crowds' mechanism (Koriat 2012). In our aggregation results, we also find that only three of the twelve models outperform the model median, which is also in line with standard accounts of wisdom of crowds. This overall suggests that the 'wisdom of the crowd' effect—in addition to applying to the human context—also applies to the silicon context. The literature on the size of the crowd needed to produce reliable 'wisdom of the crowd' effects is not very well established, though a central finding is that increasing crowd size does lead to better performance (Walter, Kölle, and Collmar 2022). As such, a natural next step in this line of research is to expand the set of models queried from the twelve we used to a substantially higher number.

However, there also do remain substantial areas of improvement for machine predictions in probabilistic forecasting. Most directly, both the aggregate and the individual models were badly calibrated, with most models showing overconfidence, i.e., they assign higher probabilities to outcomes than is warranted by the empirical facts. Improving calibration is central to providing reliable predictions over the long run (Buizza 2018), and our results of acquiescence bias suggest that this may be an actionable area for future work. Additionally, as the distance between the end of the training data and the forecasted period grows, forecasts may become less accurate as necessary background knowledge is no longer readily available to the model. Moreover, our study could draw on a well-curated set of questions from the forecasting tournament. Applications in contexts where neutral background information and question details are not available may further reduce performance.

Our results from Study 2 contribute to the literature on human-AI interactions (Kim et al. 2024; Yang et al. 2024). While previous work in the context of forecasting has looked at how LLMs can augment humans in improving prediction accuracy (Schoenegger et al. 2024), this paper provides evidence for the reverse. Specifically, our results show that machine predictions can be improved substantially by the provision of human cognition output drawn from forecasting tournaments. This finding suggests at first glance that LLM reasoning is already advanced enough to properly exploit the informational value provided by human cognition output. However, our exploratory analyses find that this process is substantially less effective than simply averaging the two estimates, suggesting that aggregation methods based on the reasoning capabilities of frontier models (in this case, GPT-4 and Claude 2) still underperform simple aggregations.

On the other hand, our findings that both frontier models (GPT-4 and Claude 2) respond as expected in their forecast updates—reducing their uncertainty when the human estimate lies within their prediction intervals, and updating in relation to the distance between their own point estimate and the human forecasts—match past theory and results pertaining to human forecasters (Atanasov et al. 2020). This overall suggests that the ability of these models to reason and act as expected—by past theory and results pertaining to human forecasters—depend on the type of task and benchmark applied. While this is not a massively strict test of their reasoning abilities—as alternative explanations of model behaviour being explained by simple expectation fulfilling remain—it does provide some evidence in favour of it.

Importantly, both studies reported in this paper test LLM capabilities in a context where it is not possible that any of the answers used to resolve the questions were part of the training data, as we queried the models in real-time alongside the human tournament. Because all question answers were unknown at the time of data collection—even to the study authors—this provides an ideal evaluation criterion for LLM capabilities: one at which our LLM ensemble approach excelled. Our findings provide evidence of LLMs' advanced reasoning capabilities, and does so in a robust way such that many of the challenges that may be raised with respect to traditional benchmarks do not apply.

In conclusion, the present paper is among the first to show that current LLMs are able to provide a human-crowd-competitive level of accurate forecasting about future real-world events. In order to do so, it is sufficient to apply the simple, practically applicable method of forecast aggregation: manifesting as the LLM ensemble approach in the so-called silicon setting. This replicates the human forecasting tournament's 'wisdom of the crowd' effect for LLMs: a phenomenon we call the 'wisdom of the silicon crowd.' Our finding opens up a number of areas for further research as well as practical applications, since the LLM ensemble approach is substantially cheaper and faster than data collection from human forecasters. Future research can aim to combine the ensemble approach with model and scaffolding progress, which may potentially result in even stronger capability gains in the domain of forecasting.

Acknowledgements

We are grateful to Lawrence Phillips and Peter Mühelbacher for helping us discover and correct a coding error in the non-preregistered equivalence test pertaining to the second null hypothesis of Study 1.

References

Abdurahman, Suhaib et al. (2023). "Perils and Opportunities in Using Large Language Models in Psychological Research". In: *PsyArXiv*. URL: https://osf.io/preprints/psyarxiv/d695y.

Abolghasemi, Mahdi, Odkhishig Ganbold, and Kristian Rotaru (2023). "Humans vs Large Language Models: Judgmental Forecasting in an Era of Advanced AI". In: *arXiv preprint arXiv:2312.06941*. URL: https://arxiv.org/abs/2312.06941.

Acemoğlu, Daron (2023). "Harms of AI". In: *The Oxford Handbook of AI Governance*. Oxford University Press. ISBN: 9780197579329. DOI: 10.1093/oxfordhb/9780197579329.013.65. URL: https://doi.org/10.1093/oxfordhb/9780197579329.013.65.

Alzahrani, Norah et al. (2024). *When Benchmarks are Targets: Revealing the Sensitivity of Large Language Model Leaderboards*. arXiv: 2402.01781 [cs.CL].

Anthropic (2023). *Model Card and Evaluations for Claude Models*. URL: https://www-cdn.anthropic.com/files/4zrzovbb/website/bd2a28d2535bfb0494cc8e2a3bf135d2e7523226.pdf.

Arora, Sanjeev and Anirudh Goyal (2023). "A Theory for Emergence of Complex Skills in Language Models". In: *arXiv preprint arXiv:2307.15936*.

Atanasov, Pavel et al. (2020). "Small steps to accuracy: Incremental belief updaters are better forecasters". In: *Proceedings of the 21st ACM Conference on Economics and Computation*, pp. 873–874.

Atari, Mohammad et al. (2023). "Which humans?" In: *PsyArXiv*. URL: https://osf.io/preprints/psyarxiv/5b26t.

Baron, Jonathan et al. (2014). "Two reasons to make aggregated probability forecasts more extreme". In: *Decision Analysis* 11.2, pp. 133–145.

Bassamboo, Achal, Ruomeng Cui, and Antonio Moreno (2018). *Wisdom of crowds: Forecasting using prediction markets*. Tech. rep. Kellogg School of Management, Northwestern University.

Beaulieu-Jones, Brendin R et al. (2024). "Evaluating capabilities of large language models: Performance of GPT-4 on surgical knowledge assessments". In: *Surgery*.

Bender, Emily M. et al. (2021). "On the Dangers of Stochastic Parrots: Can Language Models be too Big?" In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '21. Virtual Event, Canada: Association for Computing Machinery, pp. 610–623. ISBN: 9781450383097. DOI: 10.1145/3442188.3445922.

Biderman, Stella et al. (2023). *Emergent and Predictable Memorization in Large Language Models*. arXiv: 2304.11158 [cs.CL].

Brier, Glenn W. (1950). "Verification of forecasts expressed in terms of probability". In: *Monthly weather review* 78.1, pp. 1–3.

Bubeck, Sébastien et al. (2023). *Sparks of Artificial General Intelligence: Early Experiments with GPT-4*. arXiv: 2303.12712 [cs.CL].

Budescu and Chen (2015). "Identifying expertise to extract the wisdom of crowds". In: *Management science* 61.2, pp. 267–280.

Budescu, David V. (2006). "Confidence in aggregation of opinions from multiple sources". In: *Information Sampling and Adaptive Cognition*. Ed. by Klaus Fiedler and Peter Juslin. Cambridge, UK: Cambridge University Press, pp. 327–352.

Buizza, Roberto (2018). "Ensemble forecasting and the need for calibration". In: *Statistical postprocessing of ensemble forecasts*. Elsevier, pp. 15–48.

Carlini, Nicholas et al. (2023). "Quantifying Memorization Across Neural Language Models". In: *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net. URL: https://openreview.net/pdf?id=TatRHT%5C_1cK.

Cholakov, Radostin and Todor Kolev (2021). "Transformers predicting the future. Applying attention in next-frame and time series forecasting". In: *arXiv preprint arXiv:2108.08224*.

Cohen, Jacob (2013). *Statistical power analysis for the behavioral sciences*. Academic press.

Costello, Shane and John Roodenburg (2015). "Acquiescence response bias—Yeasaying and higher education". In: *The Educational and Developmental Psychologist* 32.2, pp. 105–119.

Da, Zhi and Xing Huang (2020). "Harnessing the wisdom of crowds". In: *Management Science* 66.5, pp. 1847–1867.
Davis-Stober, Clintin P. et al. (2014). "When is a crowd wise?" In: *Decision* 1.2, p. 79.
Feng, Tony Haoran et al. (2024). "More Than Meets the AI: Evaluating the performance of GPT-4 on Computer Graphics assessment questions". In: *Proceedings of the 26th Australasian Computing Education Conference*, pp. 182–191.
Fraiwan, Mohammad and Natheer Khasawneh (2023). *A Review of ChatGPT Applications in Education, Marketing, Software Engineering, and Healthcare: Benefits, Drawbacks, and Research Directions*. arXiv: 2305.00237 [cs.CY].
Gemini Team et al. (2023). *Gemini: A Family of Highly Capable Multimodal Models*. arXiv: 2312.11805 [cs.CL].
Ghirardato, Paolo (2002). "Revisiting Savage in a conditional world". In: *Economic Theory* 20, pp. 83–92.
Gneiting, Tilmann and Adrian E Raftery (2007). "Strictly proper scoring rules, prediction, and estimation". In: *Journal of the American statistical Association* 102.477, pp. 359–378.
Grove, Nathaniel P. and Stacey Lowery Bretz (2012). "A Continuum of Learning: From Rote Memorization to Meaningful Learning in Organic Chemistry". In: *Chemistry Education Research and Practice* 13.3, pp. 201–208.
Gruver, Nate et al. (2024). "Large language models are zero-shot time series forecasters". In: *Advances in Neural Information Processing Systems* 36.
Halawi, Danny et al. (2024). *Approaching Human-Level Forecasting with Language Models*. arXiv: 2402.18563 [cs.LG].
Heiding, Fredrik et al. (2023). "Devising and detecting phishing: Large language models vs. smaller human models". In: *arXiv preprint arXiv:2308.12287*.
Himmelstein, Michael, David V. Budescu, and Yoonjin Han (2023). "The wisdom of timely crowds". In: *Judgment in predictive analytics*. Springer International Publishing, pp. 215–242.
Himmelstein, Michael, David V. Budescu, and Elizabeth H. Ho (2023). "The wisdom of many in few: Finding individuals who are as wise as the crowd". In: *Journal of Experimental Psychology: General*.
Hinz, Andreas et al. (2007). "The acquiescence effect in responding to a questionnaire". In: *GMS Psycho-Social Medicine* 4.
Jiao, Wenxiang et al. (2023). *Is ChatGPT a Good Translator? Yes with GPT-4 as the Engine*. arXiv: 2301.08745 [cs.CL].
Jin, Ming et al. (2023). "Time-llm: Time series forecasting by reprogramming large language models". In: *arXiv preprint arXiv:2310.01728*.
Katz, Daniel Martin et al. (2023). "GPT-4 Passes the Bar Exam". In: *SSRN*.
Kim, Su Hwan et al. (2024). "Human-AI Collaboration in Large Language Model-Assisted Brain MRI Differential Diagnosis: A Usability Study". In: *medRxiv*, pp. 2024–02.
Kistowski, Jóakim v. et al. (2015). "How to build a benchmark". In: *Proceedings of the 6th ACM/SPEC international conference on performance engineering*, pp. 333–336.
Koriat, Asher (2012). "When are two heads better than one and why?" In: *Science* 336.6079, pp. 360–362.
Laskar, Md Tahmid Rahman et al. (2023). *A Systematic Study and Comprehensive Evaluation of ChatGPT on Benchmark Datasets*. arXiv: 2305.18486 [cs.CL].
Lichtendahl Jr, Kenneth C., Yael Grushka-Cockayne, and Phillip E. Pfeifer (2013). "The wisdom of competitive crowds". In: *Operations Research* 61.6, pp. 1383–1398.
Magar, Inbal and Roy Schwartz (May 2022). "Data Contamination: From Memorization to Exploitation". In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Dublin, Ireland: Association for Computational Linguistics, pp. 157–165. DOI: 10.18653/v1/2022.acl-short.18. URL: https://aclanthology.org/2022.acl-short.18.
Mandel, David R. and Alan Barnes (2014). "Accuracy of forecasts in strategic intelligence". In: *Proceedings of the National Academy of Sciences* 111.30, pp. 10984–10989.
Metaculus (2024). *Metaculus*. https://www.metaculus.com/home/. Accessed: 2024-02-21.
Naveed, Humza et al. (2023). *A Comprehensive Overview of Large Language Models*. arXiv: 2307.06435 [cs.CL].
Nori, Harsha et al. (2023). *Capabilities of GPT-4 on Medical Challenge Problems*. arXiv: 2303.13375 [cs.CL].
OpenAI et al. (2023). *GPT-4 Technical Report*. arXiv: 2303.08774 [cs.CL].