Ming Jin, Shiyu Wang, Lintao Ma, Zhixuan Chu, James Y. Zhang, Xiaoming Shi, Pin-Yu Chen, Yuxuan Liang, Yuan-Fang Li, Shirui Pan, and Qingsong Wen. Time-LLM: Time series forecasting by reprogramming large language models. In *International Conference on Learning Representations (ICLR)*, 2024.

Woojeong Jin, Rahul Khanna, Suji Kim, Dong-Ho Lee, Fred Morstatter, Aram Galstyan, and Xiang Ren. ForecastQA: A question answering challenge for event forecasting with temporal text data. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL)*, 2021.

Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Yucheng Li, Frank Guerin, and Chenghua Lin. An open source data contamination report for large language models. *arXiv preprint arXiv:2310.17589*, 2023.

Inbal Magar and Roy Schwartz. Data contamination: From memorization to exploitation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2022.

Nestor Maslej, Loredana Fattorini, Erik Brynjolfsson, John Etchemendy, Katrina Ligett, Terah Lyons, James Manyika, Helen Ngo, Juan Carlos Niebles, Vanessa Parli, Yoav Shoham, Russell Wald, Jack Clark, and Raymond Perrault. Artificial intelligence index report 2023. *arXiv preprint arXiv:2310.03715*, 2023.

Barbara Mellers, Eric Stone, Terry Murray, Angela Minster, Nick Rohrbaugh, Michael Bishop, Eva Chen, Joshua Baker, Yuan Hou, Michael Horowitz, Lyle Ungar, and Philip Tetlock. Identifying and cultivating superforecasters as a method of improving probabilistic predictions. *Perspectives on Psychological Science*, 2015.

Metaculus. Wisdom of the crowd vs. the best of the best of the best, 2023. URL https://www.metaculus.com/notebooks/15760/wisdom-of-the-crowd-vs-the-best-of-the-best-of-the-best/.

Yuqi Nie, Nam H Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In *International Conference on Learning Representations (ICLR)*, 2023.

Maxwell Nye, Anders Johan Andreassen, Guy Gur-Ari, Henryk Michalewski, Jacob Austin, David Bieber, David Dohan, Aitor Lewkowycz, Maarten Bosma, David Luan, Charles Sutton, and Augustus Odena. Show your work: Scratchpads for intermediate computation with language models. *arXiv preprint arXiv:2112.00114*, 2021.

OpenAI. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

David Owen. How predictable is language model benchmark performance? *arXiv preprint arXiv:2401.04757*, 2024.

Long Phan, Andrew Zeng, Mantas Mazeika, Adam Khoja, and Dan Hendrycks. Superhuman automated forecasting. https://www.safe.ai/blog/forecasting, 2024.

Sarah Pratt, Seth Blumberg, Pietro Kreitlon Carolino, and Meredith Ringel Morris. Can language models use forecasting strategies? *arXiv preprint arXiv:2406.04446*, 2024.

Clionadh Raleigh, Roudabeh Kishi, and Andrew Linke. Political instability patterns are obscured by conflict dataset scope conditions, sources, and coding choices. *Humanities and Social Sciences Communications*, 10:74, 2023.

Kashif Rasul, Arjun Ashok, Andrew Robert Williams, Arian Khorasani, George Adamopoulos, Rishika Bhagwatkar, Marin Biloš, Hena Ghonia, Nadhir Vincent Hassen, Anderson Schneider, Sahil Garg, Alexandre Drouin, Nicolas Chapados, Yuriy Nevmyvaka, and Irina Rish. Lag-Llama: Towards foundation models for time series forecasting. *arXiv preprint arXiv:2310.08278*, 2023.

Manley Roberts, Himanshu Thakur, Christine Herlihy, Colin White, and Samuel Dooley. Data contamination through the lens of time. *arXiv preprint arXiv:2310.10628*, 2023.

Oscar Sainz, Jon Campos, Iker García-Ferrero, Julen Etxaniz, Oier Lopez de Lacalle, and Eneko Agirre. NLP evaluation in trouble: On the need to measure llm data contamination for each benchmark. In *Findings of the Association for Computational Linguistics (Findings of EMNLP)*, 2023.

Ville A. Satopää, Jonathan Baron, Dean P. Foster, Barbara A. Mellers, Philip E. Tetlock, and Lyle H. Ungar. Combining multiple probability predictions using a simple logit model. *International Journal of Forecasting*, 30(2):344–356, 2014.

Philipp Schoenegger and Peter S Park. Large language model prediction capabilities: Evidence from a real-world forecasting tournament. *arXiv preprint arXiv:2310.13014*, 2023.

Philipp Schoenegger, Peter S Park, Ezra Karger, and Philip E Tetlock. AI-augmented predictions: LLM assistants improve human forecasting accuracy. *arXiv preprint arXiv:2402.07862*, 2024a.

Philipp Schoenegger, Indre Tuminauskaite, Peter S Park, and Philip E Tetlock. Wisdom of the silicon crowd: LLM ensemble prediction capabilities match human crowd accuracy. *arXiv preprint arXiv:2402.19379*, 2024b.

Philip E. Tetlock and Dan Gardner. *Superforecasting: The Art and Science of Prediction*. Crown, 2015.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Gerald Woo, Chenghao Liu, Akshat Kumar, Caiming Xiong, Silvio Savarese, and Doyen Sahoo. Unified training of universal time series forecasting transformers. In *International Conference on Machine Learning (ICML)*, 2024.

Cheng Xu, Shuhao Guan, Derek Greene, M Kechadi, et al. Benchmark data contamination of large language models: A survey. *arXiv preprint arXiv:2406.04244*, 2024a.

Ruijie Xu, Zengzhi Wang, Run-Ze Fan, and Pengfei Liu. Benchmarking benchmark leakage in large language models. *arXiv preprint arXiv:2404.18824*, 2024b.

Qi Yan, Raihan Seraj, Jiawei He, Lili Meng, and Tristan Sylvain. Autocast++: Enhancing world event prediction with zero-shot ranking-based context retrieval. In *International Conference on Learning Representations (ICLR)*, 2024.

Hugh Zhang, Jeff Da, Dean Lee, Vaughn Robinson, Catherine Wu, Will Song, Tiffany Zhao, Pranav Raja, Dylan Slack, Qin Lyu, Sean Hendryx, Russell Kaplan, Michele Lunati, and Summer Yue. A careful examination of large language model performance on grade school arithmetic. *arXiv preprint arXiv:2405.00332*, 2024.

Andy Zou, Tristan Xiao, Ryan Jia, Joe Kwon, Mantas Mazeika, Richard Li, Dawn Song, Jacob Steinhardt, Owain Evans, and Dan Hendrycks. Forecasting future world events with neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

A Licensing and maintenance

Hosting The latest leaderboards, updated nightly, are available on www.forecastbench.org. Documentation is available on the repository wiki: github.com/forecastingresearch/forecastbench/wiki.

Datasets Our datasets, distributed under the CC BY-SA 4.0 license, are available on www.forecastbench.org/datasets.html. Historical updates to the resolution datasets and leaderboards are available on github.com/forecastingresearch/forecastbench-datasets. Bi-weekly question sets are also released via this repository.

Codebase The code underlying our automated system runs on Google Cloud Platform and is available at github.com/forecastingresearch/forecastbench under the MIT license.

Participating Bi-weekly forecasting rounds are open to LLM teams. Instructions for participating are can be found on the wiki.

Maintenance & long-term preservation We ensure the long-term availability and maintenance of the benchmark as it is funded by Open Philanthropy until mid-2027. If no further funding is provided beyond that point, datasets will continue to be made available on GitHub.

B Datasets

We intend our datasets to be used for training general LLMs, fine-tuning forecasting LLMs, and for any applicable research purposes. No restrictions are placed on who may use our datasets, nor to what end.

Availability www.forecastbench.org/datasets.html

B.1 Question and resolution sets

Every question set will be published. Their resolutions will also be published such that there's a complete training set when combined with the forecast sets outlined in Section B.2. The data dictionary for the question set is outlined in Table 4 and Table 5. The data dictionary for the resolution set is outlined in Table 6 and Table 7.

Data format The question and resolution datasets are released as JSON (`.json`) files.

Ethical and responsible use There are no restrictions on use of the question and resolution datasets.

Data collection Our question and resolution datasets have been pulled, and are updated, from various, public-facing sources. From those sources where the terms of use/service prohibit the redistribution of their information (currently, Manifold Markets and Metaculus), we have obtained explicit permission to do so. Before we add new sources to our growing dataset, we will ensure the ability to distribute questions and resolutions publicly. Data sources in our question bank can be found in Table 14.

Table 4: Question set data dictionary.

Field	Description	Required	Data Type
`forecast_due_date`	Date in ISO format. e.g. `"2024-07-21"`	✓	string
`question_set`	The name of the file that contains the question set. e.g. `"2024-07-21-llm.json"`	✓	string
`questions`	A list of questions to forecast on, as defined in Table 5.	✓	array<object>