included detailed question background, resolution criteria, and question text as they were posed on the public forecasting tournament, see Figure 1.

> Full Prompt
>
> In this chat, you are a superforecaster that has a strong track record of accurate forecasts of the future. As an experienced forecaster, you evaluate past data and trends carefully and aim to predict future events as accurately as you can, even though you cannot know the answer. This means you put probabilities on outcomes that you are uncertain about (ranging from 0 to 100%). You aim to provide as accurate predictions as you can, ensuring that they are consistent with how you predict the future to be. You also outline your reasons for this forecasting. In your reasons, you will carefully consider the reasons for and against your probability estimate, you will make use of comparison classes of similar events and probabilities and take into account base rates and past events as well as other forecasts and predictions. In your reasons, you will also consider different perspectives. Once you have written your reasons, ensure that they directly inform your forecast.
> Then, you will provide me with a number between 0 and 100 (up to 2 decimal places) that is your best prediction of the event. Take a deep breath and work on this problem step-by-step. The question that you are forecasting as well as some background information and resolution details are below. Read them carefully before making your prediction.
> **Background:**
> **Resolution:**
> **Question:**

Figure 1: Full prompt for Study 1

For every set of machine forecasts, we also recorded the publicly available median human crowd prediction at the end of the day that the machine forecast was entered. If the prediction was entered on the first day, we collected the human crowd predictions at the end of the second day that the question was open to allow for higher participation rates. This was done to ensure a fair comparison of machine and human forecasts, as many LLMs can recall the current date, thus making timed forecasts of the nature studied here potentially sensitive to asynchronous queries while also introducing bias with respect to the human crowd. For roughly half the questions, the human forecasters were not able to see the human crowd forecast, though there is significant heterogeneity when the community predictions were made available to human forecasters. In 15 out of 31 questions, our data was collected prior to the revelation of the community prediction to the human forecasters.

For the human forecasts, we took the publicly available median forecast for each question. For the LLM ensemble approach, we computed the median from all non-missing forecasts on each question. We also computed the median forecast on each question for each model to enable cross-model comparisons. See Figure 2 for an overview of our LLM ensemble approach.

2.2 Study 2

In Study 2, we focused exclusively on two frontier models, GPT-4 and Claude 2. We used the same real-world forecasting tournament as in Study 1 as our study context, functioning as a source of questions and human forecasts. For Study 2, we employed a within-model research design that collected two forecasts (pre- and post-intervention) per model run for each question, with each question being posed three times at the standard temperature settings, resulting in six forecasts per model for each question. Our goal was to investigate LLM updating behaviour with respect to human cognitive output, i.e., whether and how LLMs take into account the human prediction estimates that forecasting tournament aggregates provide. We queried GPT-4 and Claude 2 via the OpenAI and Anthropic websites respectively.

We used a significantly longer and more elaborate set of prompts than in Study 1. The first prompt built on the '10 commandments of superforecasting' (Tetlock and Gardner 2016) as well as the literature on forecasting and updating, instructing models to carefully consider distinguishing different degrees of doubt, strike the correct balance between under- and overconfidence, and break difficult problems into sub-problems that are easier to solve, among other instructions. The second prompt, the intervention,

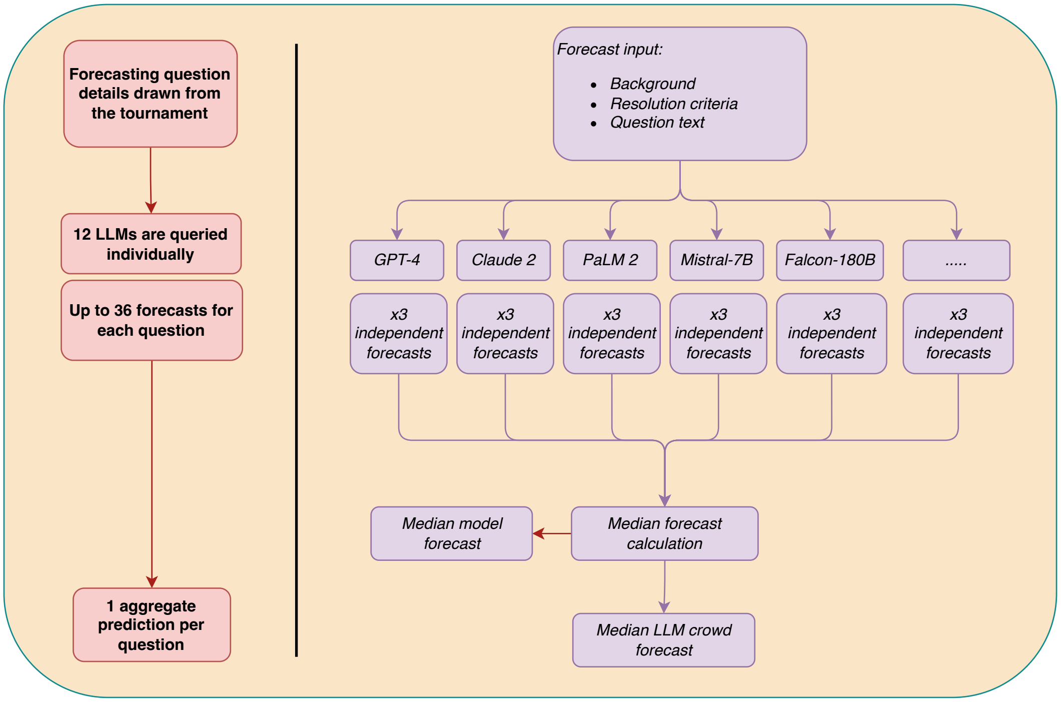
Forecasting question details drawn from the tournament
12 LLMs are queried individually
Up to 36 forecasts for each question
1 aggregate prediction per question
Forecast input:
Background
Resolution criteria
Question text
GPT-4
Claude 2
PaLM 2
Mistral-7B
Falcon-180B
.....
x3 independent forecasts
Median model forecast
Median forecast calculation
Median LLM crowd forecast

Figure 2: LLM Ensemble Mechanism Overview

informed the model of the respective human crowd's median forecast and asked it to update, if necessary, and to outline the reasons for the update (if any). For a full text of both prompts, see Figure 3 and Figure 4.

For both prompts, we collected forecasts not as point estimates but as probability ranges between 0% and 100% with two decimal point specificity. For further analysis, we treat the midpoint of this range as the point estimate and the provided predictions as upper and lower estimates. The human crowd median that is provided to the models is collected within 48 hours of the community prediction being revealed to allow human forecasters to learn about it and update their forecasts accordingly, generally leading to more well-calibrated predictions. Because of the time difference, the human forecasts are more accurate than those used in Study 1.

3 Results

3.1 Study 1

We collected a total of 1007 individual forecasts over the 31 questions from twelve LLMs that make up the ensemble. For 109 forecasts that we did not collect, this was due to technical problems with the model or interface at the time of forecast collection (in the case of Falcon-180B and PaLM 2),or because other models selectively chose not to answer certain questions, presumably due to their content restriction policies (this was the case for Coral (Command) and Qwen-7B-Chat). We also recorded some missing forecasts for Bard, which was due to the fact that the underlying model powering the interface was changed to Gemini Pro. To ensure consistency and allow comparisons between the different contexts of PaLM 2, we stopped collecting data at this point.

Across all models and questions, we observe a minimum raw forecast value of 0.1% and a maximum raw forecast value of 99.5%, with a median forecast of 60%. This indicates that the LLM models are more likely to make predictions above the 50% mid-point, with the mean forecast value of the crowd M=57.35 (SD=20.93) being significantly above the 50% mark, t(1006)=86.20, p<0.001. Importantly, the total question set resolved close to evenly, with 14/31 questions resolving positively. This imbalance thus suggests that LLM predictions generally favour positive resolutions above and beyond the appropriate empirical expectation, with just over 45% of questions resolving positively.

Initial Prompt

In this chat, you are a superforecaster who has a strong track record of accurate forecasting. You evaluate past data and trends carefully for potential clues to future events, while recognising that the past is an imperfect guide to the future so you will need to put probabilities on possible future outcomes (ranging from 0 to 100%). Your specific goal is to maximize the accuracy of these probability judgments by minimising the Brier scores that your probability judgments receive once future outcomes are known. Brier scores have two key components: calibration (across all questions you answer, the probability estimates you assign to possible future outcomes should correspond as closely as possible to the objective frequency with which outcomes occur) and resolution (across all questions, aim to assign higher probabilities to events that occur than to events that do not occur).

You outline your reasons for each forecast: list the strongest evidence and arguments for making lower or higher estimates and explain how you balance the evidence to make your own forecast. You begin this analytic process by looking for reference or comparison classes of similar events and grounding your initial estimates in base rates of occurrence (how often do events of this sort occur in situations that look like the present one?). You then adjust that initial estimate in response to the latest news and distinctive features of the present situation, recognising the need for flexible adjustments but also the risks of over-adjusting and excessive volatility. Superforecasting requires weighing the risks of opposing errors: e.g., of failing to learn from useful historical patterns vs. over-relying on misleading patterns. In this process of error balancing, you draw on the 10 commandments of superforecasting (Tetlock & Gardner, 2015) as well as on other peer-reviewed research on superforecasting:

1. Triage
2. Break seemingly intractable problems into tractable sub-problems
3. Strike the right balance between inside and outside views
4. Strike the right balance between under- and overreacting to evidence
5. Look for the clashing causal forces at work in each problem
6. Strive to distinguish as many degrees of doubt as the problem permits but no more
7. Strike the right balance between under- and overconfidence, between prudence and decisiveness
8. Look for the errors behind your mistakes but beware of rearview-mirror hindsight biases
9. Bring out the best in others and let others bring out the best in you
10. Master the error-balancing bicycle

Once you have written your reasons, ensure that they directly inform you forecast.

Then, you will provide me with your forecast that is a range between two numbers, each between between 0 and 100 (up to 2 decimal places) that is your best range of prediction of the event. Output your prediction as "My Prediction: Between XX.XX% and YY.YY%". Take a deep breath and work on this problem step-by-step.

The question that you are forecasting as well as some background information and resolution criteria are below. Read them carefully before making your prediction.

Background:

Resolution Criteria:

Question:

Figure 3: Initial prompt for Study 2

Such a bias towards more positive predictions may be a function of the machine-equivalent of acquiescence bias (Costello and Roodenburg 2015), where human responders tend to favour the positive/agreement option irrespective of question content (Hinz et al. 2007). See Figure 5 for a scatter plot of all model forecasts across all questions that shows heterogeneity between models of forecast distribution, ranges, and acquiescence bias.