significantly outperform the no-information forecasting strategy of uniform random guessing (Schoenegger and Park 2023). However, more recent work has found that aggregating a set of diverse LLM forecasts (Schoenegger et al. 2024b) or retrieval-augmented (RAG) systems that enable the model to access additional information (Halawi et al. 2024) can approach human-level performance. Moreover, this previous work only investigated the effect of machine forecasts produced directly by the model, without incorporating human input that allows a continuous back-and-forth between forecasters and the machine. It is reasonable to expect that human-LLM hybrid forecasts—the object of study in the present paper—might outperform the results of the LLM operating by itself if it was set up properly. 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 our human forecasters can engage in a back-and-forth dialogue with a specifically instructed forecasting 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 struggle most with: be it synthesizing data, making coherent forecasts, or attaching numbers to intuitions, thus increasing the potential effect of this augmentation. Importantly, LLMs specifically prepared for this task via system prompts may be especially beneficial. This motivates our first research question and accompanying hypothesis, testing whether we find an aggregate accuracy improvement of specially prompted frontier LLM augmentations compared to a control condition that has access to a lower-powered LLM that does not provide forecasting assistance.

We test two treatments, one where the human has access to an LLM with a 'superforecasting' (Tetlock and Gardner 2016) prompt that draws on well-studied principles of good forecasting practice. 'Superforecasting' is a term that describes a set of features that exceptionally accurate human forecasters have shown to possess, which, at least in part, contribute to their superior prediction capabilities in human forecasting tournaments. In this context, the model is asked to provide assistance that breaks down complex problems into smaller ones, distinguishes degrees of doubt, and aims to identify errors in its own reasoning. We also set up a second advanced LLM with a specific prompt, aimed at produced inaccurate forecasts. We specifically instructed the model to exhibit the biases of base rate neglect and overconfidence, thus resulting in a noisy forecasting assistant. Both models are instructed to assist forecasters in whatever way is requested, ranging from providing point estimates to offering feedback on forecasts. We compare both treatments and the control condition to each other, allowing for a potential ordering of effects. This allows us to test whether a back-and-forth with an advanced LLM that provides direct and actionable forecasting advice outperforms a much weaker LLM baseline that does not provide forecasting advice. We predicted that the superforecasting LLM augmentation would outperform the noisy 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 (noisy) LLM augmentation and the control.

Recent work in other areas 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, Li, and Raymond 2023), creative writers (Doshi and Hauser 2023), office workers who write memos (Noy and Zhang 2023), law school students who write exams (Choi and Schwarcz 2024), and programmers (Peng et al. 2023). The underlying reason differs by context, but the general suggestion is that low-performing individuals can increase their performance by substituting LLM output for human output, which is more likely to improve results if one's own output is not as high-quality. 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). One potential explanation for such an effect may be that low-performing individuals might be comparatively less able to spot LLM weaknesses and failure modes, whereas those more familiar with the task could selectively use the LLM augmentation to greater effect. This heterogeneity of results suggests that any effects of LLM augmentation on forecasting are likely to be distinct across the skill distribution, with lower-skill forecasters potentially relying to a greater degree on LLM augmentation, which may help alleviate biases in their predictions that would otherwise have led them to make badly calibrated judgments. This motivates our second hypothesis, which directly tests whether the LLM augmentation has disparate impacts on forecasters of different skill levels. In line with much of the previous literature, we predicted a greater positive effect on lower-skill forecasters.

> **Null Hypothesis 2**: The effect of the superforecasting (noisy) 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 collect data allowing us to look at 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 aggregated forecast that is more accurate than the majority of forecasts given by most individuals,

even across heterogeneous types of forecasters who may have different skill levels (Budescu and Chen 2015; Mannes, Soll, and Larrick 2014). However, this aggregation tends to be most effective when there is a diversity of independent forecasts, not if the forecasts share a common source of variation and are thus intercorrelated. If the LLM augmentations anchor many human forecasters on the same or very similar point forecast for a given question, it could reduce the value of aggregation as the independence of forecasts is reduced, inducing a potential group think effect. If this is the case, this would provide a substantial source of concern for applications of LLM augmentations in practice. To look at this, we test whether LLM augmentation homogenizes forecasts in this way, motivating our third hypothesis, where we predicted a reduction in group-level accuracy.

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

Finally, we compare the effect LLM forecasting augmentation has on prediction performance on questions of different difficulty levels. There are a number of reasons why the difficulty of the forecasting question may be an important factor. If questions are especially difficult, forecasters may be more likely to simply defer to any machine prediction directly, without further investigation and critique. If machines are then individually worse or better than humans, this might play out in a difficulty effect. Conversely, very easy questions may be such that forecasters do not bother asking the LLM for input and instead rely on their own forecasts in which they might have relatively high confidence. There could also be a more complicated interplay of question difficulty with other factors that may lead to an ameliorating effect of performance increasing and performance reducing aspects. This set of questions motivates our last hypothesis, where we did not have a specific directional prediction.

> **Null Hypothesis 4**: There is no difference in the effect of the superforecasting (noisy) 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 conduct after having seen the data or having done other analyses. This study received ethics approval prior to data collection.[2]

We recruited a total of 1,152 participants from Prolific, an online research platform that gives researchers access to people willing to participate in research in exchange for a participation fee. For participating in our study, participants were paid $5 for participation and could earn an additional $100 based on their accuracy. We paid three such accuracy prizes to randomly selected participants who scored in the top-10 of forecasters. We used this level of randomization to account for incentive concerns of paying out prizes only to the top performers might then be likely to extremize their predictions (Witkowski et al. 2023) by choosing values significantly above or below their true beliefs, thus distorting the incentive compatibility of the forecast elicitation. We preregistered the following a priori power analysis to determine the sample size of our study: 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 noisy 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 1,152 participants, meeting our goal.

We collected participant forecasts on a set of six forecasting questions that ranged from questions on finance, geopolitics, and cryptocurrency to ones on aviation, artificial intelligence, and foreign exchange. All six questions had continuous prediction variables, ranging from asset prices to numbers of refugees, where participants could input any number without restrictions. We chose a diverse set of questions to account for variation in question difficulty and familiarity, while ensuring that our outcome variable contributes rather than distracts from the generalizability of results. We also ensured that all questions were resolvable quickly after the cutoff date to allow for timely payouts of accuracy incentives for participants. The question set was drawn from an early question set used in the Forecasting Proficiency Test (Himmelstein et al. 2024). For a full list, see Table 1. Data collection happened on November 21, 2023, over five weeks prior to forecast question resolution.

Our main outcome variable is forecasting accuracy. Our accuracy measure is the error between participant forecasts and the true value of the forecasted question. We computed the error 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 $D_i = |F_i - A_i|$. To ensure participant comprehension, participants read a detailed explanation of this measure of accuracy, as well as an example, and then completed a one-question quiz on it, without which they were not able to continue in the experiment. Throughout the paper, unless specifically specified otherwise, when we refer to 'accuracy', we mean the error rating arrived by using absolute differences between the forecast and the truth value. As such, in all our analyses, lower values indicate higher accuracy, and higher values indicate lower accuracy.

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

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

To account for outliers in our data, which we expected due to the free entry data collection of forecasting problems that permit substantial uncertainty, we conducted an initial winsorisation process of accuracy values at the 5% levels by removing all values at the bottom 5% and top 5%.[3] Then, we standardized the values across questions by dividing them by the standard deviation of the control group for the respective question, allowing for inter-question comparability in accuracy scores. Lastly, we conducted a second winsorisation step, this time at the level of 3 standard deviations.

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?

Our secondary variables of question difficulty and forecaster skill were collected as follows: A randomly selected 10% of control group participants were tasked not only with providing forecasts for each question but also with rating their perceived difficulty for each question on a 5-point Likert scale ranging from 'Very easy' to 'Very difficult'. Questions 2 and 3 received the highest difficulty ratings and were therefore identified as being the most challenging in our analyses, to be compared with the other four questions.

Prior to the main forecasting tasks, participants were also asked a series of smaller, lower-effort forecasting questions. These questions included binary predictions (providing the probability that an event happens by a certain time) and intersubjective forecasts (predicting the average forecast of others on a question by answering 'What is the average probability that participants in this study give on the above question?'), to evaluate their 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 based on the two rankings: The top half of participants based on this composite measure was classified as relatively higher-skill forecasters. This forecasting skill measure is an abridged attempt at capturing two dimensions of forecasting skill, accuracy and intersubjective accuracy (Himmelstein et al. 2024). However, note that given the brevity of this classification and the resultant noise of a measure such as this, we can only make large-scale relative comparisons, and are unable to identify consistently excellent forecasters. For the set of questions used for the skill measures, see Table 2.

Participants were randomly selected into one of three conditions—Treatment (including the superforecasting prompt), Treatment (Noise) (including a prompt instructing the model to respond with a variety of biases, resulting in noisy assistance), and Control—with a a participant allocation ratio of 1.5/1/1. We presented

[3]We report the following deviation from our preregistered analysis plan: We applied the 5% winsorisation step to all groups, rather than solely to the control group. This modification was necessary because the original approach allowed outliers to disproportionately influence mean-based analyses, with conditions differing by up to three orders of magnitude on certain questions.