

**No fine-tuning and no retrieval.** We evaluate our optimal system without any news retrieval and using the base GPT-4-1106-Preview model. The ablated system attains a Brier score of .206.

Recall that in our baseline evaluation (Section 3.4), the lowest Brier score attained by any model is .208. Our ablated system essentially deteriorates to this baseline level. Indeed, without any fine-tuning or retrieval, the only expected advantage of our system over the baseline evaluation setup is its reasoning prompt, found through searching a set of candidate prompts (Section 5). The experiment suggests that this gives fairly a minor improvement.

## 8 Conclusion

Our work presents the first ML system that can forecast at near human levels. We develop a novel retrieval mechanism that uses a LM to determine which information to source and how to evaluate its relevance. We also give a self-supervised fine-tuning method to generate reasonings with accurate predictions.

To facilitate further research, we release our dataset: the largest and most recent forecasting dataset compiled from 5 real-world forecasting competitions. We discuss a few opportunities to improve these systems further.

**Iterative self-supervision.** With a larger training corpus, our self-supervised fine-tuning approach can be used for iterative self-improvement. Specifically, after fine-tuning a model on its previous optimal predictions and reasonings, we can generate more fine-tuning data by using the same model again, which can be repeated until training data is exhausted.

**Data.** While our forecasting benchmark is a good initial corpus to train a system, we believe that it is possible to use LMs with later training cut-offs to teach an earlier LM. This could be done by using later LMs to generate questions it knows the answer to but an earlier LM does not (postdiction). In addition, while we source questions from forecasting platforms, it is possible to collect historical data in the wild and re-formulate them as forecasting questions, leading to a larger training set.

**Domain-adaptive training.** In Section B.3, we observe that in the baseline evaluations, the Brier scores across categories are correlated with models’ pre-training knowledge. This suggests that we may be able to specialize models to areas of particular interests by fine-tuning them on domain knowledge.

**LMs get better at forecasting naturally.** We observe that as LMs improve, they naturally also become better at forecasting. In particular, in Section 3.4, we see that newer generations of models forecast better than older ones. For example, GPT-4-1106, released in 2023, outperforms GPT-4-0613, released in 2021, by .02 with respect to the Brier score. If we were to have fine-tuned the more recent model, we would expect better performance.

At a high level, our results suggest that in the near future, LM-based systems may be able to generate accurate forecasts at the level of competitive human forecasters. We hope that our work paves the way for automated, scalable forecasting that can help to inform institutional decision making.

## Acknowledgments

We thank Jean-Stanislas Denain, Erik Jones, Ezra Karger, Jacob Pfau and Ruiqi Zhong for helpful discussions, and Jean-Stanislas Denain, Owain Evans, Dan Hendrycks, Horace He and Andy Zou for comments and feedbacks on an early draft of the paper. DH was supported by an award from the C3.ai Digital Transformation Institute. FZ was supported by NSF award CCF-2311648. JS was supported by the National Science Foundation SaTC CORE Award No. 1804794 and the Simons Foundation.

## References

- Abolghasemi, M., Ganbold, O., and Rotaru, K. (2023). Humans vs large language models: Judgmental forecasting in an era of advanced AI. *arXiv preprint arXiv:2312.06941*.
- Adam, D. (2020). Special report: The simulations driving the world’s response to COVID-19. *Nature*, 580(7802):316–319.

- Anthropic (2023). Model card and evaluations for Claude models. <https://www-cdn.anthropic.com/files/4zrzovbb/website/5c49cc247484cecf107c699baf29250302e5da70.pdf>.
- Armstrong, J. S. (2001). *Principles of Forecasting: a Handbook for Researchers and Practitioners*. Springer.
- Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3.
- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language models are few-shot learners. In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Chen, X., Aksitov, R., Alon, U., Ren, J., Xiao, K., Yin, P., Prakash, S., Sutton, C., Wang, X., and Zhou, D. (2023). Universal self-consistency for large language model generation. *arXiv preprint arXiv:2311.17311*.
- Das, A., Kong, W., Sen, R., and Zhou, Y. (2024). A decoder-only foundation model for time-series forecasting. *arXiv preprint arXiv:2310.10688*.
- Deng, Y., Zhang, W., Chen, Z., and Gu, Q. (2023). Rephrase and respond: Let large language models ask better questions for themselves. *arXiv preprint arXiv:2311.04205*.
- Dooley, S., Khurana, G. S., Mohapatra, C., Naidu, S. V., and White, C. (2023). ForecastPFN: Synthetically-trained zero-shot forecasting. In *Advanced in Neural Information Processing Systems (NeurIPS)*.
- Gawlikowski, J., Tassi, C. R. N., Ali, M., Lee, J., Humt, M., Feng, J., Kruspe, A., Triebel, R., Jung, P., Roscher, R., Shahzad, M., Yang, W., Bamler, R., and Zhu, X. X. (2021). A survey of uncertainty in deep neural networks. *arXiv preprint arXiv:2107.03342*.
- Gemini Team (2023). Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.
- Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378.
- Gruver, N., Finzi, M. A., Qiu, S., and Wilson, A. G. (2023). Large language models are zero-shot time series forecasters. In *Advanced in Neural Information Processing Systems (NeurIPS)*.
- Hanson, R. (2007). Logarithmic markets coring rules for modular combinatorial information aggregation. *The Journal of Prediction Markets*, 1(1):3–15.
- Hendrycks, D., Carlini, N., Schulman, J., and Steinhardt, J. (2021). Unsolved problems in ML safety. *arXiv preprint arXiv:2109.13916*.
- Izacard, G. and Grave, É. (2021). Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*.
- Jiang, A. Q., Sablayrolles, A., Roux, A., Mensch, A., Savary, B., Bamford, C., Chaplot, D. S., de las Casas, D., Hanna, E. B., Bressand, F., Lengyel, G., Bour, G., Lample, G., Lavaud, L. R., Saulnier, L., Lachaux, M.-A., Stock, P., Subramanian, S., Yang, S., Antoniak, S., Scao, T. L., Gervet, T., Lavril, T., Wang, T., Lacroix, T., and Sayed, W. E. (2024). Mixtral of experts. *arXiv preprint arXiv:2401.04088*.
- Jin, M., Wang, S., Ma, L., Chu, Z., Zhang, J. Y., Shi, X., Chen, P.-Y., Liang, Y., Li, Y.-F., Pan, S., and Wen, Q. (2024). Time-LLM: Time series forecasting by reprogramming large language models. In *International Conference on Learning Representations (ICLR)*.
- Jin, W., Khanna, R., Kim, S., Lee, D.-H., Morstatter, F., Galstyan, A., and Ren, X. (2021). 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)*.

Lewis, P. S. H., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., and Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Manifold (2022). Maniswap. <https://manifoldmarkets.notion.site/manifoldmarkets/Maniswap-ce406e1e897d417cbd491071ea8a0c39>.

Metaculus (2023). Wisdom of the crowd vs. the best of the best of the best. <https://www.metaculus.com/notebooks/15760/wisdom-of-the-crowd-vs-the-best-of-the-best-of-the-best>.

Min, S., Zhong, V., Zettlemoyer, L., and Hajishirzi, H. (2019). Multi-hop reading comprehension through question decomposition and rescoring. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Nakano, R., Hilton, J., Balaji, S., Wu, J., Ouyang, L., Kim, C., Hesse, C., Jain, S., Kosaraju, V., Saunders, W., Jiang, X., Cobbe, K., Eloundou, T., Krueger, G., Button, K., Knight, M., Chess, B., and Schulman, J. (2021). WebGPT: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*.

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

Nye, M., Andreassen, A. J., Gur-Ari, G., Michalewski, H., Austin, J., Bieber, D., Dohan, D., Lewkowycz, A., Bosma, M., Luan, D., Sutton, C., and Odena, A. (2021). Show your work: Scratchpads for intermediate computation with language models. *arXiv preprint arXiv:2112.00114*.

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

Polymarket (2023). Polymarket/poly-market-maker: Market Maker Keeper for the polymarket CLOB. <https://github.com/Polymarket/poly-market-maker>.

Rasul, K., Ashok, A., Williams, A. R., Khorasani, A., Adamopoulos, G., Bhagwatkar, R., Biloš, M., Ghonia, H., Hassen, N. V., Schneider, A., Garg, S., Drouin, A., Chapados, N., Nevmyvaka, Y., and Rish, I. (2023). Lag-Llama: Towards foundation models for time series forecasting. *arXiv preprint arXiv:2310.08278*.

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

Shuster, K., Poff, S., Chen, M., Kiela, D., and Weston, J. (2021). Retrieval augmentation reduces hallucination in conversation. In *Findings of the Association for Computational Linguistics (Findings of EMNLP)*.

Tetlock, P. E. and Gardner, D. (2015). *Superforecasting: The Art and Science of Prediction*. Crown.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., Fuller, B., Gao, C., Goswami, V., Goyal, N., Hartshorn, A., Hosseini, S., Hou, R., Inan, H., Kardas, M., Kerkez, V., Khabsa, M., Kloumann, I., Korenev, A., Koura, P. S., Lachaux, M.-A., Lavril, T., Lee, J., Liskovich, D., Lu, Y., Mao, Y., Martinet, X., Mihaylov, T., Mishra, P., Molybog, I., Nie, Y., Poulton, A., Reizenstein, J., Rungta, R., Saladi, K., Schelten, A., Silva, R., Smith, E. M., Subramanian, R., Tan, X. E., Tang, B., Taylor, R., Williams, A., Kuan, J. X., Xu, P., Yan, Z., Zarov, I., Zhang, Y., Fan, A., Kambadur, M., Narang, S., Rodriguez, A., Stojnic, R., Edunov, S., and Scialom, T. (2023). Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Wang, C. (2023). Calibration in deep learning: A survey of the state-of-the-art. *arXiv preprint arXiv:2308.01222*.

Woo, G., Liu, C., Kumar, A., Xiong, C., Savarese, S., and Sahoo, D. (2024). Unified training of universal time series forecasting transformers. *arXiv preprint arXiv:2402.02592*.