in this context, i.e., how reliably their probability estimates match the fraction of real outcomes. In Figure 8, calibration curves for each model and their aggregate are plotted against the ideal 45-degree dotted line. This dotted line represents perfect calibration, where predicted probabilities match observed frequencies. Deviations from this line indicate calibration errors: curves above the line suggest underconfidence (predicting events as less likely than they actually are), while those below indicate overconfidence (predicting events as more likely than they actually are). Figure 8 visually represents how closely the models' predictions align with actual outcomes. We also calculate the Calibration Index (CI), which quantifies this deviation, with lower values indicating better calibration. CI is calculated using the formula:

$$CI = \frac{1}{N} \sum_{k=1}^{K} N_k (f_k - o_k)^2$$

where N is the total number of forecasts, K the number of bins, N_k the number of forecasts in bin k, f_k the mean forecast probability in bin k, and o_k the observed relative frequency in bin k. This weights each bin's contribution to the Calibration Index (CI) by the number of forecasts it contains. This approach ensures that bins with more forecasts, which provide a more statistically reliable estimate of forecasting accuracy, have a proportionately greater impact on the overall CI.

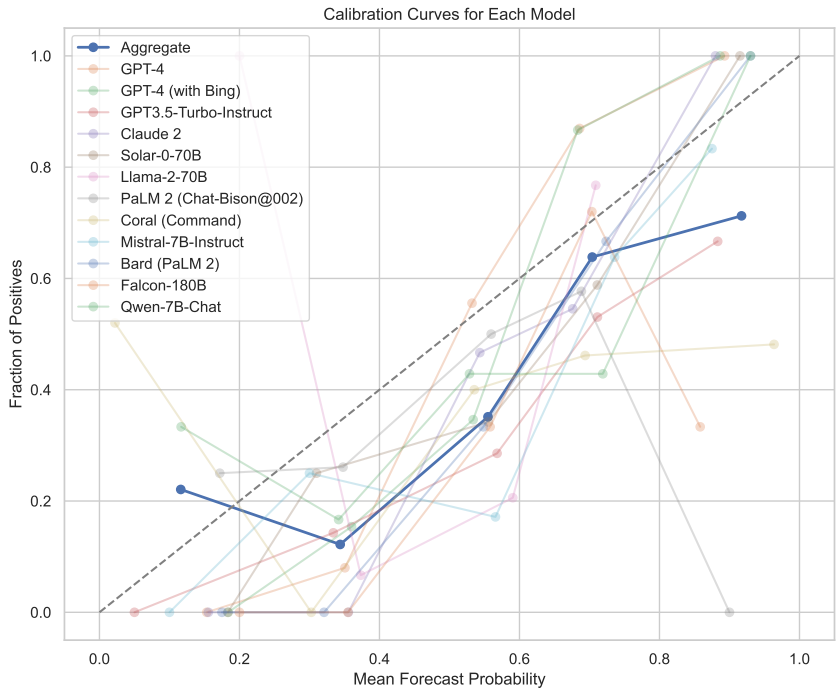

Figure 8: Calibration plot for all LLM models as well as the aggregate (bolded)

Our results demonstrate poor calibration of most models and overconfidence of the aggregate, suggesting that models overpredict outcomes compared to their actual rate of occurrence, see Figure 8. This is in line with the finding that we find a acquiescence bias of LLMs on a question set where less than half of questions resolve positively. We also find generally poor calibration across all models. However, there are substantial differences in the CI scores, with some models having substantially better calibration than others, see Table 3. This suggests that a further line of research may build upon improving calibration of models in an attempt to improve machine prediction capabilities and reliability further.

Table 3: Calibration index values for all LLM models.

Model	Calibration Index
Falcon-180B	0.027
Qwen-7B-Chat	0.055
PaLM 2 (Chat-Bison@002)	0.068
Bard (PaLM 2)	0.071
Llama-2-70B	0.071
GPT-4	0.075
Mistral-7B-Instruct	0.080
Solar-0-70B	0.081
Claude 2	0.082
GPT-4 (with Bing)	0.088
GPT3.5-Turbo-Instruct	0.106
Coral (Command)	0.212
Aggregate	0.041

3.2 Study 2

For Study 2, we collected a total of 186 primary forecasts and 186 updated forecasts from both frontier models (GPT-4 and Claude 2) over the 31 binary questions studied. Neither model refused to provide a forecast or failed to respond to our querying.

First, we test whether exposure to the human crowd median improves model accuracy. We are able to reject the first null hypothesis of Study 2 for both models: For GPT-4, there is a statistically significant difference in Brier Scores before and after exposure to the human median, with an average Brier score for the primary forecast of 0.17 (SD: 0.13) and an updated score of 0.14 (SD: 0.11), $p = 0.003$. For Claude 2, we also find a statistically significant difference in Brier Scores before and after exposure to the human median, improving from 0.22 (SD: 0.19) to 0.15 (SD: 0.14), $p < 0.001$. This suggests that the provision of human cognition in the form of crowd forecasts can improve model prediction capabilities.

We also find that, testing our second hypothesis, the size of the prediction interval narrows after exposure to human crowd predictions that lie within the probability range provided by the model, as would be predicted by theory: The prediction intervals for GPT-4 become significantly narrower after exposure to the human median, ranging from an average interval size of 17.75 (SD: 5.66)to 14.22 (SD: 5.97), $p < 0.001$. The prediction intervals for Claude 2 also become significantly narrower after exposure to the human median forecast, narrowing from 11.67 (SD: 4.201 to 8.28 (SD: 3.63), $p < 0.001$. This suggests that the models appropriately reduce their prediction uncertainty when the human forecast is already included in the LLM's, incorporating this additional information and adjusting their uncertainty. See Figure 9 for a graphical illustration of LLM forecasts for either model before and after exposure to the human forecasts.

Lastly, with respect to our third hypothesis, we analyse whether LLMs' updates are proportional to the distance between their point forecast and that of the human benchmark. We are able to reject our null hypothesis for both models, finding significant correlation between the initial deviation and the magnitude of forecast adjustment for GPT-4, $r=0.88$, $p < 0.001$ as well as for Claude 2 $r=0.87$, $p < 0.001$. This suggests that models move their predictions roughly in accordance with how large the difference between their prediction and the human median is.

As in Study 1, we use the Benjamini-Hochberg procedure for controlling multiple comparisons, given our three hypotheses each tested for each model, resulting in six tests. The original p-values were [0.001, 0.001, 0.001, 0.001, 0.001, 0.003]. After applying the Benjamini-Hochberg adjustment, the p-values were [0.006, 0.006, 0.006, 0.006, 0.006, 0.003], all of which were below the 0.05 FDR threshold. This indicates that, post-adjustment, the results from all tests remained statistically significant.

We also conduct the following exploratory analysis. Instead of comparing the LLM forecast after having been exposed to the human median to the LLM forecast before this exposure as preregistered, we compare this updated prediction to a simple average of the machine and human predictions as a

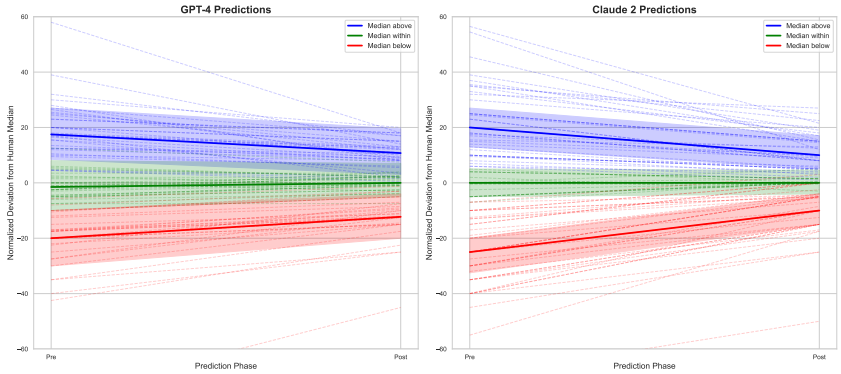

Figure 9: LLM forecasts for GPT-4 (left) and Claude 2 (right) before and after exposure to the human forecast. Colours distinguish first forecasts above, below, or within 20 percentage points of the human median forecast. Highlighted changes and intervals are of the respective median forecast within that group.

naive benchmark using straightforward aggregation. This allows us to test whether the improvements the models make are due to understanding of the need to appropriately update or simply as an agreement-focused response. We find in paired t-tests that for both GPT-4 at a Brier score of 0.13, $t(92) = 2.583, p = .011$, and Claude 2 at a Brier score of 0.14, $t(92) = 3.530, p = .001$, their updated forecasts are significantly less accurate than a simple average between the machine and the human median forecasts. This suggests that the updating itself is directionally correct but fails to improve upon a simple benchmark.

4 Discussion

Our results show that LLM prediction capabilities can rival the gold standard of the human crowd tournament method, if they themselves draw on what we call the 'wisdom of the silicon crowd.' Previous results on single models (Halawi et al. 2024; Schoenegger and Park 2023) showed that LLMs not only underperformed compared to a human crowd in a probabilistic forecasting context, but also failed to clear simple benchmarks; while others (Abolghasemi, Ganbold, and Rotaru 2023) failed to find evidence in favour of the LLMs outperforming humans in the context of time-series forecasting.[4] However, taking into account more sophisticated systems built on top of LLMs, such as combined retrieval and reasoning systems (Halawi et al. 2024), human-level prediction accuracy may already be considered matched in some aspects. We propose that the capabilities jump in moving from single frontier models to crowds of simple models in the same probabilistic forecasting context is a benefit that can be exploited in a variety of real-world contexts, as this aggregation approach remains simple to implement and does not require additions like that of news retrieval on each question. Our finding opens the door for simple, practically applicable steps like forecast aggregation to increase current AI models' forecasting ability—to predict future events in politics, economics, technology, and other real-world subjects—to a level on par with the human crowd. This opens up a lot of directly applied work, given that LLM prediction capabilities can inform decision-makers and businesses in circumstances where accurate probabilistic forecasts are difficult or expensive to acquire. Furthermore, since both our finding and the finding of Halawi et al. (2024) suggest that placing individual LLMs in advanced systems can increase their forecasting ability to a market-competitive level, it is natural to expect LLM predictions to be more widely applied across society in the near future.

Importantly, our finding holds despite the presence of an acquiescence bias (Costello and Roodenburg 2015; Hinz et al. 2007) in model predictions, in that our models' predictions are more likely to be above 50%, despite the resolution rate of all questions being almost even. This suggests that the 'wisdom of crowds' effect using median as our aggregation is able to counteract even this acquiescence bias that is present in the majority of individual models, a robustness feature of the 'wisdom of

[4]For more applications of LLMs in time-series forecasting see additional work (Cholakov and Kolev 2021; Gruver et al. 2024; Jin et al. 2023)