ForecastBench: A Dynamic Benchmark of AI Forecasting Capabilities

Ezra Karger[1,2] Houtan Bastani[1] Chen Yueh-Han[3]
Zachary Jacobs[1] Danny Halawi[4] Fred Zhang[4] Philip E. Tetlock[1,5]

[1]Forecasting Research Institute [2]Federal Reserve Bank of Chicago [3]New York University
[4]University of California, Berkeley [5]University of Pennsylvania

ICLR 2025: April 24-28

Outline

Motivation

- Forecasting is important and useful
  - Evolution of a pandemic
  - Economic indicators
  - Geopolitical events
  - ...
- Human forecasting is time-consuming and expensive $\rightarrow$ LLM forecasters
  - When will LLMs forecast as well as humans?
- Previous forecasting benchmarks have been static
  - Made obsolete once training cutoffs are after question resolution dates
  - Knowledge cutoffs are imprecise
  - Risk test set contamination

ForecastBench

- Continuously updated with questions about future events → immune to look-ahead bias
- Periodic surveys of the general public and superforecasters → human comparison
- Fully automated benchmark with open source codebase
- Publicly available leaderboards updated nightly
- Datasets released regularly
  - forecast questions
  - forecasts by LLMs and humans, with rationales
  - resolutions
- Will be maintained at least until mid-2027 thanks to a grant from Open Philanthropy!

Links

- Benchmark: https://www.forecastbench.org
- Code (MIT License): https://github.com/forecastingresearch/forecastbench
- Data (CC-BY-SA-4.0 License):
  - https://github.com/forecastingresearch/forecastbench-datasets
  - https://huggingface.co/datasets/forecastingresearch/forecastbench-datasets

Outline

Automated System

Our automated system manages the benchmark, from updating the question bank, to generating and releasing question sets, to resolving forecasts and updating the leaderboard.

- **Question Bank**: updated nightly
- **Question Sets**: generated every 2 weeks
- **Eliciting Forecasts**: every 2 weeks from LLMs, periodically from the general public and superforecasters
- **Leaderboard**: updated nightly

Question Bank

The Question Bank stores all questions to sample from. There are 2 types of questions:

1. **Market questions**: pulled from 4 forecasting platforms: Manifold, Metaculus, Polymarket, and the RAND Forecasting Initiative.
2. **Dataset questions**: generated from 5 datasets: ACLED, DBnomics, FRED, Wikipedia, and Yahoo! Finance.

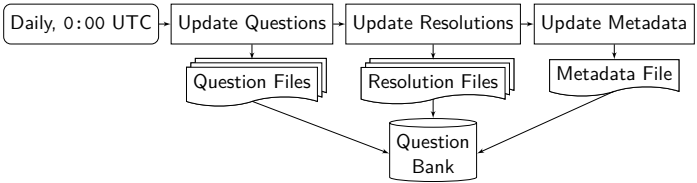


Figure: The Question Bank updating flowchart.

Question Sets

Every two weeks we sample 1,000 questions from the Question Bank to create the LLM Question Set.

- 500 standard questions: 250 market and 250 dataset questions.
- 500 *combination* questions: 250 market and 250 dataset questions.
  - Each combination question is just a pair of standard questions from the same source
  - We ask for forecasts on the Boolean combinations of these questions: $P(Q1 \cap Q2)$, $P(\neg Q1 \cap Q2)$, $P(Q1 \cap \neg Q2)$, and $P(\neg Q1 \cap \neg Q2)$

We sample 200 questions from the LLM Question Set to create the Human Question Set.

- 200 standard questions: 100 market and 100 dataset questions.
- No combination questions. Their combination forecasts are generated by treating $Q1$ and $Q2$ as independent, putting them at a disadvantage for these forecasts.

Eliciting forecasts (1/2)

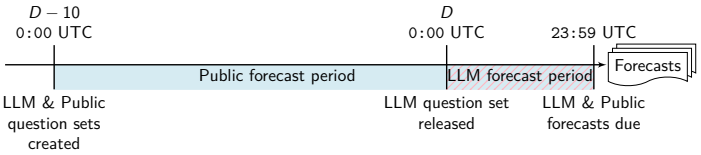


Figure: Timeline from question set generation to forecast due date, D.

- LLMs
  - Forecast all 1,000 questions in the Question Set within 24 hours
  - 17 models including GPT-4o, Claude Sonnet 3.5, Gemini 1.5 Pro, Qwen1.5 110B Chat, ...
  - Prompting strategies: zero-shot, scratchpad, scratchpad with retrieval (i.e., news)
    - ★ For market questions, with or without crowd forecasts provided (what we term "freeze values")
- General public and superforecasters
  - Forecast 200 questions sampled from the Question Set
  - Surveys take 10 days to run

Eliciting forecasts (2/2): Forecast Sets

Forecasts are returned in **Forecast sets**. Each question requires a different number of forecasts to be produced.

Question Type	**Users forecast...**	**№ of Forecasts**
Standard market questions	...the final outcome	1
Combination market questions	...the final outcome for all Boolean combinations of the questions	4
Standard dataset questions	...the outcome n days in the future, where $n \in N, N = \{7, 30, 90, 180, 365, 1095, 1825, 3650\}$	8
Combination dataset questions	...the outcome for all Boolean combinations of the questions at all forecast horizons in N	32

$\Rightarrow$ for each Question Set, LLMs provide approximately $11,250$ forecasts in the corresponding Forecast Set: $250 \times (1 + 4 + 8 + 32)$.

Leaderboard

The forecast sets produced bi-weekly are resolved every night and the leaderboard is updated.

- Market questions are scored against the crowd forecast until resolution, when they're scored against ground truth.
- Dataset questions are resolved against ground truth as datasets are updated and revised.

Outline

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.
2. For resolved market questions, forecasts are compared against ground truth while for unresolved market questions, they are compared to community aggregates.
3. The overall score is calculated as the average of the mean dataset Brier score and the mean market Brier score.
4. 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.
5. Pct. more accurate than No. 1: The percent of questions where this forecaster had a better overall score than the best forecaster.

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.
2. For resolved market questions, forecasts are compared against ground truth while for unresolved market questions, they are compared to community aggregates.
3. The overall score is calculated as the average of the mean dataset Brier score and the mean market Brier score.
4. 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.
5. Pct. more accurate than No. 1: The percent of questions where this forecaster had a better overall score than the best forecaster.

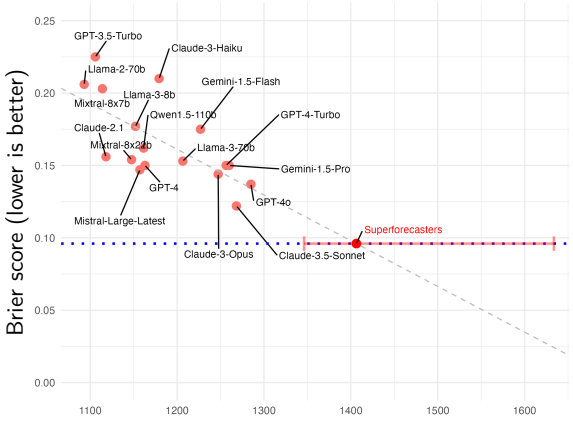
LLM/Human Combo Leaderboard (top 10)

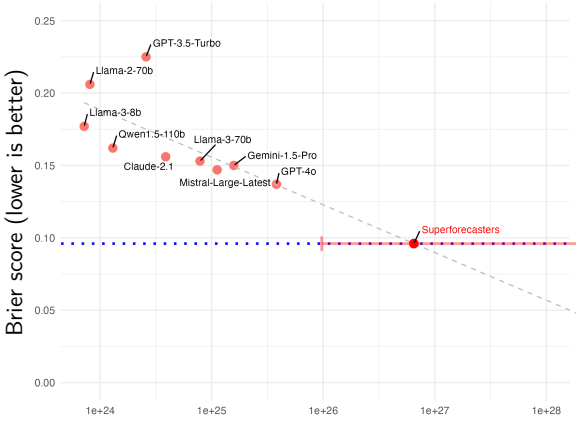
Model	Organization	Information provided	Prompt	Brier Score ↓ Dataset (N=1,754)	Market (N=296)	Overall (N=2,050)	Confidence Interval	Pairwise p-value comparing to No. 1	Pct. more accurate than No. 1
Superforecaster median forecast	ForecastBench	–	–	0.091	0.062	0.076	[0.067, 0.086]	–	0%
Public median forecast	ForecastBench	–	–	0.119	0.072	0.096	[0.086, 0.105]	<0.001	23%
GPT-4o	OpenAI	Freeze values	Scratchpad	0.175	0.085	0.130	[0.119, 0.141]	<0.001	24%
Claude-3-5-Sonnet-20240620	Anthropic	Freeze values	Scratchpad	0.154	0.107	0.131	[0.118, 0.143]	<0.001	24%
GPT-4-Turbo-2024-04-09	OpenAI	Freeze values	Scratchpad	0.164	0.101	0.133	[0.121, 0.145]	<0.001	23%
GPT-4o	OpenAI	News with freeze values	Scratchpad	0.171	0.104	0.137	[0.125, 0.149]	<0.001	20%
Gemini-1.5-Pro	Google	Freeze values	Scratchpad	0.152	0.130	0.141	[0.130, 0.152]	<0.001	21%
Gemini-1.5-Pro	Google	News with freeze values	Scratchpad	0.154	0.133	0.143	[0.133, 0.154]	<0.001	21%
Claude-3-5-Sonnet-20240620	Anthropic	News with freeze values	Scratchpad	0.160	0.130	0.145	[0.132, 0.158]	<0.001	20%
Claude-3-5-Sonnet-20240620	Anthropic	Freeze values	Zero shot	0.174	0.119	0.146	[0.133, 0.160]	<0.001	22%

Notes:

1. This shows performance on all 200 standard questions from the human question set *plus* those combination questions from the LLM question set where humans provided forecasts on both components ($Q1$ and $Q2$). LLM scores are only for this combined question set. Human forecasts for combination questions are generated from their forecasts on the component questions by assuming independence (which is not always the case, putting humans at a disadvantage). Evaluated at the 7-, 30-, 90-, and 180-day forecast horizons.
2. For resolved market questions, forecasts are compared against ground truth while for unresolved market questions, they are compared to community aggregates.
3. The overall score is calculated as the average of the mean dataset Brier score and the mean market Brier score.
4. 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.
5. Pct. more accurate than No. 1: The percent of questions where this forecaster had a better overall score than the best forecaster.

When could AI achieve superforecaster-level capabilities?





Contact us to benchmark your model!

forecastbench@forecastingresearch.org

https://www.forecastbench.org