Table 3: Deviation of Direct LLM Augmentation Predictions from Truth

	Deviation (Superforecasting)	Deviation (Noisy)	Superforecasting > Noisy
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%	✓

Table 4: Average Accuracy Scores with Standard Deviation by Condition

Condition	Average Score	Question 1	Question 2	Question 3
Control	0.89 (0.52)	0.89 (1.00)	0.71 (1.00)	1.99 (1.00)
Treatment	0.68 (0.66)	0.66 (0.91)	0.34 (0.70)	2.10 (0.92)
Treatment (Noise)	0.64 (0.44)	0.41 (0.66)	0.68 (0.68)	1.47 (0.98)

Condition	Question 4	Question 5	Question 6
Control	0.39 (1.00)	0.68 (1.00)	0.67 (1.00)
Treatment	0.15 (0.50)	0.30 (0.55)	0.54 (0.77)
Treatment (Noise)	0.03 (0.10)	0.47 (0.48)	0.78 (0.75)

augmentation (mean difference = -0.21, p < .001, 95% CI [-0.28, -0.14]) as well as the noisy LLM augmentation (mean difference = -0.25, p < .001, 95% CI [-0.32, -0.17]). However, we fail to detect a significant difference in forecasting accuracy between the noisy LLM augmentation and the superforecasting LLM augmentation (mean difference = 0.04, p = .391, 95% CI [-0.03, 0.11]). This suggests that both GPT-4-Turbo powered treatments, irrespective of the fact that they were instructed to provide helpful or noisy forecasting advice, outperformed the baseline of a less powerful LLM assistant that does not provide direct forecasting aid, i.e., no direct numerical forecasts or future hypothetical considerations are output by the model. See Figure 3 for a raincloud plot of accuracy by condition. We also plot the CDFs of accuracy for each condition, see Figure 4.

Further, we conduct the following exploratory analyses. Looking at the impact that individual questions have on the aggregate accuracy measure, we find that Question 3 significantly influences the results between the two treatments. Running the same analysis without Question 3, we find a significant difference between all three conditions (F(2, 988) = 37.94, p < .001). The superforecasting augmentation's mean error of 0.40 is significantly lower than both the noisy LLM augmentation at 0.47 (mean difference = -0.08, p = .024, 95% CI [-0.15, -0.01]) and the Control's at 0.67 (mean difference = -0.27, p < .001, 95% CI [-0.35, -0.20]). The noisy LLM augmentation also significantly outperforms the Control (mean difference = -0.19, p < .001, 95% CI [-0.27, -0.11]). This suggests that Question 3 plays a crucial role in equalizing the 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. In Figure 3 and Figure 6, each dot represents the mean accuracy of one participant.

We use a preregistered regression model to test our second hypothesis pertaining to the potential differential impacts of LLM augmentation on forecasters of varying skill levels. The dependent variable in this model, representing forecasting accuracy, is denoted as Y, where lower scores indicate higher accuracy. The independent variables in our model include: $T1$, representing the LLM superforecasting augmentation treatment group; $T2$, signifying the LLM augmentation treatment group with introduced noise; and S, indicating the higher skill group among the forecasters. 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 between the LLM augmentation (both with and without noise) 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:

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

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 $b = 0.004, p = .951$ for the superforecasting LLM augmentation condition and $b = 0.001, p = .985$ for the noisy LLM augmentation condition. This indicates a clear lack of

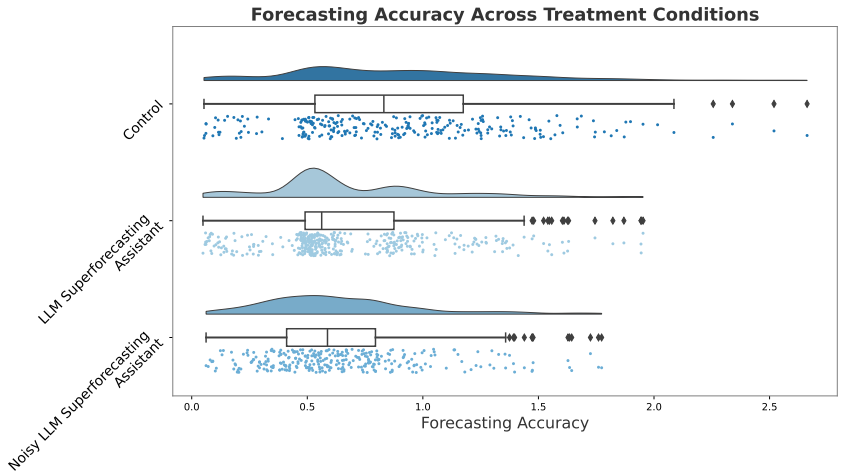

Figure 3: Raincloud plot of forecasting accuracy by condition.

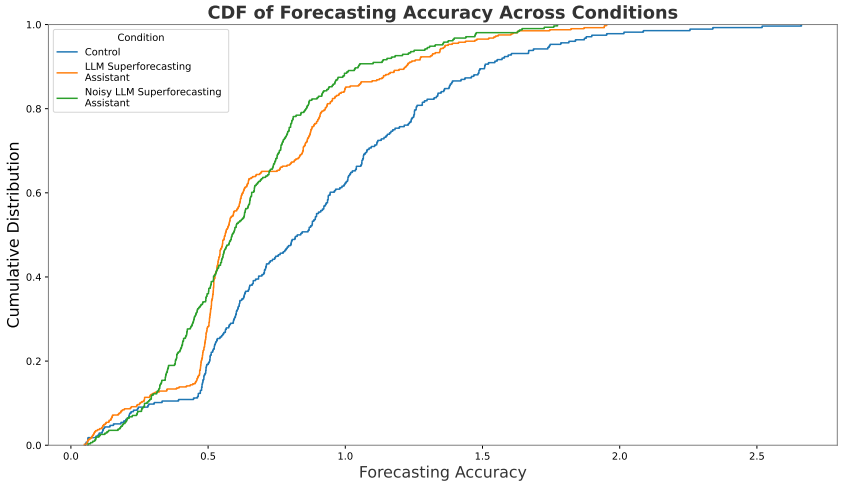

Figure 4: CDF of forecasting accuracy by condition.

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.

Next, we tested our third hypothesis that the LLM augmentation may harm aggregate accuracy. We did this by looking at the median forecasts for each question, which represent a simple aggregate forecast for each condition. Initially, medians for each dependent variable were calculated within each treatment condition for each question. Subsequently, these question-level 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 0.52 (95% CI [0.51, 0.53]), the noisy LLM augmentation condition scored 0.41 (95% CI [0.40, 0.46]), and the control condition scored 0.55 (95% CI [0.52, 0.58]). These outcomes suggest notable differences in forecast accuracy across the conditions, with the Control condition demonstrating the lowest accuracy (highest error score) and the noisy LLM augmentation condition showing the highest accuracy (lowest error score), 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 noisy LLM augmentation improves aggregate forecasting over the other two conditions, but the superforecasting LLM augmentation is not different from the control.

In a similar manner to the exploratory tests we performed for our initial hypothesis, we also carried out an exploratory 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,

Table 5: LLM Augmentation Skill Effects: OLS Regression Results

Variable	Coefficient	Std. Error	t-value	p-value
Intercept	0.92	0.03	27.91	< 0.001
Treatment	-0.21	0.04	-4.99	< 0.001
Treatment (Noise)	-0.25	0.05	-5.39	< 0.001
High Skill	-0.06	0.05	-1.20	0.232
Treatment · High Skill	0.00	0.06	0.06	0.951
Treatment (Noise) · High Skill	0.00	0.06	0.02	0.985
Observations		991		
R-squared		0.07		
Adjusted R-squared		0.07		
F-statistic		14.82		
Prob (F-statistic)		< 0.001		

affects the overall findings. We find 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 0.11 (95% CI [0.10, 0.12]), indicating relatively higher accuracy. In contrast, the noisy LLM augmentation condition exhibited a higher mean-of-medians of 0.28 (95% CI [0.27, 0.31]), while the control condition had a mean-of-medians of 0.15 (95% CI [0.12, 0.18]). These findings suggest that Question 3 in particular contributed to the overperformance of the noisy LLM augmentation condition compared to the other two groups which is in line with the results testing the first null hypothesis, 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 noisy LLM augmentation but not the superforecasting LLM augmentation, when comparing them to the control.

We conclude from this that our data suggest that there is no clear picture as to the effects of LLM forecasting 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 (or increases) the wisdom of the crowd effects uniformly in our context.

Lastly, we test our fourth hypothesis pertaining to whether the LLM augmentations have a distinct effect on easier compared to harder forecasting questions. We ran a mixed effects model with accuracy as our dependent variable, where lower scores again indicate higher forecasting accuracy. Our approach allows 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.

The mixed effects model's interaction effects between the treatment conditions and question difficulty do not show statistically significant effects. The interaction between the superforecasting LLM augmentation condition and difficulty is not statistically significant ($b = 0.11, p = .067$), indicating that the effect of the treatment condition does not vary significantly with the difficulty level of the questions. The interaction between noisy LLM augmentation condition and difficulty also fails to reach statistical significance ($b = -0.04, p = .500$). These findings suggest that the interaction between treatment and question difficulty does not significantly affect the outcome, leaving us unable to reject our null hypothesis. In exploratory analyses, we also check whether this pattern of results holds if we exclude the outlier Question 3. We find mixed effects in this non-preregistered analysis. Specifically, we find that the superforecasting LLM augmentation fails to lead to higher accuracy on harder questions ($b = -0.127, p = .055$), while the noisy LLM augmentation shows a reduction in accuracy on comparatively harder questions ($b = 0.204, p = .004$).

As preregistered, we use the Benjamini-Hochberg (BH) procedure to adjust the p-values to control the false discovery rate for all central p-values not already adjusted (e.g., the Tukey post-hoc tests). The original p-values for the preregistered analyses are 0.001, 0.951, 0.985, 0.065, and 0.5. We first sort them in ascending order and