Table 6: LLM Augmentation Difficulty Effects: Mixed Effects Model Results

Variable	Coefficient	Std. Error	z-value	p-value
Intercept	0.66	0.03	23.32	< 0.001
Treatment	-0.25	0.04	-6.75	< 0.001
Treatment (Noise)	-0.23	0.04	-6.03	< 0.001
Difficulty	0.69	0.05	14.73	< 0.001
Treatment · Difficulty	0.11	0.06	1.83	0.067
Treatment (Noise) · Difficulty	-0.04	0.07	-0.68	0.500
Observations		5946		
No. Groups		991		
Log-Likelihood		-7898.97		

Notes. Group Var = 0.015. Scale = 0.8168. Random intercepts applied at participant level.

rank them accordingly. The adjusted p-values are computed using the Benjamini-Hochberg procedure, which calculates the adjusted p-value for the i-th hypothesis as

$$\min\left\{1, \frac{p_i \cdot m}{\text{rank}_i}\right\}$$

where p_i is the i-th p-value in the sorted list, m is the total number of hypotheses tested, and $rank_i$ is the rank of the i-th p-value in the sorted list. The adjusted p-values are 0.005, 0.985, 0.985, 0.163, and 0.833, showing that our results are robust to this adjustment, with the p-value pertaining to our first hypothesis remaining significant at p=0.005, with all others remaining non-significant.

4 Discussion

Our investigation of an LLM forecasting augmentation as a tool for judgemental forecasts offers a number of results. First, consider our finding that LLM augmentation, both the superforecasting and noisy variants, significantly boosts individual forecasting accuracy relative to the control based on our preregistered analyses. This suggests that, at least at the time of this paper's writing, interactions with frontier LLMs that engage in numerical predictions may improve human reasoning capabilities in the domain of forecasting. Moreover, with LLM system's prediction performance increasing (Halawi et al. 2024; Schoenegger et al. 2024b), this synergistic effect is likely to improve going forward. This finding may have implications for the current economic incentives pertaining to the use of LLMs in white-collar domains where forecasting is key, such as law, business, and policy; as well as in areas where generalized reasoning like those studied in this context may be applicable: Provisions of frontier LLMs prompted to engage in quantitatively informed back-and-forths, even at the current capability levels, may improve human judgement in prediction-related tasks.

However, this does not mean that this pattern of human-in-the-loop systems will continue in the face of potentially more capable AI systems released in the future. To illustrate, consider that in chess, human performance was much stronger than AI performance before 1994, could serve as the key difference as the human-in-the-loop in the ten years between 1994 and 2004, and was much weaker than AI performance after 2004 (Kasparov 2010). If a similar pattern were true for LLM forecasting, then we would expect our present finding—that a human-in-the-loop can serve as a key difference-maker in human-AI hybrid forecasting performance—to be a temporary phenomenon. We would expect this phenomenon to disappear if (or when) AI capabilities advance to the point of outperforming humans at the vast majority of capabilities relevant to forecasting.

We also found that both the superforecasting and the noisy variants of LLM augmentation yield similar levels of forecasting accuracy increase compared to the control, with no statistically significant difference between them. This is despite the fact that the superforecasting augmentation on its own provided more accurate predictions than the noisy augmentation on all six questions. Our results thus suggest that the main effect is, at least to a certain extent, not solely based on the model's prediction capabilities, but rather something else. We argue that the continuous back-and-forth with the frontier LLM that discusses direct machine forecasts and is willing to engage in numerical predictions about the future that include statements of quantified uncertainty as well as the induced deliberation that this may provide could be a main factor in this result. Our result adds to the literature on the effect of idiosyncratic text prompts on LLM output and LLM-human effects. Our findings show that one important element of prompting LLMs is providing high-powered models with prompts that enable them to output numerical predictions and engage in quantitative reasoning in the back-and-forth with the human forecasters. The control LLM was much smaller and not able to do these interactions, making our

result a combination of advanced model reasoning capabilities and willingness/ability to engage in quantitative reasoning about the future.

However, our exploratory analyses also found that this pattern of results changes if we remove one outlier question, Question 3. Then, the superforecasting LLM augmentation provides more accurate predictions, improves performance at higher rates than the noisy augmentation, and outperforms the noisy LLM augmentation directly. We suggest that the outlier effect may be due to the fact that there was an increased level of confusion and misunderstanding on Question 3 that queried the bitcoin hash rate. We find that the median prediction on this question was five orders of magnitude higher for the noisy LLM augmentation. Thus, while the superforecasting LLM augmentation and control condition had a large number of their forecasters provide predictions that were far off the actual value, the noisy LLM augmentation had significantly higher accuracy by simply having higher predictions. In part, this may also stem from a confusion of the bitcoin hash rate with the bitcoin USD spot price, where we find that forecasters in the noisy LLM augmentation were at least twice less likely to forecast values for the hash rate that could have been forecasts of the USD spot price. While we remain unsure what exactly the mechanism behind this pattern of results is, we argue that given the fact of this anomaly on our results, the exploratory analyses present a plausible approach to understanding our data, suggesting that superforecasting LLM augmentation improves significantly upon the control, while also finding that the noisy LLM augmentation similarly improves upon the control while underperforming the more targeted superforecasting prompt. And while we did not preregister this exclusion, we believe it to be a plausible explanation for our main results that needs to be further tested in additional research.

Our next research question investigated the impact of LLM augmentation on low-skilled forecasters versus high-skilled forecasters. Past research on LLM augmentation generally suggests that provision of AI support disproportionately bolsters the performance of low-performing workers among consultants (Dell'Acqua et al. 2023), call-center agents (Brynjolfsson, Li, and Raymond 2023), creative writers (Doshi and Hauser 2023), office workers (Noy and Zhang 2023), law school students (Choi and Schwarcz 2024), and programmers (Peng et al. 2023). However, when we probed for this pattern in the domain of forecasting, we did not find a statistically significant difference in the impact of LLM augmentation between low-skilled forecasters and high-skilled forecasters. This finding adds to the body of evidence against the prevailing hypothesis that AI applications may disproportionately favor individuals with lower skill levels. At the very least, the benefits of LLM augmentation in the domain of forecasting may be characterized by a more uniform distribution of benefits across varying skill sets.

We also investigated the impact of LLM augmentation on the accuracy of aggregated forecasts. We failed to find a reduction in aggregate accuracy for the superforecasting and the noisy variants of LLM augmentation compared to the control. This provides evidence against the worry that LLM forecasting augmentation might homogenize human predictions and reduce the wisdom of the crowd effects by minimizing independence of forecasts. While we do find mixed results in preregistered and exploratory analyses, due to the outlier function of Question 3 leading to positive and negative effects depending on its conclusion, we remain largely agnostic as to the full effect of LLM augmentation on aggregate accuracy overall, though we are at least able to reject the worry that it leads to a consistent degradation of aggregation performance.

Finally, we found the effect of LLM augmentation on human forecasts does not significantly differ between easy and hard forecasting questions. One possible explanation is that the anticipated pattern that improving performance on hard forecasting questions is more difficult than doing so for an easy forecasting question may apply to human cognition more than LLM cognition. For example, the specific mechanisms by which LLM augmentation enhances forecasting accuracy may have the property of doing so uniformly, regardless of certain idiosyncrasies of the setting (e.g., difficulty of forecasting question) in question. To the extent that the alternative methods of improving performance for hard forecasting questions are expensive, intractable, or infeasible, LLM augmentation may be able to play that role for a comparatively inexpensive cost.

Our results demonstrate the potential of LLMs to augment human decision-making through interactive collaboration. The significant accuracy improvements we observed highlight the importance of designing effective human-AI interaction modes, a key challenge identified by Steyvers and Kumar (2023). Our approach, which allowed for back-and-forth engagement between users and the LLM, exemplifies how interactive prompting can enhance human performance in complex tasks like forecasting, aligning with the interaction modes described by Gao et al. (2024). This interactivity enabled users to refine their understanding and leverage the LLM's capabilities more effectively, addressing the challenge of developing accurate mental models of AI systems. Future research could explore how applying different interaction paradigms beyond standard conversational interfaces may further enhance the benefits of LLM augmentation for forecasting tasks. Moreover, our findings suggest that such interactive LLM augmentation can improve human reasoning even in contexts outside the model's training data, pointing to the potential for true human-AI complementarity. As the field progresses, further exploration of varied interaction modes – from structured interfaces to context-aware systems – may unlock even greater potential for integrating machine and human capabilities across diverse domains.

5 Limitations

There are a number of limitations to the design and results presented in this paper. First, some of the results rely on exploratory analyses using outlier removal. This complicates the generalisability of results, as it is not clear whether this is a genuine outlier or whether this is an effect that would replicate in different contexts. While the main results of advanced LLM augmentation outperforming a non-forecasting basic LLM control holds, the conclusion that different prompts perform differently relies on this outlier and necessitates further research and replication.

Second, there are concerns that online samples like the one used in this study reduce the generalisability of results, as participants might be systematically biased. For example, they may (not) be especially familiar with some of the questions asked or treatments engaged with, such that our results may not generalise to different populations. While some concerns with online samples remain, we argue that recent work has shown Prolific participants to be substantially higher quality than other online recruitment platforms (Douglas, Ewell, and Brauer 2023), suggesting that while online samples may not be optimal, they are unlikely to be systematically biased in a way that reduces the validity of our results.

Third, it is possible that LLM assistants could have an overall negative effect on forecasting accuracy compared to human forecasters without an LLM. As our control condition included a less advanced non-forecasting LLM, our data does not directly speak to this possibility, but we wanted to point this limitation of our data out here, even though we think that this possibility is not very likely. Further research may want to test this comparison specifically.

References

Abdurahman, Suhaib et al. (2023). "Perils and opportunities in using large language models in psychological research". In: *OSF Preprints* 10.

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.

Agarwal, Nikhil et al. (July 2023). *Combining Human Expertise with Artificial Intelligence: Experimental Evidence from Radiology*. Working Paper 31422. National Bureau of Economic Research. DOI: 10.3386/w31422. URL: http://www.nber.org/papers/w31422.

Antony, Victor Nikhil and Chien-Ming Huang (2023). "ID. 8: Co-Creating Visual Stories with Generative AI". In: *ACM Transactions on Interactive Intelligent Systems*.

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. (2017). "Distilling the wisdom of crowds: Prediction markets vs. prediction polls". In: *Management science* 63.3, pp. 691–706.

Atari, Mohammad et al. (2023). "Which humans?" In.

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, 610–623. ISBN: 9781450383097. DOI: 10.1145/3442188.3445922.

Benjamin, Daniel M. et al. (2023). "Hybrid Forecasting of Geopolitical Events". In: *AI Magazine*.

Bernard, David Rhys and Philipp Schoenegger (2024). "Forecasting Long-Run Causal Effects". In: *Available at SSRN 4702393*.

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.

Brynjolfsson, Erik, Danielle Li, and Lindsey R Raymond (Apr. 2023). *Generative AI at Work*. Working Paper 31161. National Bureau of Economic Research. DOI: 10.3386/w31161. URL: http://www.nber.org/papers/w31161.

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

Budescu, David V and Eva Chen (2015). "Identifying Expertise to Extract the Wisdom of Crowds". In: *Management Science* 61.2, pp. 267–280.

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_1cK.