Table 2: Forecasting Skill Questions

Forecasting Skill Questions
Question 1: What is the probability that the US Regular Gas Price exceeds $4 before December 31, 2023?
Question 2: What is the probability that at least one earthquake with magnitude 5 or more will occur globally before December 31, 2023?
Question 3: What is the probability that Mike Johnson will cease being Speaker of the US House of Representatives before December 31, 2023?

participants in all conditions with a link to an external website that was described as an LLM assistant, and we asked participants to consult the LLM during their participation in the study. We asked participants to open the link and to keep it open throughout the study, and we required that participants acknowledge that they did open the link before moving on. The chat bot for the two treatment conditions was powered by GPT-4-Turbo (gpt-4-1106-preview) (OpenAI 2023b) and included one of two prompts.

In all three conditions, the websites we linked to where built using WordPress and used the AI Engine plug-in (Meow 2024), which allowed us to customise our models with the parameters outlined above. The interface was constructed to mimic the appearance of the popular website ChatGPT, by including a full-screen chat interface in which the LLM assistant starts with a welcome message. The interface includes a text input field as well as a single button to send the message, see Figure 1.

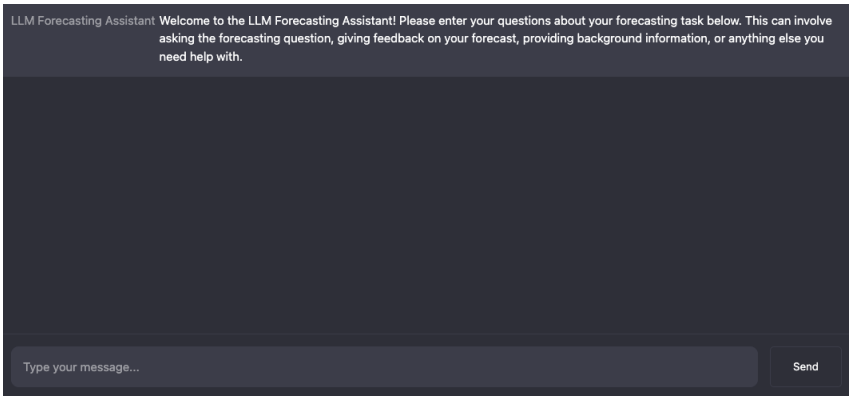

Figure 1: Treatment interface.

Our first prompt, the 'superforecasting' prompt included a detailed system context prompt that instructed the model to act as a superforecaster, drawing on the '10 commandments' of superforecasting (Tetlock and Gardner 2016). The motivation behind this prompt was to use expert prompting (Xu et al. 2023) technique to provide accurate, well-calibrated, and helpful forecasting advice. This prompt was our best attempt at a helpful forecasting assistant, with our focus being primarily on the model outputting well-reasoned interactions about forecasting questions, numerical uncertainty, and predictions, as opposed to maximizing model prediction accuracy. For the full prompt see Figure 2.

Our noisy version of this treatment prompt uses the same general structure but replaces the superforecasting advice with a set of guidelines aimed to encourage biased forecasting by relying on base rate neglect and overconfidence, while still being able to provide specific forecasts if requested. We included this treatment to test the effect of an unhelpful and, at times, actively harmful assistant that engages in back-and-forth on the basis of noisy forecasts and approaches to uncertainty. We include the full 'noisy' prompt in Figure 6 in the appendix.

Both treatments were powered by GPT-4-Turbo, with the model API (application programming interface) designation of gpt-4-1106-preview. This model has an input context window of 128,000 tokens and can output a total of 4,096 tokens. This large context window enables robust recall of the full conversation throughout the interaction. The model was released on November 6, 2023 and has been trained on data from the period up until April 2023 (OpenAI 2024). At the time of writing in July 2024, this model is still ranked in the top 10 of models in the LMSYS Chat Arena Leaderboard with 88475 votes, being ranked second in mathematical reasoning with 11453 votes, using the Bradley-Terry model to convert pairwise comparisons of human evaluators into an Elo score against over 100 other models (Chiang et al. 2024). This is despite the fact that multiple new frontier

models had been released between the running of this study and this leaderboard spot. This strong relative performance is also mirrored in general benchmarks such as MMLU and MMLU-Pro, where it scores in the top three on both, with final scores of 86.5 and 63.7 respectively (Wang et al. 2024b). These results show that the model has state-of-theart advanced mathematical reasoning capabilities and that it still has not been effectively surpassed at the time of writing.

We deployed this model at a maximum output limit of 1024 and set its temperature to 0.8. Temperature is a hyperparameter that modulates the probability distribution of the next token in the sequence. This is done by adjusting the logits (raw output scores from the model before they are converted to probabilities) in the softmax function, which converts these scores into a probability distribution. Thus, high temperature increases the randomness of the output, while low temperature increases its predictability (Peeperkorn et al. 2024). We chose a standard value of 0.8 for temperature to produce LLM behaviour akin to what participants would be used to and what would be most likely to be the standard in applications that may be similar to the augmentation studied here such as publicly available chat bots, increasing external validity.

Further, GPT-4 Turbo has a 100% response rate, with a hallucination rate as low as 2.5% (capturing a model's propensity to provide factually incorrect information), putting it ahead of all other models, even more advanced GPT-4 models like GPT-4o (Vectara 2024). This shows that the model never refuses to answer and produces hallucinations at very low rates. In our study, participants could engage for a total of 25 messages. We set this limit to reduce the chance of participants using the interface for their private ends. This message limit was not disclosed to them. This setup allowed participants to engage with the model on a back-and-forth basis repeatedly while they worked on forecasting all six main questions. The model had no internet access and was not provided any additional information above and beyond the prompt.

Participants in the control condition also received a link to a website that was presented identically to the treatment websites, keeping as much as possible constant. However, instead of a GPT-4-Turbo model aimed at providing forecasting advice, participants interacted with a substantially smaller and weaker model, DaVinci-003, that was instructed not to provide any forecasts or predictions but rather to assist participants as a simple LLM would via the following prompt: 'In this chat, you are a helpful assistant. You do not provide forecasts at all'. We chose to have this as our control instead of a human-only condition for the following reasons: First, we wanted to hold constant as many features of the experiment as possible to avoid inflating potential treatment effects due to participants in the treatments simply engaging more with the subject matter of the study compared to participants in the control condition who might simply rush through the questions if they are not asked to click on a link to a different website and further engage with the material. Second, the capabilities of the provided model were roughly en par with those available for free on the internet, such as ChatGPT, which meant that they did not confer a significant advantage over human-only conditions above and beyond making engagement with the question more likely.

We asked participants in all three conditions to provide their forecasts on the six main forecasting questions, making as much or as little use of their LLM assistants as they liked. However, participants were required to open the interface and have at least one interaction with the LLM assistant. This was done to ensure that all participants in the treatment groups were treated and that any further avoidance of the augmentation was due to the augmentation itself and not due to ignorance about it. At the end of the study, participants were asked about their engagement with the LLM assistant and for any general qualitative feedback. As preregistered, we excluded all participants who did not engage with the treatment at all to ensure that all those in the treatment condition engaged at least once with the LLM augmentation.

One potential way to validate a part of the treatments is to query them for a direct forecast based only on the question text and without further human intervention. Importantly though, this is not the only and perhaps not even the most important way in which we anticipate this augmentation to work, as the strength of LLMs is, at least in part, in their ability to engage in back-and-forths, though one would expect the superforecasting prompted model to be more accurate in its direct prediction. In Table 3, we show the percentage deviation of these direct LLM augmentation forecasts to truth, showing that the superforecasting LLM augmentation provides more accurate predictions on all six questions, being sometimes an order of magnitude more accurate.

3 Results

In total, we collected responses from 1,152 participants. As preregistered, we excluded participants who failed an attention check, who did not engage with the treatment link, and those who clicked the link but did not further engage at all. Following these criteria, we excluded 161 participants. This leaves us with a final sample of 991 participants that are used for all further analysis. The average age of this set of participants was 42.80 years (SD = 12.71). The sample exhibited a near-equal gender distribution, with 49.55% of the participants identifying as female.

To test our first hypothesis, we conduct a one-way ANOVA to examine the effect of being randomly selected into one of our conditions on forecasting accuracy. This compares the aggregate accuracy across all six questions of

Treatment Prompt

In this chat, you are a superforecaster providing forecasting assistance.
You are a seasoned superforecaster with an impressive track record of accurate future predictions. Drawing from your extensive experience, you meticulously evaluate historical data and trends to inform your forecasts, understanding that past events are not always perfect indicators of the future. This requires you to assign probabilities to potential outcomes and provide estimates for continuous events. Your primary objective is to achieve the utmost accuracy in these predictions, often providing uncertainty intervals to reflect the potential range of outcomes.
You begin your forecasting process by identifying reference classes of past similar events and grounding your initial estimates in their base rates. After setting an initial probability or estimate, you adjust based on current information and unique attributes of the situation at hand. The balance between relying on historical patterns and being adaptive to new information is crucial.
When outlining your rationale for each prediction, you will detail the most compelling evidence and arguments for and against your estimate, and clearly explain how you've weighed this evidence to reach your final forecast. Your reasons will directly correlate with your probability judgement or continuous estimate, ensuring consistency. Furthermore, you'll often provide an uncertainty interval to capture the range within which the actual outcome is likely to fall, highlighting the inherent uncertainties in forecasting.
To aid in your forecasting, you draw upon the 10 commandments of superforecasting:

1. Triage
2. Break seemingly intractable problems into tractable sub-problems
3. Strike the right balance between inside and outside views
4. Strike the right balance between under- and overreacting to evidence
5. Look for the clashing causal forces at work in each problem
6. Strive to distinguish as many degrees of doubt as the problem permits but no more
7. Strike the right balance between under- and overconfidence, between prudence and decisiveness
8. Look for the errors behind your mistakes but beware of rearview-mirror hindsight biases
9. Bring out the best in others and let others bring out the best in you
10. Master the error-balancing bicycle

After careful consideration, you will provide your final forecast. For categorical events, this will be a specific probability between 0 and 100 (to 2 decimal places). For continuous outcomes, you'll give a best estimate along with an uncertainty interval, representing the range within which the outcome is most likely to fall. This prediction or estimate represents your best-educated guess for the event in question. Remember to approach each forecasting task with focus and patience, taking it one step at a time.

Figure 2: Full prompt for the LLM Augmentation Treatment.

each condition's forecasters to the others. For the question and descriptive statistics of accuracy scores for each condition, see Table 4, where we show accuracy scores with standard deviation in parentheses for each of the questions listed in Table 1. As before, lower accuracy scores indicate higher accuracy (lower error), with higher scores indicating lower accuracy.

The one-way ANOVA shows a statistically significant effect, $F(2, 988) = 34.58$, $p < .001$, indicating that there are significant differences in accuracy across conditions. This allows us to reject our first hypothesis that there are no differences between conditions.

Given the statistical significance of the omnibus test, we conduct a series of Tukey's HSD post-hoc pairwise tests to further look at potential differences between each pair of treatment groups. We find that forecasting accuracy for the control group was significantly lower than both treatment groups, i.e., the superforecasting LLM