Question:
{question}
Question Background:
{background}
Resolution Criteria:
{resolution_criteria}
Question close date: {close_date}
Instructions:
1. Given the above question, rephrase and expand it to help you do better answering. Maintain all information in the original question.
{{ Insert rephrased and expanded question.}}
2. Provide a few reasons why the answer might be no. Rate the strength of each reason.
{{ Insert your thoughts }}
3. Provide a few reasons why the answer might be yes. Rate the strength of each reason.
{{ Insert your thoughts }}
4. Aggregate your considerations. Think like a superforecaster (e.g. Nate Silver).
{{ Insert your aggregated considerations }}
5. Output an initial probability (prediction) given steps 1-4.
{{ Insert initial probability }}
6. Evaluate whether your calculated probability is excessively confident or not confident enough. Also, consider anything else that might affect the forecast that you did not before consider.
{{ Insert your thoughts }}
7. Output your answer (a number between 0 and 1) with an asterisk at the beginning and end of the decimal. (For example, if there are n resolution dates, you would output different *p* for each resolution date) Do not output anything else.
{{ Insert your answer }}

Figure 5: Scratchpad Prompt modified from Halawi et al. (2024)

I AGGREGATING HUMAN FORECASTS

We provide aggregated forecasts on benchmark questions from 500 members of the general public and 39 superforecasters, as described in Section 4:

- Public: www.forecastbench.org/datasets/forecast_sets/2024-07-21/2024-07-21.ForecastBench.human_public.json
- Superforecasters: www.forecastbench.org/datasets/forecast_sets/2024-07-21/2024-07-21.ForecastBench.human_super.json

As described in Section 4, members of the general public were recruited via Prolific and Facebook. First, participants completed an introductory survey designed to gather demographic information and evaluate performance on a few forecasting and comprehension tasks. Then, they were invited to take part in the main survey, featuring 20 random benchmark questions. Some participants were disqualified from participating in the main survey based on suspicious elements of their presurvey responses (e.g., answering questions rapidly, large numbers of submissions from the same IP address, etc.).

Superforecasters were invited to participate directly and were prompted to give forecasts for at least 20 questions from the benchmark (though, many chose to participate on additional questions). Superforecasters were also given access to other Superforecasters' forecasts and rationales and were allowed to comment on and update based on others' forecasts.

For each group, the median forecast for each forecasting question was taken to create the aggregated forecast sets. Code and individual forecasts are forthcoming once we can ensure that the text responses have been fully anonymized.

J REPRODUCE LLM FORECASTS

We evaluate 17 LLMs on our initial benchmark: GPT-3.5-Turbo-Instruct (Brown et al., 2020), GPT-4 (OpenAI, 2023), GPT-4o, Llama-2-70B (Touvron et al., 2023), Llama-3-7B, Llama-3-70B, Mistral-7B, Mistral-8x7B (Jiang et al., 2024a), Mistral-8x22B, Mistral-Large, Qwen1.5-110B-Chat, Claude-2.1 (Anthropic, 2023), Claude-3-Haiku, Claude-3.5-Sonnet, Claude-3-Opus (Anthropic, 2024), Gemini 1.5 Flash, and Gemini 1.5 Pro (Gemini Team, 2023).

To make inferences, we use APIs. For the GPT-suite, we use OpenAI's API; for Gemini-suite, we use Google's API; for Llama-suite, Mistral-7B, Mixtral-8x7B, Mixtral-8x22B and Qwen1.5-110B-Chat,

Question: {question}
Question Background: {background}
Resolution Criteria: {resolution_criteria}
Question close date: {close_date}
We have retrieved the following information for this question: {retrieved_info}
Instructions:
1. Given the above question, rephrase and expand it to help you do better answering. Maintain all information in the original question. {{ Insert rephrased and expanded question.}}
2. Let's start by coming up with a base-rate that could be helpful for forecasting this question. Come up with the best reference-class you can for this sort of event, and give a general base-rate that doesn't take into account factors unique to this question.
For instance, if the question were about the probability of a new technology being widely adopted within five years, you might look at historical data on the adoption rates of similar technologies as a reference class. Come up with a base-rate that could be relevant for this question.
The base-rate must be formatted as a clear probability (or number, in cases where you believe that to be more useful than a probability). For instance, imagine you are forecasting the probability that an incumbent president will be re-elected in an upcoming election in a hypothetical country. The past data shows that the incumbent has been elected 60% of the time.
Here, you would write 'The reference class I have chosen is the incumbent being elected. My base-rate is that the probability of the incumbent being re-elected is 0.6.' Give a justification for the base-rate, as well as a clear number.
Importantly, the base-rate should be as specific as it's possible to be without losing confidence that the number is correct. For instance, if you were forecasting on the probability of a hypothetical democratic country going to war in the next year, you should ideally produce a base-rate for a democratic country going to war in a given year, rather than simply thinking about a given country going to war.
{{ Insert your base rate }}
3. Now, let's think about factors specific to this question that may give us a good reason to deviate from the base-rate. Please give some reasons that the probability of this question resolving positively may be higher than the base rate. Please note specifically how they affect your forecast in terms of percentage point change. {{ Insert your thoughts }}
4. Now, let's think about reasons that the probability of this question resolving positively may be lower than the base rate. Please note specifically how they affect your forecast in terms of percentage point change. {{ Insert your thoughts }}
5. Consider any other factors that may affect the probability of this question resolving positively or negatively, that you have not already discussed in the previous two steps. {{ Insert your thoughts }}
6. Aggregate your considerations. Think like a superforecaster (e.g. Nate Silver). Give a ranking to each consideration based on how much you believe it ought to affect your forecast. {{ Insert your aggregated considerations }}
7. Are there any ways in which the question could resolve positively or negatively that you haven't considered yet, or that require some outside-the-box thinking? For example, if the question was 'Will Microsoft have a market capitalization of over $5tn by 2030', you might consider questions like:
How likely is it that Microsoft no longer exists in 2030? How likely is it that inflation erodes that value of the dollar as such that $5n is worth significantly less than it is today? How likely is it that there is a merger between Microsoft and another large company? How likely is it that Microsoft is broken up, as it is perceived to have monopoly power?
Here, we're thinking about things that are probably quite unlikely to happen, but should still be integrated into your forecast. Write up some possibilities and consider how they should be integrated into your final forecast. {{Insert your thoughts and considerations about how this should affect your forecast}}
8. Output an initial probability (prediction) given steps 1-7. {{ Insert initial probability. }}
9. Okay, now let's think about some other ways to consider how to forecast on this question. What would you say are the odds that if you could fast-forward and find out whether that statement is true or false, you would find out it's true? You must give an odds ratio. This odds ratio probably shouldn't be purely on the basis of the considerations in the previous steps, but you should think again about what you would expect to see if you could fast-forward into the future. If it helps, imagine that you're taking a bet. {{ Insert your odds ratio. }}
10. Given your rephrased statement from step 1, think of 2-3 statements that if you conditioned on their being TRUE, you would think it more or less likely that your statement would be TRUE as well. These statements must not DETERMINE OR BE LOGICALLY EQUIVALENT to the original statement. Be creative! {{ Insert 2 to 3 related statements. }}
11. For each of your related statements, give new odds of the original statement conditional on the related statement being TRUE. {{ For each related statement, insert new odds for the original statement. }}
12. Now consider each of your odds from the previous steps(steps 9 - 11), and come up with your all-things-considered odds ratio for the original statement. {{ Insert final odds for the original statement. }}
13. Now, convert that odds ratio to a probability between 0 and 1. {{Insert a probability}}
14. Now, consider the probability that you came up with in step 8, as well as the probability that you came up with in step 13. Which of these probabilities do you lean towards? How do you weigh them against one another? Write up your thoughts on which probability is more likely to be "correct", and then decide on a FINAL probability that will be used as your forecast. {{Insert your thoughts AND a final probability}}
15. Output your answer (a number between 0 and 1) with an asterisk at the beginning and end of the decimal. {{ Insert your answer }}

Figure 6: Superforecaster prompt 1

we use Together AI's API; for Mistral-Large, we use Mistral AI's API; and for the Claude-suite, we use Anthropic's API. To reproduce our results, people will need to gather these API keys.

We then record model predictions using several different methods: zero-shot prompting, scratchpad prompting, scratchpad prompting with retrieval augmentation, and scratchpad prompting with retrieval augmentation and aggregate human forecasts. For the scratchpad and information retrieval setting, we use the retrieval infrastructure from Halawi et al. (2024) and provide relevant news articles to the models in-context to reason about. Additionally, only models with a context window larger than 8,000 tokens were evaluated under the retrieval setting due to the inclusion of news articles in the prompt.

Question: {question}
Question Background: {background}
Resolution Criteria: {resolution_criteria}
Here's some related information from the news that I've collected for you: {retrieved_info}
Question close date: {close_date}
Instructions:
1. Rephrase the question as a statement about the future, e.g. you would rephrase "Will Biden be the U.S. president on January 1 2025?" as "Biden is the U.S. president on January 1 2025." {{ Insert question rephrased as a statement. }}
2. What would you say are the odds that if you could fast-forward and find out whether that statement is true or false, you would find out it's true? You must give an odds ratio. If it helps, imagine that you're taking a bet. {{ Insert your odds ratio. }}
3. Given your rephrased statement, think of 2-3 statements that if you conditioned on their being TRUE, you would think it more or less likely that your statement would be TRUE as well. These statements must not DETERMINE OR BE LOGICALLY EQUIVALENT to the original statement. Be creative! {{ Insert 2 to 3 related statements. }}
4. For each of your related statements, give new odds of the original statement conditional on the related statement being TRUE.insert new odds for the original statement. }}
5. Now consider each of your odds from the previous steps and come up with your all-things-considered odds ratio for the original statement. Output your answer (a number between 0 and 1) with an asterisk at the beginning and end of the decimal. {{ Insert final odds for the original statement. }}

Figure 7: Superforecaster prompt 2

Question: {question}
Question Background: {background}
Resolution Criteria: {resolution_criteria}
Relevant information we retrieved from news articles: {retrieved_info}
Question close date: {close_date}
Instructions:
1. Given the above question, rephrase and expand it to help you do better answering. Maintain all information in the original question. {{ Insert rephrased and expanded question.}}
2. Provide a few reasons why the answer might be no. Rate the strength of each reason. For now, ignore the evidence, ideas, and perspectives contained in the attached news articles. {{ Insert your thoughts }}
3. Provide a few reasons why the answer might be yes. Rate the strength of each reason. For now, ignore the evidence, ideas, and perspectives contained in the attached news articles. {{ Insert your thoughts }}
4. Aggregate the considerations you developed in the previous steps. Think like a superforecaster (e.g. Nate Silver). {{ Insert your aggregated considerations }}
5. Output an initial probability (prediction) given steps 1-4. {{ Insert initial probability. }}
6. Now, consider the perspectives, ideas, and evidence that was provided in the retrieved news articles. How should these affect your judgment of the probability of the question resolving positively? List all reasons why these news articles might increase the probability of the question resolving positively. {{Insert your thoughts}}
7. Now, let's focus on how the ideas, perspectives, and evidence provided in the news articles might decrease the probability of the question resolving positively. {{Insert your thoughts}}
8. Given what you've thought about in the previous two steps, update your probability from the initial probability you gave in step 5. {{Insert updated probability}}
9. Evaluate whether your calculated probability is excessively confident or not confident enough. Also, consider anything else that might affect the forecast that you did not before consider. {{ Insert your thoughts }}
10. Output your answer (a number between 0 and 1) with an asterisk at the beginning and end of the decimal. Do not output anything else. {{ Insert your answer }}

Figure 8: Superforecaster prompt 3

Additionally, to speed up the inference, we use multithreading with 50 workers, which requires a high rate limit and requests for better subscription plans from each source. However, one can run inference sequentially by setting it as 1 worker, but this requires longer time to generate all the baselines.

J.1 Zero-shot and scratchpad baselines

Prompts We use the zero-shot and scratchpad prompts shown in Appendix F.

Hyperparameters For the zero-shot setting, we set the maximum output token length to 50 since we only request probabilistic forecasts. For the scratchpad prompt, we increase the maximum output token length to 1300 as it requires reasoning and probabilistic forecasts. We initially considered a high token length of 3000, but after observing that the maximum response length was around 1250, we settled on 1300 as the optimal maximum token length. In both cases, the model temperature is set to 0 to ensure stable outputs.

How to Reproduce To run zero-shot and scratchpad baselines, follow the steps below:

1. Insert all the necessary API keys in `src/helpers/keys.py`.