, repeat: Data contamination and evaluation malpractices in closed-source LLMs. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, 2024.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. Chatbot arena: An open platform for evaluating llms by human preference, 2024. URL https://arxiv.org/abs/2403.04132.

Peter Christensen, Kenneth Gillingham, and William Nordhaus. Uncertainty in forecasts of long-run economic growth. *Proceedings of the National Academy of Sciences*, 115(21):5409–5414, 2018.

Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting. In *International Conference on Machine Learning (ICML)*, 2024.

Samuel Dooley, Gurnoor Singh Khurana, Chirag Mohapatra, Siddartha Venkat Naidu, and Colin White. ForecastPFN: Synthetically-trained zero-shot forecasting. In *Advanced in Neural Information Processing Systems (NeurIPS)*, 2023.

Yanai Elazar, Akshita Bhagia, Ian Helgi Magnusson, Abhilasha Ravichander, Dustin Schwenk, Alane Suhr, Evan Pete Walsh, Dirk Groeneveld, Luca Soldaini, Sameer Singh, et al. What's in my big data? In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024.

Epoch AI. Data on notable ai models, 2024. URL https://epoch.ai/data/notable-ai-models. Accessed: 2024-11-22.

Lukas Fluri, Daniel Paleka, and Florian Tramer. Evaluating superhuman models with consistency checks. *IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, 2024.

Gemini Team. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Mononito Goswami, Konrad Szafer, Arjun Choudhry, Yifu Cai, Shuo Li, and Artur Dubrawski. Moment: A family of open time-series foundation models. In *International Conference on Machine Learning (ICML)*, 2024.

Nate Gruver, Marc Anton Finzi, Shikai Qiu, and Andrew Gordon Wilson. Large language models are zero-shot time series forecasters. In *Advanced in Neural Information Processing Systems (NeurIPS)*, 2023.

Danny Halawi, Fred Zhang, Chen Yueh-Han, and Jacob Steinhardt. Approaching human-level forecasting with language models. *arXiv preprint arXiv:2402.18563*, 2024.

Dan Hendrycks, Nicholas Carlini, John Schulman, and Jacob Steinhardt. Unsolved problems in ML safety. *arXiv preprint arXiv:2109.13916*, 2021.

Alon Jacovi, Avi Caciularu, Omer Goldman, and Yoav Goldberg. Stop uploading test data in plain text: Practical strategies for mitigating data contamination by evaluation benchmarks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5075–5084, 2023.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023. URL https://arxiv.org/abs/2310.06825.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024a.

Minhao Jiang, Ken Ziyu Liu, Ming Zhong, Rylan Schaeffer, Siru Ouyang, Jiawei Han, and Sanmi Koyejo. Investigating data contamination for pre-training language models. *arXiv preprint arXiv:2401.06059*, 2024b.