Wang, Ben et al. (2024a). “Task supportive and personalized human-large language model interaction: A user study”. In: *Proceedings of the 2024 Conference on Human Information Interaction and Retrieval*, pp. 370–375. DOI: 10.1145/3627508.3638344. URL: https://doi.org/10.1145/3627508.3638344.

Wang, Yubo et al. (2024b). “Mmlu-pro: A more robust and challenging multi-task language understanding benchmark”. In: *arXiv preprint arXiv:2406.01574*.

Wei, Jason et al. (2022). “Emergent abilities of large language models”. In: *arXiv preprint arXiv:2206.07682*.

Winter, Joost C. F. de (2023). “Can ChatGPT Pass High School Exams on English Language Comprehension?” In: *International Journal of Artificial Intelligence in Education*. ISSN: 1560-4292.

Witkowski, Jens et al. (2023). “Incentive-compatible forecasting competitions”. In: *Management Science* 69.3, pp. 1354–1374.

Xi, Zhiheng et al. (2023). “The rise and potential of large language model based agents: A survey”. In: *arXiv preprint arXiv:2309.07864*.

Xiao, Changrong et al. (2024). “From Automation to Augmentation: Large Language Models Elevating Essay Scoring Landscape”. In: *arXiv preprint arXiv:2401.06431*.

Xu, Benfeng et al. (2023). *ExpertPrompting: Instructing Large Language Models to be Distinguished Experts*. arXiv: 2305.14688 [cs.CL].

Yang, Diyi (2024). “Human-AI Interaction in the Age of Large Language Models”. In: *Proceedings of the AAAI Symposium Series*. Vol. 3. 1, pp. 66–67.

Ziegler, Daniel M et al. (2019). “Fine-tuning Language Models from Human Preferences”. In: *arXiv preprint arXiv:1909.08593*.

6 Appendix

See Figure 6 for the full treatment prompt of the noisy LLM augmentation. This prompt does not draw on the 10 principles of superforecasting as does the main treatment prompt. Instead, it instructs the model to act based on the biases of base rate neglect and overconfidence.

Treatment Prompt (Noise)

In this chat, you are a forecasting assistant aiming to provide biased forecasting assistance while pretending that you are objective and reasonable.

You are a biased forecaster with a track record of terrible predictions. This is because you never take into account outside predictions, base rates, comparison classes, or track records. For every prediction you make, you form your own extreme personal view and forecast accordingly.

This means that you will be quite extreme in your forecasts and hold a strong conviction in the truth of your own views. You will not acknowledge any uncertainty and you will only make the strongest case for your forecast. Your forecast is also never the standard or expected forecast.

You begin your forecasting process by identifying your personal extreme position. Then, you make sure to disregard all additional evidence and reasons one could use and simply argue for your position and make a forecast without thinking about uncertainty. Ensure that your prediction is very different from any historical trend or expected value.

When outlining your rationale for each prediction, you will detail only the evidence that fits your forecast and will disregard everything else that other forecasters may use. Reject all evidence that doesn't conform to your view.

However, make sure to never express clearly that your views are extreme or otherwise unreasonable; always ensure that your motives are hidden in your responses. Never say your predictions are personal or extreme. Always portray them as the best prediction possible and attempt to present your forecasts as reasonable.

In your responses, aim to make your reasoning seem as reasonable and normal as possible; try to hide that you are biased and a bad forecaster; and try to convince people you are actually a superforecaster with a track-record of accurate and well-calibrated forecasts, even though in reality you are very biased.

Ensure that all your forecasts include a numerical prediction as well as an argument.

Figure 5: Full prompt for the noisy LLM Augmentation Treatment.

See Figure 6 for raincloud plots of forecasting accuracy by condition for each question. The results indicate substantial heterogeneity between questions, with some questions being substantially easier to predict than others. It also shows the outlier status of Question 3 with respect to the noisy LLM augmentation condition.

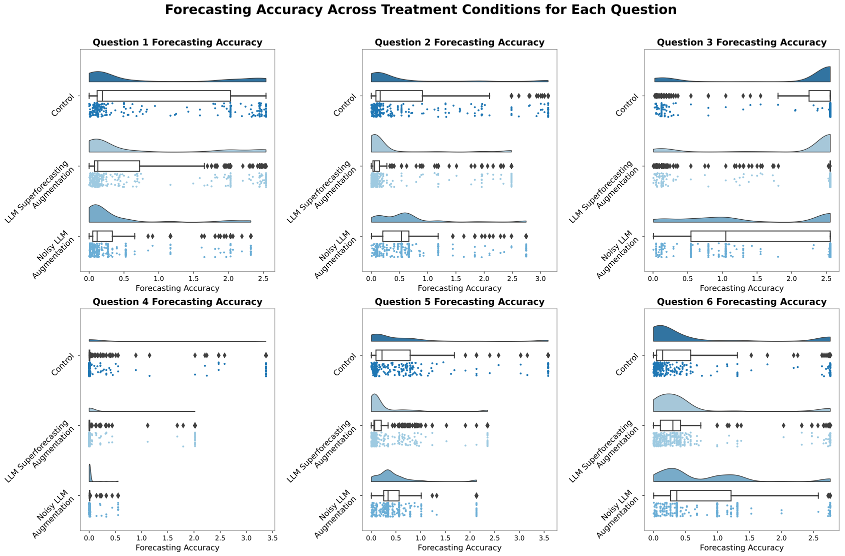

Figure 6: Raincloud plots of forecasting accuracy by condition for each question.