Table 20: Leaderboard: Human question set with LLM question set combination questions (top 50)

Model	Organization	Information provided	Prompt	Brier Score ↓			Confidence Interval	Pairwise p-value comparing to No. 1	Pct. more accurate than No. 1
				Dataset (N=1,754)	Market (N=296)	Overall (N=2,050)			
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%
Gemini-1.5-Pro	Google	–	Scratchpad	0.152	0.143	0.148	[0.137, 0.158]	<0.001	20%
GPT-4-Turbo-2024-04-09	OpenAI	–	Scratchpad	0.164	0.132	0.148	[0.138, 0.158]	<0.001	17%
Gemini-1.5-Pro	Google	News	Scratchpad	0.154	0.143	0.148	[0.137, 0.160]	<0.001	21%
Claude-3-5-Sonnet-20240620	Anthropic	–	Scratchpad	0.154	0.143	0.149	[0.137, 0.160]	<0.001	20%
GPT-4o	OpenAI	–	Scratchpad	0.175	0.122	0.149	[0.138, 0.159]	<0.001	19%
Claude-3-Opus-20240229	Anthropic	Freeze values	Zero shot	0.173	0.124	0.149	[0.135, 0.162]	<0.001	21%
GPT-4o	OpenAI	News	Scratchpad	0.171	0.127	0.149	[0.138, 0.160]	<0.001	18%
GPT-4-Turbo-2024-04-09	OpenAI	Freeze values	Zero shot	0.200	0.100	0.150	[0.138, 0.162]	<0.001	24%
Qwen1.5-110B-Chat	Qwen	Freeze values	Scratchpad	0.171	0.131	0.151	[0.140, 0.162]	<0.001	16%
Claude-3-5-Sonnet-20240620	Anthropic	News	Scratchpad	0.160	0.149	0.154	[0.143, 0.166]	<0.001	19%
Imputed Forecaster	ForecastBench	–	–	0.250	0.059	0.155	[0.147, 0.163]	<0.001	22%
GPT-4	OpenAI	Freeze values	Zero shot	0.213	0.099	0.156	[0.144, 0.168]	<0.001	21%
Gemini-1.5-Pro	Google	Freeze values	Zero shot	0.205	0.110	0.157	[0.144, 0.171]	<0.001	20%
Claude-3-5-Sonnet-20240620	Anthropic	News	Superforecaster 2	0.167	0.149	0.158	[0.146, 0.169]	<0.001	17%
GPT-4	OpenAI	Freeze values	Scratchpad	0.190	0.125	0.158	[0.145, 0.171]	<0.001	19%
Claude-3-Opus-20240229	Anthropic	Freeze values	Scratchpad	0.185	0.134	0.159	[0.148, 0.171]	<0.001	18%
LLM Crowd	ForecastBench	News	–	0.241	0.080	0.161	[0.153, 0.168]	<0.001	18%
GPT-4-Turbo-2024-04-09	OpenAI	News with freeze values	Scratchpad	0.209	0.114	0.161	[0.149, 0.173]	<0.001	20%
LLM Crowd	ForecastBench	News	–	0.242	0.083	0.162	[0.155, 0.170]	<0.001	18%
LLM Crowd	ForecastBench	News	–	0.243	0.082	0.162	[0.155, 0.170]	<0.001	18%
Gemini-1.5-Pro	Google	News	Superforecaster 1	0.176	0.151	0.164	[0.153, 0.175]	<0.001	19%
Mistral-Large-Latest	Mistral AI	Freeze values	Scratchpad	0.185	0.143	0.164	[0.154, 0.175]	<0.001	16%
GPT-4	OpenAI	–	Scratchpad	0.190	0.140	0.165	[0.156, 0.174]	<0.001	15%
Qwen1.5-110B-Chat	Qwen	–	Scratchpad	0.171	0.161	0.166	[0.156, 0.175]	<0.001	15%
Gemini-1.5-Pro	Google	–	Zero shot	0.205	0.128	0.167	[0.154, 0.179]	<0.001	19%
Gemini-1.5-Flash	Google	Freeze values	Scratchpad	0.179	0.154	0.167	[0.153, 0.180]	<0.001	18%
Claude-2.1	Anthropic	–	Scratchpad	0.228	0.105	0.167	[0.157, 0.177]	<0.001	20%
Llama-3-70b-Chat-Hf	Meta	Freeze values	Zero shot	0.205	0.132	0.168	[0.155, 0.182]	<0.001	19%
Llama-3-70b-Chat-Hf	Meta	Freeze values	Scratchpad	0.208	0.129	0.169	[0.158, 0.179]	<0.001	17%
Claude-3-Opus-20240229	Anthropic	–	Zero shot	0.173	0.165	0.169	[0.156, 0.183]	<0.001	18%
Gemini-1.5-Flash	Google	Freeze values	Zero shot	0.217	0.122	0.169	[0.155, 0.183]	<0.001	23%
GPT-4-Turbo-2024-04-09	OpenAI	–	Zero shot	0.200	0.139	0.169	[0.157, 0.182]	<0.001	19%
GPT-4-Turbo-2024-04-09	OpenAI	News	Scratchpad	0.209	0.131	0.170	[0.159, 0.180]	<0.001	17%
Gemini-1.5-Flash	Google	–	Scratchpad	0.179	0.161	0.170	[0.159, 0.181]	<0.001	16%
GPT-4-Turbo-2024-04-09	OpenAI	News	Superforecaster 2	0.202	0.139	0.170	[0.160, 0.181]	<0.001	17%
Qwen1.5-110B-Chat	Qwen	News with freeze values	Scratchpad	0.198	0.146	0.172	[0.161, 0.183]	<0.001	16%
Claude-3-5-Sonnet-20240620	Anthropic	–	Zero shot	0.174	0.171	0.172	[0.158, 0.187]	<0.001	17%
Mistral-Large-Latest	Mistral AI	Freeze values	Zero shot	0.203	0.145	0.174	[0.160, 0.188]	<0.001	19%
Claude-2.1	Anthropic	Freeze values	Scratchpad	0.228	0.120	0.174	[0.162, 0.186]	<0.001	20%
GPT-4o	OpenAI	News	Superforecaster 3	0.206	0.145	0.175	[0.165, 0.186]	<0.001	16%

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. 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.

`NUM_SUMMARIES_THRESHOLD`: The threshold number of summaries to generate. This is set to 10.

`PRE_FILTER_WITH_EMBEDDING`: A boolean flag indicating whether to pre-filter articles using embeddings. This is set to `True`.

`PRE_FILTER_WITH_EMBEDDING_THRESHOLD`: The threshold for pre-filtering articles using embeddings. This is set to 0.32.

`RANKING_MODEL_NAME`: The name of the model used for ranking articles. We use `gpt-3.5-turbo-1106`.

`RANKING_TEMPERATURE`: The temperature setting for the ranking model, which controls the randomness of the output. We set this to 0.0 for deterministic outputs.

`RANKING_PROMPT_TEMPLATE`: The template used for ranking articles. In our configuration, we use `PROMPT_DICT["ranking"]["0"]`.

`RANKING_RELEVANCE_THRESHOLD`: The relevance threshold for ranking articles. This is set to 4.

RANKING_COSINE_SIMILARITY_THRESHOLD: The cosine similarity threshold used in ranking. This is set to 0.5.

SORT_BY: The criterion used to sort articles. We sort by `date`.

RANKING_METHOD: The method used for ranking articles. We use `llm-rating`.

RANKING_METHOD_LLM: The specific method for ranking articles using the LLM. We use `title_250_tokens`, meaning ranking articles based on their titles and the first 250 tokens.

NUM_SUMMARIES_THRESHOLD: The threshold number of summaries to generate for final output. This is set to 20.

EXTRACT_BACKGROUND_URLS: A boolean flag indicating whether to extract background URLs from the articles. This is set to `True`.

Inference Hyperparameters: We set the maximum output token length to 2000 to accommodate reasoning and probabilistic forecasts. We set the model temperature to 0 to ensure stable outputs.

How to reproduce To run the Scratchpad with Information Retrieval baseline, follow these steps:

1. To run the information retrieval part:
   (a) Insert all the necessary API keys in `llm_retrieval/forecasting-llm-retrieval/config/keys.py`. Specifically, add the Newscatcher and OpenAI API keys.
   (b) Run `llm_retrieval/notebooks/retrieval_cache.ipynb`.
   (c) Save all the retrieved news under a folder called `news`.
2. To run the scratchpad with the information retrieval baseline:
   (a) Insert all the necessary API keys in `src/helpers/constants.py`.
   (b) Place the "news" folder in the same directory as `src/base_eval/all_recurrent_llm_baselines/main.py`.
   (c) Run `src/base_eval/all_recurrent_llm_baselines/main.py`.

J.3 LLM "ENSEMBLE" BASELINE

To produce the LLM ensemble forecast, we query three models: GPT-4o, Claude-3.5-Sonnet, and Gemini-1.5-Pro. We use three prompts crafted by superforecasters who were given explicit instructions to write prompts that would help an LLM produce accurate forecasts This results in $3 \times 3 = 9$ forecasts per question. We then show 3 LLM crowd baselines using the median, geometric mean, and geometric mean of log odds.

Prompts We use the 3 superforecaster-written prompts shown in the appendix of our paper as Superforecaster Prompt 1-3.

Inference hyperparameters We set the maximum output token length to 2000 to accommodate reasoning and probabilistic forecasts. We initially considered a high token length of 3000, but after observing that the maximum response length was around 1950, we finalized 2000 as the optimal maximum token length. We set the model temperature to 0 to ensure stable outputs.

How to reproduce To run LLM "Ensemble" baseline, follow the below steps:

1. Insert all the necessary API keys in `src/helpers/constants.py`.
2. Place the "news" folder in the same directory as `src/base_eval/llm_crowd/notebook.ipynb`.
3. Run `src/base_eval/llm_crowd/notebook.ipynb`.

K REPRODUCE RESOLUTION AND LEADERBOARD

Given the forecast files output from Appendix I and Appendix J, the forecasts can be resolved and the leaderboard created as outlined below, after first having downloaded the benchmark codebase.

The Google Cloud Run Job in `src/resolve_forecasts/main.py` resolves all forecasts on the questions from the question set in Section B.1. To do this, it depends on:

- the forecast files provided in Section B.2.1 and Section B.2.3;
- the complete resolution files from our Question Bank on GCP Cloud Storage, which we cannot distribute freely because some providers do not allow us to distribute their data directly, rather only modifications of their data. However, the code to create these resolution files is provided under `src/questions` and can be created given the API keys to the data sources.

Having resolved the forecasts for the day, either to ground truth if it was a forecast on a dataset question, or the resolution value or market value for market questions, we can now create the leaderboard. To do this, we use the Google Cloud Run Job defined in `src/leaderboard/main.py`.

L General public survey demographics

We collected demographic information from the 500 human forecasters in the general public survey. Summaries of participants' age, gender, ethnicity, and country of residence are shown in the tables below.

Table 21: Age Distribution

Age	Percentage
18–24 years old	32.0%
25–34 years old	43.4%
35–44 years old	14.4%
45–54 years old	5.4%
Over 55	4.8%

Table 22: Gender Distribution

Gender	Percentage
Male	53.4%
Female	46.2%
Prefer not to say	0.4%

Table 23: Ethnicity Distribution

Ethnicity	Percentage
White	48.6%
Black	33.4%
Mixed	8.6%
Asian	4.4%
Other	3.4%
Prefer not to say	1.6%

We did not collect similar demographic information from the superforecasters participating in the study, but are reasonably certain that the superforecasters in this study are roughly representative of superforecasters as a whole. Describing forecasters previously recruited by Good Judgment Project, Mellers et al. (2015) noted that they "tended to be men (83%) and U.S. citizens (74%), with an average age of 40 years."

M Performance breakdown

Tables Table 25 and Table 26 show the performance of the top LLM—Claude 3.5 Sonnet (using the Scratchpad prompt with freeze values)—compared with Superforecasters, evaluated by forecast category and horizon.