more direct comparison of two aggregated forecasts and would present a result that had so far not been achieved.

Null hypothesis 2, Study 1: The average of median LLM forecasts is neither statistically significantly more nor less accurate than the average of median human forecasts, $H_{0_2} : \mu_{LLM} = \mu_{Human}$.

Lastly for Study 1, we test for differences in forecasting accuracy between the twelve models. Some of these models are variations of each other, like GPT-4 and GPT-4 with Bing access, PaLM2 and PaLM2 in Bard, or Llama-2-70B and Solar-0-70B; while others differ on more fundamental grounds. Testing whether we find differences between models with different capabilities, endpoints, fine-tunings, sizes, etc. might provide further insight into which aspects help or hinder prediction accuracy.

Null hypothesis 3, Study 1: There are no statistically significant differences in the average accuracy across the different LLMs and humans, $H_{0_3} : \mu_1 = \mu_2 = \ldots = \mu_k$.

In Study 2, we investigate the ability of two frontier models (GPT-4 and Claude 2) to integrate human intelligence into their forecasting updating processes. This contributes to work on the interactions between humans and AI. While previous work has focused on how AI can improve human predictions (Schoenegger et al. 2024), this study looks at the reverse; how human forecasts can improve LLM predictions. This is studied in a context where models update their forecasts in response to receiving the human crowd prediction. This investigation of updating behavior is grounded in the premise that access to external information, such as the median forecast of a human crowd, can serve as a valuable reference point for recalibrating predictions. This process builds on Bayesian principles (Ghirardato 2002; Park 2022; Savage 1972) where prior beliefs (in this case, initial forecasts) are adjusted in light of new evidence (the human crowd median) to produce updated posterior beliefs (revised forecasts). The interaction between human and machine intelligence in this context is of particular interest, as it exemplifies the potential synergies that can emerge from integrating the intuitive, experience-based judgments of humans with the data-processing capabilities of LLMs.

We first investigate whether for each of the two LLMs, its average forecast becomes more accurate after being presented with the human crowd's median forecast. This is the most straightforward test of whether human cognitive output in this setting can augment machine-generated forecasts, as measured by forecasting accuracy.

Null hypothesis 1, Study 2: There is no statistically significant difference in the average accuracy of either LLM model before and after having been provided the human crowd median, $H_{0_1} : \mu_{\text{pre}} = \mu_{\text{post}}$.

We next investigate the impact of human median forecasting exposure on the precision of LLM forecasts. Specifically, we investigate whether the prediction intervals become narrower, indicating increased confidence in the forecasts: an effect that would suggest that the human predictions—to which the LLMs have been exposed—have nontrivial information value.

Null hypothesis 2, Study 2: The size of the prediction intervals do not become narrower after exposure to the human crowd median, $H_{0_2} : \Delta_{\text{range}} \geq 0$.

Finally, we investigate the relationship between the initial deviation of LLM forecasts from the human median and the magnitude of subsequent adjustments. This probes the extent to which larger discrepancies prompt more significant forecast revisions as would be expected.

Null hypothesis 3, Study 2: The magnitude of LLM forecast adjustments is not correlated with the initial deviation of their forecasts from the human crowd median, $H_{0_3} : \rho = 0$.

Both studies jointly provide the next step in LLM prediction capabilities research. Building on previous work (Schoenegger and Park 2023; Schoenegger et al. 2024), the present paper examines an LLM ensemble approach instead of a single model. Additionally, while other work (Schoenegger et al. 2024) has looked at how AI predictions can improve human accuracy, the present paper also tests the converse, thereby helping complete the picture of how humans and AI systems may interact in real-world contexts that require accurate forecasting.

2 Methods

All analyses were preregistered on the Open Science Framework[1]. We clearly label all exploratory and non-preregistered analyses as such throughout the paper to indicate which tests were decided on after having seen the data.

2.1 Study 1

In Study 1, we collected data from a total of twelve diverse large language models to simulate the LLM crowd. Specifically, these twelve models were GPT-4, GPT-4 (with Bing), Claude 2, GPT3.5-Turbo-Instruct, Solar-0-70b, Llama-2-70b, PaLM 2 (Chat-Bison@002), Coral (Command), Mistral-7B-Instruct, Bard (PaLM 2), Falcon-180B, and Qwen-7B-Chat. We accessed each model through a web interface and did not query any models via their APIs to hold the query method constant, thus using default parameters (e.g., temperature) for all models. These web interfaces included company-specific interfaces like those offered for the models by OpenAI, Anthropic, Cohere, and Google, as well as interfaces provided by other third parties such as Poe, Huggingface, and Modelscope that provided access to the remaining models. We took this approach to maximise the number of models that we could reliably query throughout the study period that we collected data for while retaining heterogeneity of model specifications as our goal was to draw on a diverse set of models. Additionally, this also kept this study in the context of publicly available and easily accessible models. The final set of models includes frontier proprietary models (GPT-4, Claude 2) as well open-source models (e.g., Llama-2-70b, Mistral 7B-Instruct) from a variety of demographically diverse companies originating from China, France, United Arab Emirates, South Korea, Canada, and the United States. We also have a variety of models with internet access (e.g., GPT-4 with Bing, Bard, Coral) and a large diversity of model sizes, ranging from 7 billion parameters to an estimated 1.6 trillion.[2] For a full list of all models and their central specifications, see Table 1 below.

In order to assess the prediction capabilities of these models, we drew on a set of forecasting questions that were asked in real time to a public forecasting tournament that ran from October 2023 to January 2024 on the platform Metaculus, where over the course of this tournament, 925 human forecasters provided at least one prediction. In this tournament, forecasters were able to sign up to Metaculus (Metaculus 2024) and predict on as many questions as they wanted. The questions posed ranged from conflict in the Middle East, interest rates, literary prizes, and English electoral politics to Indian air quality, cryptocurrency, consumer technology, and space travel. We focused exclusively on binary probabilistic forecasts, collecting a total of 31 questions. Each question included a question title, a background section detailing the context of the question being asked, and a resolution passage that spelled out in detail how the question will resolve. We drew on the same set of questions and used the publicly available human median predictions for each question as the human benchmark. For a full list of the questions, see Table 4 in the appendix.

For every probabilistic question, within 48 hours of the question opening, we queried each model three independent times and recorded their predictions at the default settings. We recorded both the quantitative forecast and the qualitative rationale for all entries. If a model was unresponsive because of a technical reason, we attempted to collect a forecast 24 hours after the first failed attempt. If a model failed to provide a forecast for non-technical reasons like model censorship/content restrictions after several attempts, we did not reattempt data collection and recorded the prediction as missing. For each question, we prompted each model three times and recorded all predictions.[3] For cases in which a model failed to provide a forecast for the second or third run after having provided a forecast before, we continued to query the model until all three forecasts were provided.

Our prompt that we used for all models included instructions on how to format the output as well as a number of prompting techniques that include instructing the model to respond as a superforecaster and to approach these questions step-by-step as is current best prompting practice. Each prompt also

[1] https://osf.io/sb6mw/?view_only=395ab8faccba419c91f5f12dcaf97ce6

[2] We monitored updates to the original models at the web interfaces and responded as follows to changes: In response to the release of GPT-4-Turbo, from Nov 6, we queried the 'Classic' model instead. For the upgrade to Claude 2.1, we did not switch the query method from Nov 21 . When Bard switched, at least in part, to Gemini Pro from PaLM 2, we ceased data collection of this model via the Bard interface from Dec 6.

[3] If a model only responded with 'Yes' or 'No' as their prediction, we coded this as 99% and 1% respectively, though we note that this happened in less than 1% of cases across models.

Table 1: Model Details

Model	Company	Internet Access	Open Source	Hosting Platform	Country of Company
GPT-4	OpenAI	No	No	OpenAI	United States
GPT-4 Bing	OpenAI	Yes	No	OpenAI	United States
Claude 2	Anthropic	No	No	Anthropic	United States
GPT-3.5-Turbo-Instruct	OpenAI	No	No	OpenAI	United States
Solar-0-70B	Upstage	No	Yes	Poe	South Korea
Llama-2-70B	Meta	No	Yes	Poe	United States
PaLM 2 (Chat-Bison@002)	Google	No	No	Poe	United States
Coral (Command)	Cohere	Yes	No	Cohere	Canada
Mistral-7B-Instruct	Mistral	No	Yes	Poe	France
Bard (PaLM 2)	Google	Yes	No	Google	United States
Falcon 180B	Technology Innovation Institute	No	No	Huggingface	United Arab Emirates
Qwen-7B-Chat	Alibaba Cloud	No	Yes	Modelscope	China