AI-Augmented Predictions: LLM Assistants Improve Human Forecasting Accuracy

Philipp Schoenegger
London School of Economics
and Political Science

Peter S. Park
MIT

Ezra Karger
Federal Reserve Bank of Chicago

Philip E. Tetlock
University of Pennsylvania

Abstract

Large language models (LLMs) show impressive capabilities, matching and sometimes exceeding human performance in many domains. This study explores the potential of LLMs to augment judgement in forecasting tasks. We evaluated the impact on forecasting accuracy of two GPT-4-Turbo assistants: one designed to provide high-quality advice ('superforecasting'), and the other designed to be overconfident and base-rate-neglecting. Participants (N = 991) had the option to consult their assigned LLM assistant throughout the study, in contrast to a control group that used a less advanced model (DaVinci-003) without direct forecasting support. Our preregistered analyses reveal that LLM augmentation significantly enhances forecasting accuracy by 23% across both types of assistants, compared to the control group. This improvement occurs despite the superforecasting assistant's higher accuracy in predictions, indicating the augmentation's benefit is not solely due to model prediction accuracy. Exploratory analyses showed a pronounced effect in one forecasting item, without which we find that the superforecasting assistant increased accuracy by 43%, compared with 28% for the biased assistant. We further examine whether LLM augmentation disproportionately benefits less skilled forecasters, degrades the wisdom-of-the-crowd by reducing prediction diversity, or varies in effectiveness with question difficulty. Our findings do not consistently support these hypotheses. Our results suggest that access to an LLM assistant, even a biased one, can be a helpful decision aid in cognitively demanding tasks where the answer is not known at the time of interaction.

*Any views expressed in this paper do not necessarily reflect those of the Federal Reserve Bank of Chicago or the Federal Reserve System.

1 Introduction

Recent advances in artificial intelligence (AI), and large language models (LLMs) specifically, demonstrate that AI systems have impressive abilities across a large number of complex and economically valuable tasks (Naveed et al. 2023). This development challenges previously held beliefs about the necessity of human cognition for many of these tasks (Bubeck et al. 2023), and raises the possibility of significant negative effects of AI systems on the (human) labor market in large parts of the knowledge economy (George and Baskar 2023). Understanding the current ability of LLMs to interface with economically central tasks requires a broad empirical study across domains. However, most knowledge-work jobs require substantial reasoning capabilities that use data and patterns of observations beyond any model's training data. This makes finding a suitable study context central in any attempt to understand how LLMs might impact advanced economies.

LLMs are comprised of a vast array of parameters, built using the Transformer architectural paradigm (Vaswani et al. 2017) and trained on a large amount of data (Shen et al. 2023). These advanced AI systems excel at next-token prediction: the ability to predict the next word or subword (token). But this general training objective of next-token prediction also equips these models with a variety of specialized advanced skills, often in an emergent way that cannot have been fully predicted before training due to nonlinearities in capabilities (Wei et al. 2022). These skills include marketing (Fraiwan and Khasawneh 2023), translation (Jiao et al. 2023), high levels of reading comprehension (Winter 2023), teaching (Fraiwan and Khasawneh 2023; Sallam et al. 2023), summarization (T. Goyal, J. J. Li, and Durrett 2023), abstract categorization of objects (Atari et al. 2023), programming (Bubeck et al. 2023), spear phishing cyber attacks (Hazell 2023; Heiding et al. 2023), robotics (Vemprala et al. 2023), medical reasoning (Bubeck et al. 2023; Nori et al. 2023; Sallam et al. 2023), legal reasoning (Bubeck et al. 2023; Katz et al. 2023), deception (Park, Goldstein, et al. 2023), and others. LLMs' many capabilities substantially increase the amount of money and talent going into LLM research and products (Sutton 2023), suggesting further growth in capabilities in the near future.

Crucially, LLMs are not naturally autonomous for many of the relevant tasks (Xi et al. 2023), although they can be made generally autonomous with an agent framework such as AutoGPT (Firat and Kuleli 2023) or other scaffolding approaches. Moreover, future iterations of models may enable such behavior directly (Kinniment et al. 2023), making agency, the ability to take actions and achieve goals independently more widely accessible. At the time of writing, however, LLMs are not economically useful as autonomous agents due to prominent limitations like inefficiency, forgetting, and hallucinations (Firat and Kuleli 2023). Instead, they are generally used in conjunction with human labor as a hybrid technology that requires human input at several stages (Dell'Acqua et al. 2023). This hybrid use of LLMs enables humans to combine their strengths with those of the models to produce output that is, at least in some aspects, more efficient than the output produced by either humans or machines alone. For example, LLM augmentations have been shown to improve performance of human graders (Xiao et al. 2024) as well as that of programmers (Peng et al. 2023).

In this paper, we study the application of present-era LLMs as a hybrid augmentation technology in the context of forecasting future events. This allows us to test their ability to augment human decision-making in a domain robust to in-sample overfitting of training data, since no one, including LLMs, can know the answer to prospective forecasting questions at the time of data collection. This context is also practically relevant as accurate forecasting is essential to many aspects of economic activity, especially within white-collar occupational domains such as law, business, and policy: fields that may be disrupted or even replaced by LLM capabilities (Acemoğlu 2023; Park and Tegmark 2023; Summers and Rattner 2023). If the use of present or future AI systems increases the forecasting accuracy of humans and organizations, the efficiency and productivity gains to the relevant industries' individuals and businesses are clear.

Our specific object of interest in this study is human judgment forecasting, where humans provide forecasts of future events, such as the probability that inflation will hit a certain milestone over the next twelve months or the anticipated number of barrels in the Strategic Petroleum Reserve at the end of the year. The science of forecasting has found that aggregated forecasts of a crowd of forecasters can be surprisingly accurate (P. E. Tetlock and Gardner 2016), can impact policy debates (P. E. Tetlock, B. A. Mellers, and Scoblic 2017), and can affect businesses (Schoemaker and P. E. Tetlock 2016). Previous work on the topic focuses on a variety of other topics, ranging from the identification of skilled forecasters (Himmelstein, Budescu, and Han 2023; B. Mellers et al. 2015; P. E. Tetlock and Gardner 2016) and novel aggregation methods (P. Atanasov et al. 2017) to improvements of forecasting accuracy (Chang et al. 2016; Karger, P. D. Atanasov, and P. Tetlock 2022).

Similar to our project, some previous work focuses on human-machine hybrid forecasting in the context of IARPA's 'Hybrid Forecasting Competition.' Benjamin et al. (2023) report the results of 'SAGE,' a hybrid forecasting system designed to combine human- and machine-generated forecasts (such as ARIMA forecast outputs). They find that their hybrid forecasting system outperformed their human-only baseline, suggesting that cost savings and accuracy increases of these hybrid systems may be "a viable approach for maintaining a competitive level of accuracy" (Benjamin et al. 2023, p. 113). Similarly, P. Atanasov et al. (2017) introduce a 'Human Forest' method that enables human forecasters to define custom reference classes, draw on historical databases, and review base rates in their forecasting. They find that these forecasters outperform statistical model predictions. However, both approaches used pre-LLM methods as their machine counterparts, which makes

them more static and predictable than frontier LLMs and their potential hybrid functions that go beyond previous machine capabilities.

In this paper, we update this literature in light of recent breakthroughs in frontier LLMs, enabling a free exchange between the human and the model in a way that previous technology did not. Those interacting with the model could query it to fill their own gaps in knowledge or perceived weaknesses. They could ask it to produce a forecast for them, they could input their own reasoning and predictions into the model for feedback, or they could do a combination of these and other approaches they might find helpful. Our goal is to probe whether LLM-augmentation can be a cheap, scalable, and effective method of improving human judgement forecasting. Inference costs for LLMs remain low, at a few cents per 1000 tokens, making LLM augmentation a prime candidate for a generalized hybrid system that can boost individual proficiency in many valuable tasks at costs far below a human assistant equivalent, if such an equivalent existed.

While traditional measures of AI proficiency often rely on task benchmarks, we argue that evaluating forecast accuracy in real-world scenarios presents a more comprehensive assessment of reasoning capabilities. This method also increases the likelihood of these results generalizing to different and even out-of-distribution settings (Arora and A. Goyal 2023). Our approach diverges from conventional task benchmarks, as it focuses on the LLM's ability to apply its knowledge and understanding to novel settings, rather than settings represented in some shape or form in its training data. Even if an LLM excels at a given task benchmark, it is unclear whether this reveals a deep understanding of the process behind the task, instead of rote memorization of the task benchmark's answers in the training data (Bender et al. 2021; Biderman et al. 2023; Carlini et al. 2023; Magar and Schwartz 2022). The difficulty in disentangling true understanding from training data memorization is non-trivial. Deep understanding, after all, also originates from exposure to relevant content within the training dataset. However, the success or failure to generalize outside of the training data appears central to this disentangling (Grove and Bretz 2012). In our study, we analyze human forecasting behavior on a set of prediction questions that resolve in the future such that it is impossible for any human forecaster or AI-based system to access the answer at the time of data collection.

Past work found that the frontier model GPT-4 significantly underperformed the median human-crowd forecast in a real-world forecasting tournament, failing to even significantly outperform the no-information forecasting strategy of uniform random guessing (Schoenegger and Park 2023). However, this previous study investigated the static machine forecasts produced solely by the LLM, without incorporating human input. It is reasonable to expect that human-LLM hybrid forecasts—the object of study in the present paper—would outperform the poor results of the LLM operating by itself. While hybrid forecasting approaches have been previously studied—for example, in making predictions on geopolitical questions (Benjamin et al. 2023) and in radiology (Agarwal et al. 2023)—our approach is arguably more meaningfully hybrid, in that a human forecaster can engage in a back-and-forth dialogue with an LLM to fill gaps in knowledge, understanding, and data that differ on a person-by-person level. This back-and-forth LLM augmentation may allow forecasters to use the model for the parts of forecasting that they themselves struggle most with: be it synthesizing data, making coherent forecasts, or attaching numbers to intuitions. This motivates our first research question and accompanying hypothesis, testing whether we find an aggregate accuracy improvement of LLM augmentation. We test two treatments, one where the human has access to an LLM with a 'superforecasting' (P. E. Tetlock and Gardner 2016) prompt and the other using a biased LLM prompt to exhibit base rate neglect and overconfidence. Both models are instructed to assist forecasters in whatever way is requested, ranging from providing point estimates to offering feedback on forecasts. We predicted that the superforecasting LLM augmentation would outperform the biased LLM augmentation, and that both hybrid treatment arms would have higher aggregate accuracy than the control.

> **Null Hypothesis 1**: There is no difference in forecasting accuracy between the superforecasting (biased) LLM augmentation and the control.

Recent work has also shown that less skilled individuals benefit the most from LLM augmentation. For example, LLM augmentation boosted the performance of low-performing professionals more than that of high-performing professionals in studies where it was provided to management consultants (Dell'Acqua et al. 2023), customer-support agents (Brynjolfsson, D. Li, and Raymond 2023), creative writers (Doshi and Hauser 2023), office workers who write memos (Noy and W. Zhang 2023), law school students who write exams (Choi and Schwarcz 2024), and programmers (Peng et al. 2023). However, other work in the context of medicine found that human-AI hybrid decisions are not associated with increased diagnostic quality, suggesting that the effects of AI may show substantial heterogeneity across subject domains and implementation details (Agarwal et al. 2023). This suggests that any effects of LLM-augmentation on forecasting are likely to be heterogeneous across the skill distribution, with lower-skill forecasters relying to a greater degree on LLM augmentation which may help alleviate biases in their predictions. This motivates our second hypothesis, which directly tests whether the LLM augmentation has disparate impacts on forecasters of different skill levels. In line with previous literature, we predicted a greater effect on lower skill forecasters.

> **Null Hypothesis 2**: The effect of the superforecasting (biased) LLM augmentation on forecasting accuracy does not differ between high- and low-skilled forecasters.

In addition to investigating the effects of LLM augmentation on individual forecasts and on forecasters of different levels of skill, we also investigate its potentially adverse effects on aggregate forecasts. Due to the 'wisdom of the crowd' effect, aggregation—such as taking the median forecast—tends to result in an overall forecast that is more accurate than the forecasts given by most individuals, even across heterogeneous types of forecasters who may have different skill levels (Budescu and E. Chen 2015; Mannes, Soll, and Larrick 2014). However, this aggregation tends to be most effective when there is a diversity of forecasts. If the LLM augmentation anchors human forecasters on the same forecast for a given question, it could reduce the value of aggregation. We test whether LLM augmentation homogenizes forecasts: motivating our third hypothesis, where we predicted a reduction in accuracy.

> **Null Hypothesis 3**: There is no difference in aggregate level forecasting accuracy between the superforecasting (biased) LLM augmentation and the control.

Finally, we compare the effect LLM augmentation has on forecasting questions of different difficulty levels. There are a number of reasons why the difficulty of the forecasting question may be an important factor. To illustrate, consider the plausibility of the following mechanism. Without Bayesian-rational calibration, human forecasts tend to only partially take into account the difficulty of a given forecasting question in their estimate (Lichtenstein and Fischhoff 1977; Moore and Healy 2008; Park 2022). In addition, there is a 'correct answer' effect in LLM output (Abdurahman et al. 2023; Park, Schoenegger, and Zhu 2024; Solaiyappan et al. 2023), where an LLM can answer even non-straightforward questions with near-zero or zero variance: with a predetermined answer. If this 'correct answer' effect were to also affect LLM augmentation, then a plausible interplay is that while human forecasters successfully answer easy forecasting questions, they do not sufficiently take into account the difficulty of hard forecasting questions and instead are led astray by the LLM's overconfidently predetermined 'correct answers.' This would increase the degree of groupthink and thereby reduce the benefit of LLM augmentation on forecasting accuracy, potentially even to the degree of decreasing forecast accuracy. This motivates our last hypothesis, where we did not have a specific prediction.

> **Null Hypothesis 4**: There is no difference in the effect of the superforecasting (biased) LLM augmentation on forecasting accuracy between hard and easy questions.

2 Methods

All analyses were preregistered on the Open Science Framework[1]. We clearly label all exploratory/non-preregistered analyses as such throughout the paper to indicate which analyses we decided to investigate after having seen the data. This study has received ethics approval prior to data collection.[2]

For our study, we recruited a total of 1152 participants from Prolific, an online research platform gives researchers access to people willing to participate in research. Participants were paid $5 for participation and could earn an additional $100 based on their accuracy (we paid three such prizes to randomly selected participants who scored in the top-10 of forecasters). We preregistered the following a priori power analysis: Using Cohen's d=0.20 as our smallest effect size of interest as a conventionally small effect, with an allocation ratio of 1.5/1/1 between the main treatment, the secondary bias treatment, and the control, aiming for 80% power, we needed to recruit 492 participants for the Main treatment and 328 for the other two conditions, resulting in a final participant count of 1148. We recruited a total of 1152 participants, meeting our goal.

Our central dependent variable throughout this study was aggregate forecasting accuracy on a set of six continuous forecasting questions that ranged from financial questions to geopolitical ones. For a full list, see Table 1. Data collection happened on November 21, 2023, over five weeks prior to forecast question resolution. We computed participant's accuracy by comparing their forecasts to the true value of the forecasted question. We computed the initial accuracy calculation for each forecasting question i as the absolute difference D_i between the participant's forecast F_i and the actual value A_i, expressed as $D_i = |F_i - A_i|$. As preregistered, we conducted a 5% winsorisation process. Then, we standardized the values by dividing them by the standard deviation of the control group for the respective question to normalise the values, allowing for inter-question comparability in accuracy scores. Lastly, we conducted a second preregistered winsorisation step, this time at the level of 3 standard deviations. For this accuracy measure, lower scores correspond to higher accuracy.

Our secondary variables, question difficulty, and forecaster skill, were determined as follows. A selected 10% of the control group participants were tasked not only with providing forecasts for each question but also with rating the perceived difficulty of each question on a 5-point Likert scale ranging from 'Very easy' to 'Very difficult'. Questions 2 and 3, receiving the highest difficulty ratings and we therefore identified those questions as being the most challenging. In addition, prior to the main forecasting tasks, participants were asked a series of smaller, lower-effort forecasting questions. These questions included binary predictions and intersubjective forecasts,

[1] `https://osf.io/d9rhx/?view_only=c631c477026a41f3bd4e6b7a4e546157`

[2] University of Pennsylvania Institutional Review Board IRB Protocol number: 854515

Table 1: Main Study Questions

Main Forecasting Questions
Question 1: What will be the closing value for the Dow Jones Transportation Average on December 29, 2023?
Question 2: How many refugees and migrants will arrive in Europe by sea in the Mediterranean between December 1, 2023 and December 31, 2023?
Question 3: What will Bitcoin's network hash rate per second be (in TH/s) according to the performance rates posted by blockchain.com on December 31, 2023?
Question 4: How many commercial flights will be in operation globally on December 31, 2023?
Question 5: How many AI papers will be published on ArXiv during the month between December 1, 2023 and December 31, 2023?
Question 6: What will be the closing value for the U.S. Dollar against the Russian Ruble (converting 1 USD to RUB) on December 30, 2023?

to evaluate their ex-ante forecasting skill. Forecaster skill was quantified in two ways: firstly, through Brier scores for binary predictions, defined as Brier Score $= \frac{1}{N}\sum_{n=1}^{N}(f_n - o_n)^2$, where f_n represents the forecast probability, o_n the actual outcome, and N the total number of binary forecasts. Secondly, intersubjective forecast accuracy was measured using the Euclidean distance formula Euclidean Distance $= \sqrt{\sum_{i=1}^{k}(p_i - q_i)^2}$, with p_i being the participant's forecast and q_i the average forecast for each question. Then, we ranked participants based on these two metrics and created a composite measure. The top half of participants based on this composite measure was classified as skilled forecasters. For the set of questions used for the skill measures, see Table 2. Brier scores were calculated with participant accuracy on these questions. Intersubjective accuracy was calculated with respect to participant responses to the question 'What is the average probability that participants in this study give on the above question?' to each of the questions in Table 2.

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?

For our main analyses, participants were randomly selected into one of three conditions—Treatment (including the superforecasting prompt), Treatment (Bias) (including a biased prompt), and Control—with an allocation ratio of 1.5/1/1. We presented participants in the Treatment and Treatment (Bias) conditions with a link to an external website that was described as an LLM forecasting 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 website itself was a full-screen chat interface in the style of the well-known ChatGPT (see Figure 1). It was powered by GPT-4-Turbo (OpenAI 2023) at a temperature setting of 0.8 and included a detailed context prompt that instructed the model to act as a superforecaster, drawing on the '10 commandments' of superforecasting (P. E. Tetlock and Gardner 2016), for the full prompt see Figure 2. Our biased version of this treatment uses the same general structure but replaces the superforecasting advice with a set of guidelines aimed to encourage

biased forecasting by relying on baserate neglect and overconfidence. We include the full 'biased' prompt in Figure 5 in the appendix.

Both treatments were powered by GPT-4-Turbo at a temperature of 0.8 with 1024 maximum tokens. Participants could engage for a total of 25 messages but this number was not disclosed to them. This allowed participants to engage with the model on a back-and-forth basis repeatedly on each of the six questions. This engagement could include both directly asking the models for forecasts, which they were explicitly instructed to provide, as well as to ask for feedback on their own forecasts or engage in dialogue of any kind.

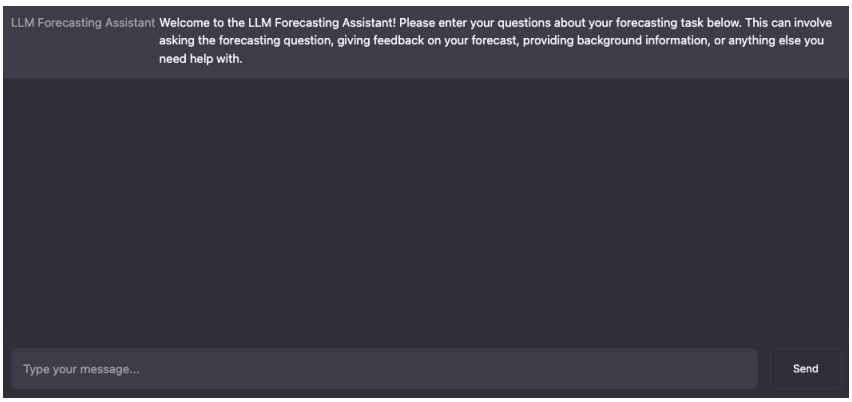

Figure 1: Treatment interface.

We gave participants in the control condition a link to a website that was presented identically to the treatment websites, but 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 forecasts but rather to assist participants as a simple LLM would. Having a model at the capability level of free LLMs like ChatGPT as the control allows for a more rigorous comparison than not providing them with an interactive model at all.

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 that 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 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. Nonetheless, 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. From looking at the chat logs of participants, we also find that both models engaged in the anticipated behavior of forecast elicitation and back-and-forth, providing further evidence in favor of the treatments working as intended.

Table 3: Deviation of Direct LLM Augmentation Predictions from Truth

	Deviation (Superforecasting)	**Deviation (Biased)**	**Superforecasting > Biased**
Question 1	-5.65%	+13.22%	✓
Question 2	+19.88%	+470.84%	✓
Question 3	-48.90%	+57.24%	✓
Question 4	-3.76%	+46.12%	✓
Question 5	-55.05%	+322.48%	✓
Question 6	-15.20%	+69.61%	✓

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 judgment 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.

3 Results

In total, we collected responses from 1152 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. In total, we excluded 161 participants. This leaves the final sample at 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-equitable gender distribution, with 49.55% of the participants identifying as female.

To test our first hypothesis, we conducted 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 of 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

Table 4: Average Accuracy Scores with Standard Deviation by Condition

Condition	Average Score	Question 1	Question 2	Question 3
Control	0.82 (0.50)	0.67 (0.76)	0.69 (0.98)	1.92 (0.96)
Treatment	0.63 (0.63)	0.50 (0.69)	0.33 (0.68)	2.02 (0.89)
Treatment (Bias)	0.60 (0.43)	0.32 (0.50)	0.67 (0.66)	1.41 (0.94)

Condition	Question 4	Question 5	Question 6
Control	0.39 (1.00)	0.66 (0.96)	0.62 (0.92)
Treatment	0.15 (0.50)	0.28 (0.53)	0.49 (0.71)
Treatment (Bias)	0.03 (0.10)	0.46 (0.46)	0.72 (0.69)

in Table 1. The one-way ANOVA found a statistically significant effect, F(2, 988) = 34.67, p < .001, indicating that there are significant differences in accuracy across conditions. This allows us to reject our first hypothesis.

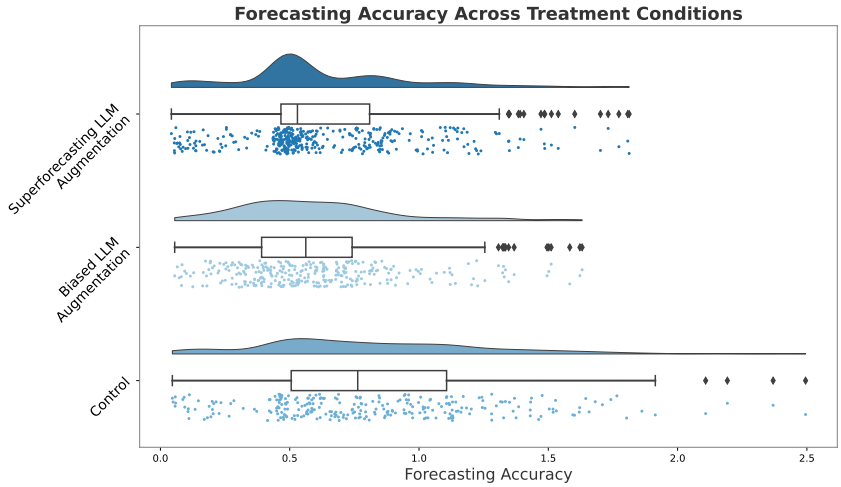

Figure 3: Raincloud plot of forecasting accuracy by condition.

Given the significance of the omnibus test, we conducted a series of Tukey's HSD post-hoc pairwise tests to further explore the differences between each pair of treatment groups. We found that that forecasting accuracy for the control group was significantly lower than both treatment groups, i.e., the superforecasting LLM augmentation (mean difference = -0.20, p < .001, 95% CI [-0.26, -0.13]) as well as the the biased LLM augmentation (mean difference = -0.23, p < .001, 95% CI [-0.29, -0.16]). However, we failed to detect a significant difference in forecasting accuracy between the biased LLM augmentation and the superforecasting LLM augmentation (mean difference = 0.03, p = .506, 95% CI [-0.03, 0.09]). This suggests that both GPT-4-Turbo powered treatments, irrespective of the fact that they were instructed to provide good or biased forecasting advice, outperformed the baseline of a simple LLM assistant that does not provide direct forecasting aid. See Figure 3 for a raincloud plot of accuracy by condition. We also plot the CDFs of accuracy for each condition, see Figure 4.

We also conduct the following exploratory analyses. Looking at the impact that individual questions have on the aggregate accuracy measure, we found that it is primarily the effect of the biased LLM augmentation on Question 3 that drives these results. Running the same analysis without Question 3, we find that the LLM augmentation's mean accuracy of 0.35 is significantly lower than both the biased LLM augmentation at 0.44 and the Control's at 0.61, with the Tukey HSD post-hoc pairwise comparison p-value at 0.006 for the comparison between the two augmentations This suggests that it is primarily Question 3 that leads to the similar effects of both treatments in the preregistered aggregate analysis. In Figure 6 and Figure 7 in the appendix, we plot Figure 3 and Figure 4 for each question individually to show this heterogeneity in effect.

We use a preregistered regression model to test our second hypothesis pertaining to the differential impacts of LLM augmentation on forecasters with varying skill levels.. The dependent variable in this model, representing forecasting accuracy, is denoted as Y. The independent variables in our model include: $T1$, representing the LLM superforecasting augmentation treatment group; $T2$, signifying the LLM augmentation treatment group with introduced bias; and S, indicating the high skill group among the forecasters. Crucially, the model integrates interaction terms $\beta_4(T1 \cdot S)$ and $\beta_5(T2 \cdot S)$. These terms allow us to directly examine the interaction effect

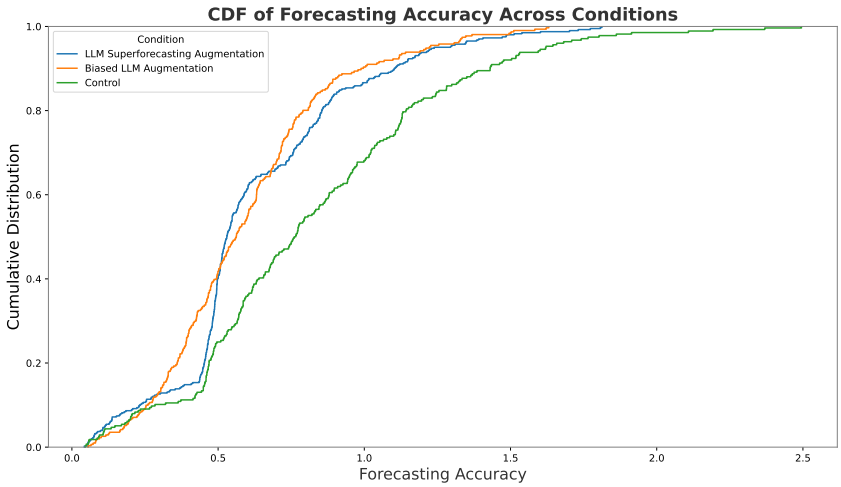

Figure 4: CDF of forecasting accuracy by condition.

between the LLM augmentation (both with and without bias) and the forecasters' skill level. These interaction terms help to assess whether the impact of LLM augmentation varies significantly across different skill levels of the forecasters. The regression model is given by:

$$Y = \beta_0 + \beta_1 T1 + \beta_2 T2 + \beta_3 S + \beta_4 (T1 \cdot S) + \beta_5 (T2 \cdot S) + \epsilon \tag{1}$$

Table 5: LLM Augmentation Skill Effects: OLS Regression Results

Variable	Coefficient	Std. Error	t-value	p-value
Intercept	0.87	0.03	30.05	< 0.001
Treatment	-0.24	0.04	-6.16	< 0.001
Treatment (Bias)	-0.24	0.04	-5.92	< 0.001
High Skill	-0.10	0.04	-2.24	0.025
Treatment · High Skill	0.09	0.06	1.63	0.103
Treatment (Bias) · High Skill	0.05	0.06	0.87	0.383
Observations		991		
R-squared		0.07		
Adjusted R-squared		0.07		
F-statistic		15.16		
Prob (F-statistic)		< 0.001		

Consistent with our previous finding, we observed significant treatment effects for both the treatment condition variables, which are associated with an increase in accuracy, as indicated by their negative coefficients, which are statistically significant (Treatment: $b = -0.24, p < .001$; Treatment (Bias): $b = -0.24, p < .001$). Furthermore, having been categorized as a high-skill forecaster was positively correlated with increased accuracy, with the skill dummy showing a significant negative coefficient ($b = -0.10, p = .025$), suggesting that individuals categorized as higher skill are indeed more accurate in their forecasts of the main task. However, we do not find statistically significant results for the main hypothesis test, i.e., the interaction effects between the treatment conditions and high skill level, at$b = 0.09, p = .103$ for the superforecasting LLM augmentation condition and $b = 0.05, p = .383$) for the biased LLM augmentation condition. This indicates a lack of evidence to support the hypothesis that the effect of the treatment on accuracy has distinct effects based on the forecasting skill level of the participants. As such, we are unable to reject the second hypothesis. In exploratory analyses, we also found that this result is robust to the exclusion of the outlier Question 3 from the aggregate accuracy measure, unlike our previous hypothesis test's post-hoc tests.

We now test our third hypothesis: that the LLM augmentation may harm aggregate accuracy. We did this by running a bootstrap analysis on the median forecasts, which represent the aggregate forecast for each condition. Initially, medians for each dependent variable were calculated within each treatment condition. Subsequently, these medians were averaged to yield a single summary measure per group. A bootstrap procedure with 10,000

resamples is used to estimate 95% confidence intervals for these estimates. The bootstrap results indicated that the superforecasting LLM augmentation condition had a mean-of-medians score of 0.49 (95% CI [0.49, 0.50]), the biased LLM augmentation condition scored 0.39 (95% CI [0.38, 0.44]), and the control condition scored 0.52 (95% CI [0.50, 0.55]). These outcomes suggest notable differences in forecast accuracy across the conditions, with the Control condition demonstrating the lowest accuracy and the biased LLM augmentation condition showing the highest accuracy, with the superforecasting LLM augmentation falling somewhere in the middle. This provides unexpected results with respect to our null hypothesis, as we do find that the biased LLM augmentation improves aggregate forecasting over the other two conditions, but the superforecasting LLM augmentation is not different from the control. This provides mixed results, as we find an increase in aggregate accuracy compared to the control in only one of the two treatments.

In a similar manner to the exploratory tests we performed for our initial hypothesis, we also carried out a sensitivity analysis. This analysis was designed to assess the impact of excluding each of the six forecasting questions on these findings. This involved examining how the removal of each item, one at a time, affects the overall findings. We found that, except for Question 3, the pattern of results remained largely consistent. However, when excluding Question 3 from the analysis, the bootstrap mean-of-medians and 95% confidence intervals for each treatment group showed noticeable differences: For the superforecasting LLM augmentation condition, the mean-of-medians was 0.10 (95% CI [0.09, 0.11]), indicating relatively higher accuracy. In contrast, the biased LLM augmentation condition exhibited a higher mean-of-medians of 0.26 (95% CI [0.26, 0.29]), while the control condition had a mean-of-medians of 0.13 (95% CI [0.10, 0.17]). These findings suggest that Question 3 in particular contributed to the overperformance of the biased LLM augmentation condition compared to the other two groups which is in line with the results testing null hypothesis 1, where we also find Question 3 to drive this pattern of results. Importantly, compared to the pre-registered analyses, here we find a significantly reduced accuracy for the biased LLM augmentation but not the superforecasting LLM augmentation.

We conclude from this that our data suggest that there is no clear picture as to the effects of LLM augmentation on aggregate level accuracy. Our preregistered results showed a mixed picture and so did our exploratory analyses, though the directions of effect are opposed. At the very least, our data do not convincingly show that the introduction of LLM augmentation reduces the wisdom of the crowd effects uniformly.

Lastly, in Study 4, we tested our fourth hypothesis pertaining to whether the superforecasting LLM augmentation has a distinct effect on easier or harder forecasting questions. It may be the case that it provides strong forecasting support for harder questions while not improving easier questions that much. We ran a mixed effects model with accuracy as our dependent variable. This approach allowed us to account for both individual differences among participants and varying levels of difficulty in forecasting questions. The model included fixed effects for the treatment conditions ($T1$, $T2$), a binary variable indicating the difficulty level of each question (D), and interaction terms between the treatment conditions and difficulty levels, represented as $\beta_4(T1 \cdot D)$ and $\beta_5(T2 \cdot D)$. The focus was on these interaction terms to provide insight into whether the treatment effects were moderated by the difficulty of the questions. The model is given by

$$Y_{ij} = \beta_0 + \beta_1 T1_j + \beta_2 T2_j + \beta_3 D_i + \beta_4(T1_j \cdot D_i) + \beta_5(T2_j \cdot D_i) + u_j + \epsilon_{ij} \tag{2}$$

where Y_{ij} is the accuracy of the i-th question for the j-th participant, $T1_j$ and $T2_j$ are the treatment dummy variables for the participant, D_i is the difficulty level of the question, u_j represents the random intercept for each participant, and ϵ_{ij} is the error term.

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

Variable	Coefficient	Std. Error	z-value	p-value
Intercept	0.58	0.03	22.50	< 0.001
Treatment	-0.23	0.03	-6.73	< 0.001
Treatment (Bias)	-0.20	0.04	-5.75	< 0.001
Difficulty	0.72	0.04	16.46	< 0.001
Treatment · Difficulty	0.09	0.06	1.65	0.099
Treatment (Bias) · Difficulty	-0.06	0.06	-1.02	0.307
Observations		5946		
No. Groups		991		
Log-Likelihood		-7450.44		

Notes. Group Var = 0.010. Scale = 0.7052. Random intercepts applied at participant level.

The mixed effects model shows significant effects of both the superforecasting LLM augmentation and the biased LLM augmentation conditions. Specifically, participants in the superforecasting LLM augmentation condition

showed an increase in accuracy as shown by a significant negative coefficient ($b = -0.23, p < .001$), and similarly for the biased LLM augmentation condition ($b = -0.20, p < .001$). This suggests that both treatment conditions were associated with an increase in accuracy measure compared to the control group. Additionally, the model indicated that the difficulty of the forecasting questions significantly influenced the dependent variable, with more difficult questions being associated with lower accuracy ($b = 0.72, p < .001$), as would be expected.

However, the interaction effects between the treatment conditions and question difficulty did not show statistically significant effects. The interaction between the superforecasting LLM augmentation condition and difficulty was not statistically significant ($b = 0.09, p = .099$), indicating that the effect of the treatment condition did not vary significantly with the difficulty level of the questions. Similarly, the interaction between biased LLM augmentation condition and difficulty also failed to reach statistical significance ($b = -0.06, p = .307$). These findings suggest that while treatment conditions and question difficulty independently influenced the outcome, their combined interaction effects did not significantly affect the outcome. In exploratory analyses, we also checked whether this pattern of results holds if we exclude the outlier Question 3. We found statistically significant effects in this non-preregistered analysis. Specifically, we found that the superforecasting LLM augmentation lead to higher accuracy on harder questions ($b = -0.139, p = .020$), with the converse being true for the biased LLM augmentation ($b = 0.177, p = .005$).

As preregistered, we used the the Benjamini-Hochberg (BH) procedure to adjust the p-values to control the false discovery rate for all p-values not already adjusted (e.g., by Tukey post-hoc tests) The original p-values for the preregistered analyses were 0.001, 0.103, 0.383, 0.099, and 0.307. We first sorted them in ascending order and then ranked accordingly. The adjusted p-values were computed using the Benjamini-Hochberg procedure, which calculates the adjusted p-value for the i-th hypothesis as $\min\left\{1, \frac{p_i \cdot m}{i}\right\}$, where p_i is the i-th p-value in the sorted list, m is the total number of hypotheses tested, and i is the rank of the p-value. The adjusted p-values are 0.005, 0.172, 0.383, 0.248, and 0.384, showing that our results are robust to this adjustment, with our first hypothesis remaining significant at p=0.005.

4 Discussion

Our investigation of LLM augmentation as a tool for human decision-making in the context of forecasting offers a number of results. Consider our finding that LLM augmentation, both the superforecasting and biased variants, significantly boosts individual forecasting accuracy relative to the control based on our preregistered analyses. Contrast this with the past finding that when GPT-4 forecasts binary-answer forecasting questions on its own, it substantially underperforms compared to human crowd performance, and in fact does not even outperform the no-information strategy of estimating 50% for each possible answer (Schoenegger and Park 2023). This suggests that, at least at the time of this paper's writing, LLM cognition may synergistically improve human cognition in the domain of forecasting when used as a human tool, even when LLM cognition by itself is somewhat ineffective. 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 studied in this context may be applicable.

Having a human-in-the-loop significantly improves LLM forecasts, propelling poor LLM forecasting performance to a level significantly higher than the human forecaster would have by themselves. However, this does not mean that this pattern will continue for the likely more capable AI systems of 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-year period between 1994 and 2004, and was much weaker than AI performance after 2004 (Kasparov 2010). If a similar pattern were true for 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 every capability relevant to forecasting.

We also found that both the superforecasting and the biased variants of LLM augmentation yields similar levels of forecasting accuracy increase, 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 biased augmentation on all six questions. Our result thus suggests that the main effect is not the model's prediction capabilities, but rather something else. Our result also contrasts with past studies' finding that adding idiosyncratic text to a prompt can have a strong effect on the output. Our contrasting finding suggests that at least in the domain of forecasting, the specific idiosyncrasies added to a prompt given to the augmenting LLM may play a lesser role than the past literature on prompt idiosyncrasies might suggest. Instead, the intrinsic reasoning capabilities of the models, irrespective of their idiosyncratic focus, seems to be the primary drivers of improved forecasting performance. This challenges the conventional understanding that much of LLMs' utility come from idiosyncratic prompt customization, which—in line with the 'correct answer' effect—suggests that LLM use may not increase society's diversity-of-thought, and may in fact decrease it (Doshi and Hauser 2023; Park, Schoenegger, and Zhu 2024).

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 does provide more accurate predictions, does improve performance at higher rates than the biased augmentation, and does outperform the biased 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 biased LLM augmentation. Thus, while the superforecasting LLM augmentation and control condition had a large number of their forecasters provide predictions that were so far off the actual value, the biased LLM augmentation had significantly higher accuracy by simply having higher predictions. In part, this may also stem from a confusion with the bitcoin USD spot price, where we find that forecasters in the biased 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 finding 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 biased LLM augmentation similarly improves upon the control while underperforming the more targeted superforecasting prompt.

Our further research question investigated the impact of LLM augmentation on low-skilled forecasters versus high-skilled forecasters. Past research on LLM augmentation generally agrees that it disproportionately bolsters the performance of low-performing workers among consultants (Dell'Acqua et al. 2023), call-center agents (Brynjolfsson, D. Li, and Raymond 2023), creative writers (Doshi and Hauser 2023), office workers writing memos (Noy and W. Zhang 2023), law school students writing exams (Choi and Schwarcz 2024), and coders (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 biased variants of LLM augmentation compared to the control. This raises the possibility that the private benefit of LLM augmentation due to its improvement of individual forecasting (e.g., a trader improving their performance via better market forecasting) may be significant, while its public benefit (e.g., market-making effects of stock market competition) may be less significant. 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 uniformly doing so, 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.

Overall, our results show the promise of augmenting human decision-making with LLMs. In both preregistered and exploratory analyses, we find significant accuracy improvements over a control that utilized a simpler non-forecasting LLM assistant. This shows that the augmentation ability of LLMs, ranging from providing answers outright to engaging with it in a back-and-forth manner can improve human performance and reasoning in contexts that are strictly outside the model's training data environment. As such, we argue that at the current margin, LLM augmentation may prove to be a valuable approach to integrating machine and human capabilities.

References

Abdurahman, Suhaib et al. (2023). "Perils and Opportunities in Using Large Language Models in Psychological Research". In.

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.

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, pp. 610–623. ISBN: 9781450383097. DOI: 10.1145/3442188.3445922.
Benjamin, Daniel M et al. (2023). "Hybrid forecasting of geopolitical events". In: *AI Magazine*.
Biderman, Stella et al. (2023). *Emergent and Predictable Memorization in Large Language Models*. arXiv: 2304.11158 [cs.CL].
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%5C_1cK.
Chang, Welton et al. (2016). "Developing expert political judgment: The impact of training and practice on judgmental accuracy in geopolitical forecasting tournaments". In: *Judgment and Decision making* 11.5, pp. 509–526.
Choi, Jonathan H and Daniel Schwarcz (2024). "AI Assistance in Legal Analysis: An Empirical Study". In: *Journal of Legal Education* 73. Forthcoming.
Dell'Acqua, Fabrizio et al. (2023). "Navigating the jagged technological frontier: Field experimental evidence of the effects of AI on knowledge worker productivity and quality". In: *Harvard Business School Technology & Operations Mgt. Unit Working Paper* 24-013.
Doshi, Anil R and Oliver Hauser (2023). *Generative artificial intelligence enhances creativity*. URL: https://ssrn.com/abstract=4535536.
Firat, Mehmet and Saniye Kuleli (2023). "What if GPT4 became autonomous: The Auto-GPT project and use cases". In: *Journal of Emerging Computer Technologies* 3.1, pp. 1–6.
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].
George, A Shaji and T Baskar (2023). "The Impact of AI Language Models on the Future of White-Collar Jobs: A Comparative Study of Job Projections in Developed and Developing Countries". In: *Partners Universal International Research Journal* 2.2, pp. 117–135.
Goyal, Tanya, Junyi Jessy Li, and Greg Durrett (2023). *News Summarization and Evaluation in the Era of GPT-3*. arXiv: 2209.12356 [cs.CL].
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.
Hazell, Julian (2023). *Spear Phishing With Large Language Models*. arXiv: 2305.06972 [cs.CY].
Heiding, Fredrik et al. (2023). "Devising and detecting phishing: Large language models vs. smaller human models". In: *arXiv preprint arXiv:2308.12287*.
Himmelstein, Mark, David V Budescu, and Ying Han (2023). "The Wisdom of Timely Crowds". In: *Judgment in Predictive Analytics*. Springer, pp. 215–242.
Jiao, Wenxiang et al. (2023). *Is ChatGPT a Good Translator? Yes with GPT-4 as the Engine*. arXiv: 2301.08745 [cs.CL].
Karger, Ezra, Pavel D. Atanasov, and Philip Tetlock (2022). "Improving judgments of existential risk: Better forecasts, questions, explanations, policies". In: *Questions, Explanations, Policies (January 17, 2022)*.
Kasparov, Garry (2010). "The chess master and the computer". In: *The New York Review of Books* 57.2, pp. 16–19.
Katz, Daniel Martin et al. (2023). "GPT-4 Passes the Bar Exam". In: *SSRN*. URL: https://ssrn.com/abstract=4389233.

Kinniment, Megan et al. (2023). "Evaluating language-model agents on realistic autonomous tasks". In: *arXiv preprint arXiv:2312.11671*.

Lichtenstein, Sarah and Baruch Fischhoff (1977). "Do those who know more also know more about how much they know?" In: *Organizational behavior and human performance* 20.2, pp. 159–183.

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.

Mannes, Albert E., Jack B. Soll, and Richard P. Larrick (2014). "The Wisdom of Select Crowds". In: *Journal of Personality and Social Psychology* 107.2, p. 276.

Mellers, Barbara et al. (2015). "The psychology of intelligence analysis: Drivers of prediction accuracy in world politics." In: *Journal of Experimental Psychology: Applied* 21.1, p. 1.

Moore, Don A and Paul J Healy (2008). "The trouble with overconfidence." In: *Psychological review* 115.2, p. 502.

Naveed, Humza et al. (2023). *A Comprehensive Overview of Large Language Models*. https://github.com/humza909/LLM_Survey.git.

Nori, Harsha et al. (2023). *Capabilities of GPT-4 on Medical Challenge Problems*. arXiv: 2303.13375 [cs.CL].

Noy, Shakked and Whitney Zhang (2023). "Experimental evidence on the productivity effects of generative artificial intelligence". In: *SSRN*. URL: https://ssrn.com/abstract=4375283.

OpenAI (2023). *New models and developer products announced at DevDay*. https://help.openai.com/en/articles/8555510-gpt-4-turbo.

Park, Peter S. (2022). "The evolution of cognitive biases in human learning". In: *Journal of Theoretical Biology* 541, p. 111031.

Park, Peter S., Simon Goldstein, et al. (2023). *AI Deception: A Survey of Examples, Risks, and Potential Solutions*. arXiv: 2308.14752 [cs.CY].

Park, Peter S., Philipp Schoenegger, and Chongyang Zhu (2024). "Diminished diversity-of-thought in a standard large language model". In: *Behavior Research Methods*, pp. 1–17.

Park, Peter S. and Max Tegmark (2023). *Divide-and-Conquer Dynamics in AI-Driven Disempowerment*. arXiv: 2310.06009 [cs.CY].

Peng, Sida et al. (2023). "The impact of ai on developer productivity: Evidence from github copilot". In: *arXiv preprint arXiv:2302.06590*.

Sallam, Malik et al. (2023). "ChatGPT applications in medical, dental, pharmacy, and public health education: A descriptive study highlighting the advantages and limitations". In: *Narra J* 3.1, e103–e103.

Schoemaker, Paul JH and Philip E Tetlock (2016). "Superforecasting: How to upgrade your company's judgment". In: *Harvard Business Review* 94.5, pp. 73–78.

Schoenegger, Philipp and Peter S. Park (2023). *Large Language Model Prediction Capabilities: Evidence from a Real-World Forecasting Tournament*. arXiv: 2310.13014 [cs.CY].

Shen, Zhiqiang et al. (2023). "SlimPajama-DC: Understanding Data Combinations for LLM Training". In: *arXiv preprint arXiv:2309.10818*.

Solaiyappan, Siddharth et al. (2023). "Utilizing Machine Learning Algorithms Trained on AI-generated Synthetic Participant Recent Music-Listening Activity in Predicting Big Five Personality Traits". In.

Summers, Lawrence H and Steve Rattner (2023). *Larry Summers on who could be replaced by AI [Interviewed by Bloomberg TV's David Westin]*. URL: https://www.youtube.com/watch?v=8Epl9yAu0gk.

Sutton, Rich (2023). *AI succession [Youtube video of talk]*. World Artificial Intelligence Conference in Shanghai. URL: https://www.youtube.com/watch?v=NgHFMolXs3U.

Tetlock, Philip E and Dan Gardner (2016). *Superforecasting: The art and science of prediction*. Random House.

Tetlock, Philip E, Barbara A Mellers, and J Peter Scoblic (2017). "Bringing probability judgments into policy debates via forecasting tournaments". In: *Science* 355.6324, pp. 481–483.

Vaswani, Ashish et al. (2017). "Attention is All You Need". In: *Advances in Neural Information Processing Systems* 30.

Vemprala, Sai et al. (2023). "Chatgpt for robotics: Design principles and model abilities". In: *Microsoft Auton. Syst. Robot. Res* 2, p. 20.

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.

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*.

5 Appendix

See Figure 5 for the full treatment prompt of the biased 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 in a biased way, drawing on baserate neglect and overconfidence.

Treatment Prompt (Bias)

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 Biased 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 biased LLM augmentation condition.

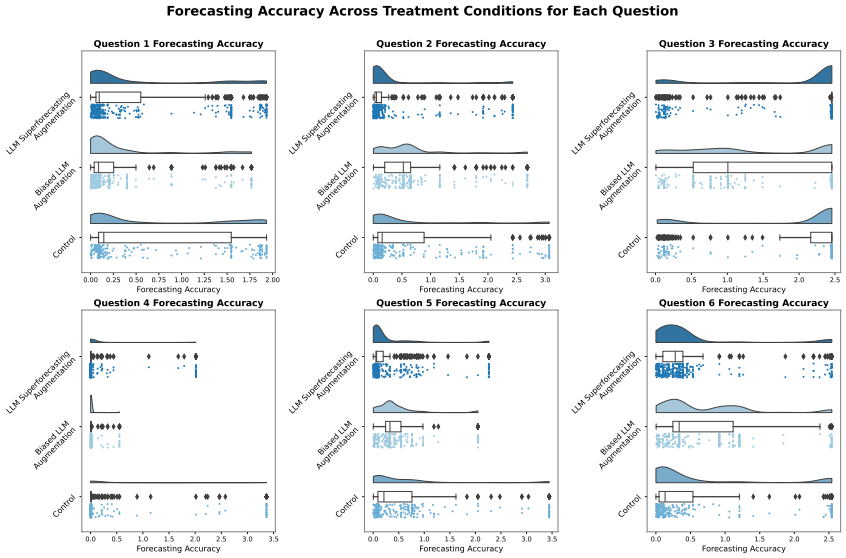

Figure 6: Raincloud plots of forecasting accuracy by condition for each question.

See Figure 7 for CDF plots of forecasting accuracy by condition for each question. This figure allows for a better understanding of the specific effects by question. For example, it shows that the majority of the accuracy advantage that the biased LLM augmentation condition enjoys over the other two conditions is due to having less predictions that were at the winsorized bound.

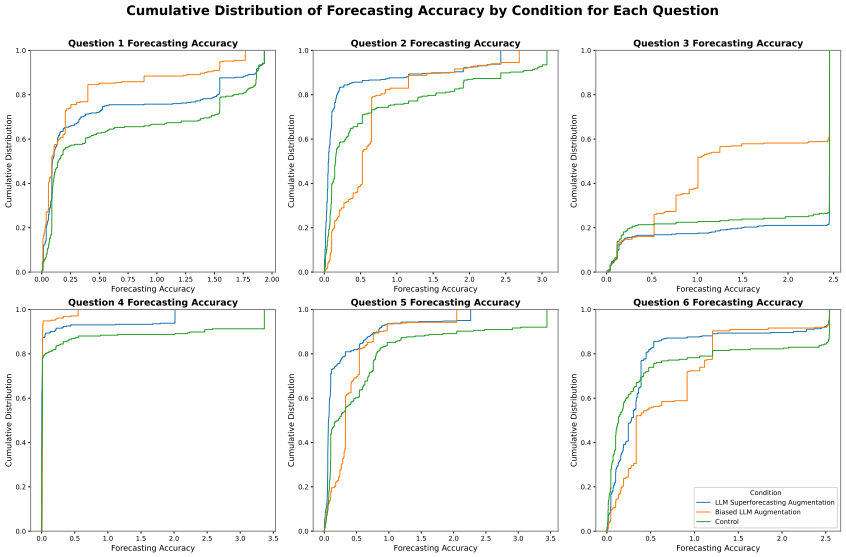

Figure 7: CDF plots of forecasting accuracy by condition for each question.