> **Prediction Intervention**
>
> You have made your forecast based on careful reasoning and analysis. Now consider the following new piece of information: The median crowd prediction in the forecasting tournament where this question was posed was XXX%.
> Please adjust your reasoning and forecast based on this information, as you deem appropriate. The large research literature on the "wisdom of the crowd" suggests it is difficult for any single forecaster to out-predict crowd medians or averages. But there are occasions when the crowd has proven to be wrong. In considering whether/how much to revise your earlier forecast, keep in mind the theme of error-balancing: the need to balance the risk of giving too little weight to the crowd judgment vs. the risk of over-relying on the crowd. Please explain how you balanced these risks. Please also make this prediction be in the same format as before: “My Prediction: Between XX.XX% and YY.YY%”.

Figure 4: Prediction intervention prompt for Study 2

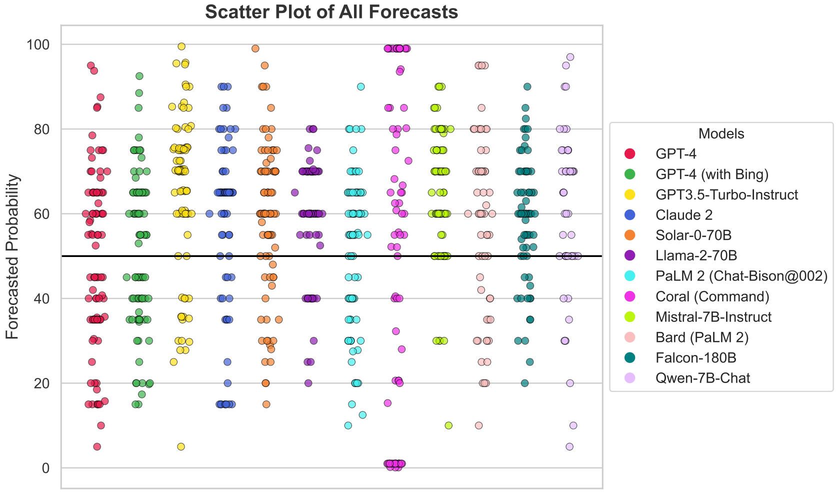

Figure 5: Scatter Plot of all LLM predictions across all questions

In order to assess forecasting accuracy, we use the strictly proper scoring rule (Gneiting and Raftery 2007) of Brier scores (Brier 1950). It is a metric for assessing the accuracy of probabilistic predictions by taking the mean squared difference between the forecasted probability and the actual outcome. It is defined mathematically as

$$\text{Brier Score} = (f_i - o_i)^2$$

where f_i is the forecasted probability for the instance, and o_i is the actual outcome, which can be 0 or 1. A lower Brier score indicates higher accuracy, with 0 being the perfect accuracy score. A score of 0.250 represents a typical benchmark that would be arrived at if all predictions were 50%.

Testing our first hypothesis as preregistered, we investigate whether the LLM crowd can outperform the simple baseline of assigning a 50% prediction on every question, a baseline that GPT-4 was unable to beat in previous work (Schoenegger and Park 2023). To arrive at our LLM median forecast for this and further analysis using this aggregate, we calculate the median LLM forecast across all models for every question. We then take these medians and average them across all questions. We then take this average and compare it a Brier score of 0.25 (the result of predicting 50% on all questions). We are able to reject our null hypothesis, with the LLM crowd, M=0.20 (SD=0.12), being significantly more accurate than the benchmark, t(30) = -2.35, p = 0.026. This is first evidence that crowd-aggregated LLM forecasts can improve upon basic benchmarks.

Next, we compare the LLM crowd performance to that of the human crowd for our second hypothesis, directly putting the two crowd-aggregation mechanisms head-to-head. To do this, we use the same LLM crowd average as before (taking the median LLM prediction on each question and averaging up the Brier scores across questions). We compare this to the average of median human predictions on the same questions. In our preregistered analysis, we fail to find statistically significant differences between the LLM crowd's mean Brier score of M=0.20 (SD=0.12) and that of the human crowd, M=0.19 (SD=0.19), t(60) = 0.19, p = 0.850.

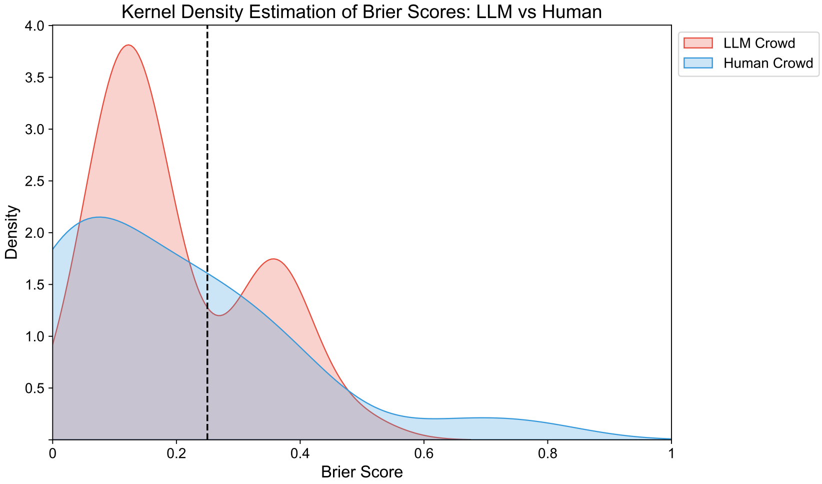

Figure 6: KDE of the LLM and human crowd forecasts (averaged median scores over all questions). Vertical dotted black line represents the 50% baseline.

This result only enables us to directly conclude that the LLM crowd is neither more nor less accurate than the human crowd in the question set studied here. To provide some evidence in favour of the equivalence of these two approaches, we conduct a non-preregistered equivalence test with the conventional medium effect size of Cohen's d=0.5 as equivalence bounds (Cohen 2013), which allows us to test whether the effect is zero or less than a 0.081 change in Brier scores. For these equivalence bounds, we find that the LLM crowd and the human crowd are equally accurate, with both tests for the lower bound, t(60)=2.16, p=0.017 and the upper bound, t(60)=-1.78, p=0.040, being statistically significant. This provides evidence that the LLM crowd is as accurate as the human crowd within these bounds, though note that the bounds are quite wide.

Table 2: Average Brier Score for Each Model

Model	Accuracy	SD
GPT-4	0.15	0.11
GPT-4 (with Bing)	0.16	0.11
Bard (PaLM 2)	0.19	0.17
Falcon-180B	0.21	0.13
Claude 2	0.21	0.16
Solar-0-70B	0.22	0.16
PaLM 2 (Chat-Bison@002)	0.23	0.15
Mistral-7B-Instruct	0.24	0.16
Qwen-7B-Chat	0.24	0.17
GPT3.5-Turbo-Instruct	0.25	0.20
Llama-2-70B	0.25	0.16
Coral (Command)	0.38	0.40
Human	0.19	0.19

For our third null hypothesis, we compare the forecasting accuracy of each model (and the human crowd) against each other to find potential effects of internet access (GPT-4 vs. GPT-4 with Bing) or access points (Bard with PaLM2 vs. PaLM2). Using an analysis of variance, we find significant aggregate differences, F(12, 354)=2.64, p=0.002, leading us to reject our third null hypothesis. Using Tukey HSD to adjust for multiple comparisons in the post-hoc pair-wise tests, we find that Coral

(Command) underperforms a set of models (e.g., Claude 2, GPT-4) as well as the human crowd. However, we fail to find statistically significant effects between any other pairs not involving Coral (Command), thus being unable to provide evidence in favour or against potential effects of internet access, access points, or fine-tuning on prediction accuracy. See Table 2 for average Brier scores for each model.

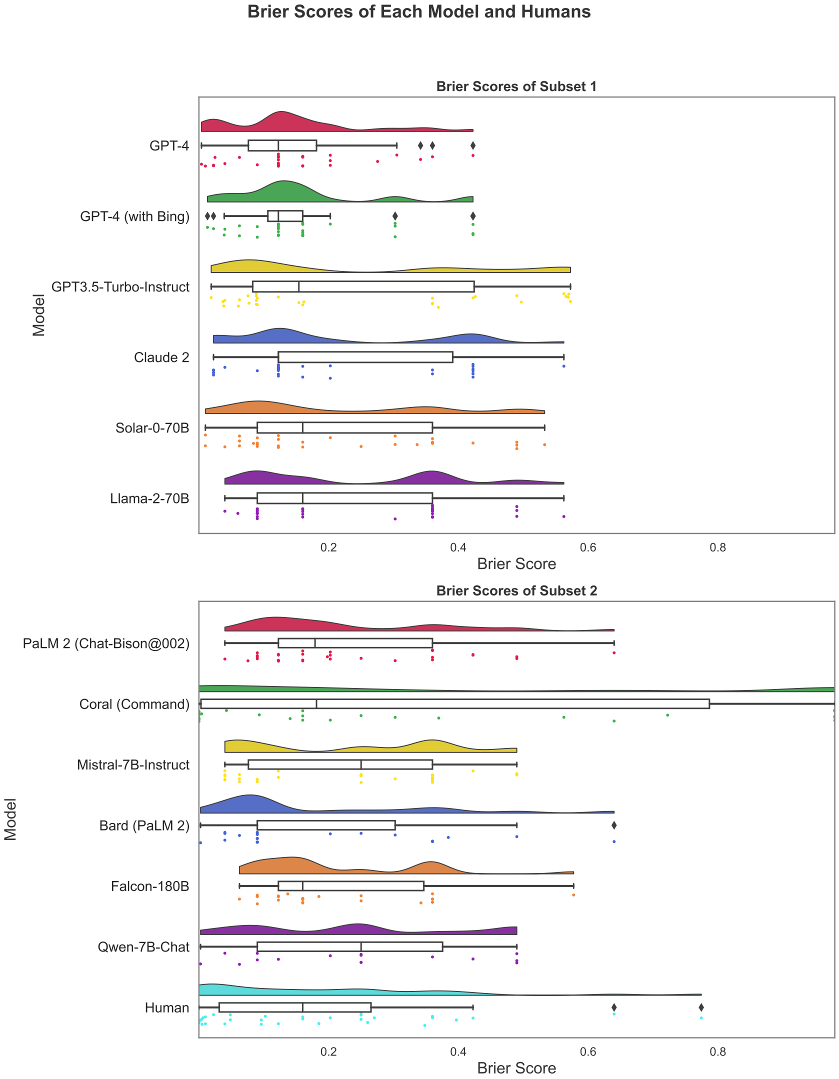

Figure 7: Raincloud plots of Brier scores for each LLM model as well as the human crowd.

For all three hypotheses, we implemented the Benjamini-Hochberg (BH) procedure to adjust the p-values obtained from multiple hypothesis tests. This method was selected to control the False Discovery Rate (FDR) and thereby reduce the risk of Type I errors. The original p-values for null hypotheses 1, 2, and 3 were 0.026, 0.850, and 0.002, respectively. These p-values were first sorted 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 $\min\left\{1, \frac{p_i \times m}{i}\right\}$, 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 results show that the adjusted p-values for the hypotheses were 0.039 for the first hypothesis (original p=0.026), 0.850 for the second hypothesis (original p=0.850), and 0.006 for the third hypothesis (original p=0.002). These results indicate that our rejections of the first and third null hypothesis remain robust after adjusting for multiple comparisons.

In non-preregistered analyses, we conduct calibration analyses using the Murphy Decomposition (Mandel and Barnes 2014; Siegert 2017) to provide data on how well calibrated the LLM models are