Superforecaster forecast dataset In addition to forecasts and rationales, the superforecasters provide pertinent information about their forecasting process, like search terms used and useful URLs consulted. See Section B.2.2.

LLM forecast dataset Similar to the general public dataset, we ask LLMs to produce forecasts on each of 1,000 forecast questions in the LLM question set. Their rationales are also included in the dataset whenever provided. See Section B.2.3 for details.

Question & resolutions dataset In creating the benchmark, we have automated question creation and resolution from all of the sources outlined in Table 1. We provide these as a dataset that can be combined with the forecast datasets mentioned above. See Section B.1 for details.

4 HUMAN FORECASTER BASELINE

To compare LLM forecasting performance to human performance, we ran surveys of two different groups: the general public and superforecasters. We scored each group's median forecast, treating it as representative of its overall performance.

4.1 GENERAL PUBLIC

We recruited 500 human forecasters via Prolific and advertisements on Facebook to participate as representatives of the general public. These human subjects completed a brief introductory survey to gather demographic information[8] and evaluate performance on a few forecasting and comprehension tasks. They then completed a one-hour survey containing 20 random questions from the 200-item human question set described in Section 3.2, providing their forecasts and rationales for each question.

The number of responses per question varied to ensure representativeness across categories and sources; at least 40 responses were gathered per question, averaging 49 responses per question.

4.2 SUPERFORECASTERS

We recruited 39 superforecasters, who have a strong track record of accurate performance on a diverse set of geopolitical questions, to participate in a 9-day forecasting experiment in which participants were prompted to give their individual forecasts for 20 random questions from the same 200-item human question set described above. Roughly halfway through the 9-day experiment, participants were moved into a group forecasting stage in which we allowed them to see one another's forecasts and rationales and to communicate about each question. They were also given the opportunity to forecast on questions beyond the 20 questions assigned to them individually.

Because of the lower number of superforecasters, questions generally had fewer responses than in the public survey; a minimum of 3 forecasts were recorded for each question, with an average of 8 responses per question.[9]

5 LLM BASELINE

In this section, we evaluate the forecasting capabilities of LLMs and report on the methodology and results.

5.1 METHODOLOGY

We evaluate a suite of instruction-following chat models without any additional fine-tuning (see Section 2 for details on the models). For each baseline outlined below, we prompt the model to generate a probabilistic forecast that the question will resolve to "Yes."

[8] See Appendix L for an overview of public participant demographics.

[9] See Figure 2 for an example question from the human surveys.

Baselines We implement seven baselines: (1) zero-shot prompting; (2) prompting with scratchpad instructions; (3) prompting with scratchpad instructions and retrieved news articles; (4) zero-shot prompting with crowd forecasts; (5) scratchpad prompting with crowd forecasts; (6) scratchpad prompting with retrieved news articles and crowd forecasts; and (7) aggregating predictions from multiple LLMs. Each baseline is described in more detail below.

1 Our first baseline prompts the model **zero-shot** to generate a forecast directly without generating other content, such as intermediate thinking (Figure 4). By prompting the model to output its forecast directly, we assess raw forecasting capability without sensitivity to prompting strategies.

2 Our second baseline, prompts the model with **scratchpad** instructions (Nye et al., 2021) that outline a procedure the model should use to reason about the question (Figure 5). Our scratchpad prompt comes from (Halawi et al., 2024), which formed its prompts through a combination of analyzing the Brier score as prompt changes were made, and by adding language to fix common errors the LLMs would make, e.g., asking them to rephrase the question for understanding.

3 Since LLMs' knowledge is not continuously updated, it is important to provide them with up-to-date information relevant to the question (Zou et al., 2022). Our fourth baseline, **scratchpad with news**, uses the same scratchpad prompt as above, supplemented with retrieved news articles. The retrieval system is the same as described in Halawi et al. (2024): an LLM generates search queries for a news API, filters articles for relevancy, and summarizes the articles.

4 The question sets we provide to LLMs contain what we term **freeze values**. For market questions these are just the crowd forecast on the market the day the question set was created, as described in Section 3.2. For dataset questions, these are baseline values relevant to the forecasting task.[10] Our third baseline is the **zero-shot with freeze values**. This is simply the zero-shot prompt from Baseline 1 supplemented with the freeze value and an explanation of the freeze value. For examples of the freeze value and its explanation, see Table 8 and Table 9.

5 Our fifth baseline is the **scratchpad with freeze values** (the scratchpad prompt from Baseline 2 supplemented with freeze values as explained in Baseline 4).

6 Our sixth baseline is the **scratchpad with news with freeze values**.

7 In our final baseline, we aggregate the predictions generated by LLMs into an **LLM "ensemble"** forecast. We do this as Metaculus (2023) shows that an ensemble of all forecasters consistently outperforms using just the top 5, 10, ..., 30 best forecasters (based on past scores).

   To produce the LLM ensemble forecast, we use 3 models (GPT-4o, Claude-3.5-Sonnet, and Gemini-1.5-Pro) and 3 prompts crafted by superforecasters (Figure 6, Figure 7, and Figure 8). This results in 9 forecasts per question. We generate 3 LLM ensemble baselines using the median, geometric mean, and geometric mean of log odds (Satopää et al., 2014). For details, see Appendix E.

5.2 Results

Comparing humans and LLMs In Table 2, we show that superforecasters achieve an overall mean Brier score of 0.096, significantly outperforming both the general public (Brier $= 0.121$, $p < 0.001$) and the top LLM performer on the 200-item subset (Claude 3.5 Sonnet: Brier $= 0.122$, $p < 0.001$).[11] The top-performing LLMs all had access to the crowd forecast on market questions (the "freeze values" from Baselines 4, 5, and 6 above). The top-performing model without access to the crowd forecast on market questions was less accurate than models with access to the human forecast with a Brier score of 0.136. The comparison between humans and LLMs relies on the 200 questions forecasted by humans, which is a random sub-sample of the 1,000 questions in the question set provided to LLMs (excluding combination questions).[12]

[10] For example, in a question generated from a Wikipedia page about whether a chess player's Elo rating will increase by a given date, the freeze value is the chess player's Elo rating on the question set generation date. An explanation of what the freeze value represents is also provided.

[11] See statistical note in Appendix G.

[12] Accuracy measures are based on more than 200 forecasts because human and LLM forecasters submitted multiple forecasts on each dataset question, one for each time horizon. The results presented here include forecasts over the 7-, 30-, 90-, and 180-day time horizons.

Table 2: LLM/Human Leaderboard (top 10)

Model	Organization	Information provided	Prompt	Brier Score ↓ Dataset (N=422)	Market (N=76)	Overall (N=498)	Confidence Interval	Pairwise p-value comparing to No. 1	Pct. more accurate than No. 1
Superforecaster median forecast	ForecastBench	–	–	0.118	0.074	0.096	[0.076, 0.116]	–	0%
Public median forecast	ForecastBench	–	–	0.153	0.089	0.121	[0.101, 0.141]	<0.001	22%
Claude-3-5-Sonnet-20240620	Anthropic	Freeze values	Scratchpad	0.138	0.107	0.122	[0.099, 0.146]	<0.001	31%
Claude-3-5-Sonnet-20240620	Anthropic	News with freeze values	Scratchpad	0.142	0.112	0.127	[0.104, 0.150]	<0.001	29%
GPT-4-Turbo-2024-04-09	OpenAI	Freeze values	Zero shot	0.162	0.095	0.128	[0.105, 0.151]	<0.001	32%
Claude-3-5-Sonnet-20240620	Anthropic	Freeze values	Zero shot	0.145	0.117	0.131	[0.103, 0.159]	<0.001	31%
GPT-4	OpenAI	Freeze values	Zero shot	0.167	0.096	0.132	[0.109, 0.155]	<0.001	31%
GPT-4o	OpenAI	News with freeze values	Scratchpad	0.162	0.105	0.133	[0.113, 0.154]	<0.001	25%
Claude-3-5-Sonnet-20240620	Anthropic	–	Scratchpad	0.138	0.133	0.136	[0.113, 0.158]	<0.001	28%
GPT-4o	OpenAI	Freeze values	Scratchpad	0.161	0.113	0.137	[0.115, 0.158]	<0.001	27%

Notes:

1. Shows performance on the 200 standard questions provided in the human question set at the 7-, 30-, 90-, and 180-day forecast horizons. See Table 18 for top 50.
2. The full leaderboard is available at www.forecastbench.org. Online results are updated nightly, so may be slightly different than the version presented here.
3. For resolved market questions, forecasts are compared against ground truth while for unresolved market questions, they are compared to community aggregates.
4. The overall score is calculated as the average of the mean dataset Brier score and the mean market Brier score.
5. Pairwise p-value comparing to No. 1 (bootstrapped): The p-value calculated by bootstrapping the differences in overall score between each model and the best forecaster under the null hypothesis that there's no difference.
6. Pct. more accurate than No. 1: The percent of questions where this forecaster had a better overall score than the best forecaster.

As a particular failure mode, we find LLMs are significantly worse at combination questions. Although our human surveys did not explicitly ask for forecasts on combination questions, we bound human performance by assuming independence of the component of each combination question. This underestimates human accuracy because a human forecaster predicting the outcome of a combination question could account for dependence between the permuted events. In Table 20, we present this comparison of human and LLM forecasts. We see that LLMs perform poorly on these combination questions, and including them in the benchmark widens the gap between human and LLM performance: superforecasters (Brier $= 0.076$) outperform the general public (Brier $= 0.096$) and the top LLM (GPT-4o, Brier $= 0.130$) significantly. To benchmark the size of this gap in performance, the 0.054 Brier score gap in performance between superforecasters and GPT-4o is significantly larger than the 0.026 gap in performance between GPT-4o and GPT-4.

Table 3: LLM Leaderboard (top 10)

Model	Organization	Information provided	Prompt	Brier Score ↓ Dataset (N=5,492)	Market (N=897)	Overall (N=6,389)	Confidence Interval	Pairwise p-value comparing to No. 1	Pct. more accurate than No. 1
Claude-3-5-Sonnet-20240620	Anthropic	Freeze values	Scratchpad	0.169	0.078	0.123	[0.117, 0.129]	–	0%
GPT-4-Turbo-2024-04-09	OpenAI	Freeze values	Scratchpad	0.172	0.080	0.126	[0.120, 0.132]	0.096	43%
GPT-4o	OpenAI	Freeze values	Scratchpad	0.186	0.069	0.128	[0.122, 0.133]	<0.01	43%
Gemini-1.5-Pro	Google	Freeze values	Scratchpad	0.162	0.106	0.134	[0.128, 0.139]	<0.001	35%
GPT-4o	OpenAI	News with freeze values	Scratchpad	0.190	0.084	0.137	[0.131, 0.143]	<0.001	39%
Gemini-1.5-Pro	Google	News with freeze values	Scratchpad	0.166	0.111	0.139	[0.133, 0.144]	<0.001	34%
Claude-3-Opus-20240229	Anthropic	Freeze values	Zero shot	0.186	0.093	0.139	[0.133, 0.146]	<0.001	41%
Qwen1.5-110B-Chat	Qwen	Freeze values	Scratchpad	0.176	0.108	0.142	[0.136, 0.148]	<0.001	30%
Claude-3-5-Sonnet-20240620	Anthropic	News with freeze values	Scratchpad	0.184	0.101	0.143	[0.137, 0.149]	<0.001	32%
Claude-3-5-Sonnet-20240620	Anthropic	Freeze values	Zero shot	0.192	0.094	0.143	[0.136, 0.150]	<0.001	42%

Notes:

1. Shows performance on the 1,000 (500 standard, 500 combination) questions in the LLM question set at the 7-, 30-, 90-, and 180-day forecast horizons. See Table 19 for top 50.
2. The full leaderboard is available at www.forecastbench.org. Online results are updated nightly, so may be slightly different than the version presented here.
3. For resolved market questions, forecasts are compared against ground truth while for unresolved market questions, they are compared to community aggregates.
4. The overall score is calculated as the average of the mean dataset Brier score and the mean market Brier score.
5. Pairwise p-value comparing to No. 1 (bootstrapped): The p-value calculated by bootstrapping the differences in overall score between each model and the best forecaster under the null hypothesis that there's no difference.
6. Pct. more accurate than No. 1: The percent of questions where this forecaster had a better overall score than the best forecaster.

Comparing LLMs Table 3 excludes humans and evaluates LLMs on the entire question set (N=1,000 questions). Here we see a similar ranking of models, with Claude 3.5 Sonnet slightly outperforming GPT-4 Turbo. As in Table 2, most of the top-performing models use the scratchpad prompt (Figure 5) and use as inputs the human crowd forecasts for market questions. Access to recent topical news related to the questions did not improve performance.

LLM performance and forecasting accuracy Figure 1a demonstrates the seemingly linear relationship between Chatbot Arena scores (Chiang et al., 2024) and the overall Brier score from Table 2. We observe a significant correlation ($r = -0.68$, $p = 0.003$), indicating that models with higher Arena scores tend to produce more accurate forecasts. The linear relationship implies that LLMs