```
Question 1: {question_1}
Question 2: {question_2}
Question 1 Background: {background_1}
Question 2 Background: {background_2}
Question 1 Resolution Criteria: {resolution_criteria_1}
Question 2 Resolution Criteria: {resolution_criteria_2}
Question 1 Current value on {freeze_datetime_1}: {value_at_freeze_datetime_1}
Question 1 Value Explanation: {value_at_freeze_datetime_explanation_1}
Question 2 Current value on {freeze_datetime_1}: {value_at_freeze_datetime_2}
Question 2 Value Explanation: {value_at_freeze_datetime_explanation_2}
Here's some related information from the news that I've collected for Question 1: {retrieved_info_1}
Here's some related information from the news that I've collected for Question 2: {retrieved_info_2}
Question resolution date: {list_of_resolution_dates}
```

Figure 9: Combination prompt that includes information about both non-market questions. The instructions are truncated and can be supplemented with any of the prompts shown above.

```
I want to assess the quality of a forecast question.
Here is the forecast question:  {question}.
Please flag questions that don't seem appropriate by outputting "flag".  Otherwise, if it seems like a
reasonable question or if you're unsure, output "ok."
In general, poorly-defined questions, questions that are sexual in nature, questions that are too
personal,
questions about the death/life expectancy of an individual should be flagged or, more generally,
questions
that are not in the public interest should be flagged.  Geopolitical questions, questions about court
cases,
the entertainment industry, wars, public figures, and, more generally, questions in the public interest
should
be marked as "ok."
Examples of questions that should be flagged:
* "Will I finish my homework tonight?"
* "Metaculus party 2023"
* "Will Hell freeze over?"
* "Heads or tails?"
* "Will I get into MIT?"
* "Will this video reach 100k views by the EOD?"
* "If @Aella goes on the Whatever podcast, will she regret it?"
* "Daily coinflip"
* "Musk vs Zuckerberg:  Will either of them shit their pants on the mat?"
Examples of questions that should NOT be flagged:
* "Will Megan Markle and Prince Harry have a baby by the end of the year?"
* "Will the Brain Preservation Foundation's Large Mammal preservation prize be won by Feb 9th, 2017?"
* "Will there be more novel new drugs approved by the FDA in 2016 than in 2015?"
* "Will Israel invade Rafah in May 2024?"
* "Will Iraq return its ambassador to Iran in the next month?"
* "Tiger Woods Will Win Another PGA Tournament"
* "Will Dwayne Johnson win the 2024 US Presidential Election?"
* "Will Oppenheimer win best picture AND Bitcoin reach $70K AND Nintendo announce a new console by EOY
2024?"
* "Will anybody born before 2000 live to be 150?"
* "Will Taylor Swift get married before Bitcoin reaches $100K USD?"
* "Will Russia's total territory decrease by at least 20% before 2028?"
* "Will Donald Trump be jailed or incarcerated before 2030?"
* "If China invades Taiwan before 2035, will the US respond with military force?"
* "Will there be a tsunami that kills at least 50,000 people before 2030?"
* "Will there be a military conflict resulting in at least 50 deaths between the United States and China
in 2024?"
* "Will an AI system be reported to have successfully blackmailed someone for >$1000 by EOY 2028?"
* "Will Vladimir Putin declare Martial Law in at least 3/4 of Russia before 2025?"
Again, when in doubt, do NOT flag the question; mark it as "ok".
Your response should take the following structure:
Insert thinking:
{{ insert your concise thoughts here }}
Classification:
{{ insert "flag" or "ok"}}
```

Figure 10: Question validation prompt

2. Run the zero-shot and the scratchpad baselines in `src/base_eval/llm_baselines/manager/main.py`.

J.2 SCRATCHPAD WITH INFORMATION RETRIEVAL BASELINE

Prompts We use the same scratchpad prompt as scratchpad baseline with an additional line "We have retrieved the following information for this question: {retrieved_info}" before the instructions begin.

Table 18: Leaderboard: Human question set (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=422)	Market (N=76)	Overall (N=498)			
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%
Claude-3-5-Sonnet-20240620	Anthropic	News	Scratchpad	0.142	0.137	0.139	[0.117, 0.161]	<0.001	26%
Claude-3-Opus-20240229	Anthropic	Freeze values	Zero shot	0.163	0.115	0.139	[0.114, 0.164]	<0.001	23%
Claude-3-5-Sonnet-20240620	Anthropic	News	Superforecaster 2	0.158	0.123	0.140	[0.120, 0.161]	<0.001	25%
GPT-4o	OpenAI	–	Scratchpad	0.161	0.125	0.143	[0.125, 0.160]	<0.001	24%
Claude-3-5-Sonnet-20240620	Anthropic	News	Superforecaster 1	0.150	0.135	0.143	[0.120, 0.166]	<0.001	25%
GPT-4o	OpenAI	News	Scratchpad	0.162	0.126	0.144	[0.122, 0.165]	<0.001	22%
Claude-3-Opus-20240229	Anthropic	Freeze values	Scratchpad	0.160	0.129	0.144	[0.125, 0.163]	<0.001	23%
Mistral-Large-Latest	Mistral AI	Freeze values	Zero shot	0.173	0.117	0.145	[0.121, 0.169]	<0.001	23%
Gemini-1.5-Pro	Google	News with freeze values	Scratchpad	0.163	0.130	0.146	[0.127, 0.165]	<0.001	23%
Gemini-1.5-Pro	Google	–	Scratchpad	0.162	0.131	0.147	[0.129, 0.164]	<0.001	23%
Mistral-Large-Latest	Mistral AI	Freeze values	Scratchpad	0.161	0.133	0.147	[0.129, 0.165]	<0.001	22%
GPT-4-Turbo-2024-04-09	OpenAI	–	Zero shot	0.162	0.137	0.149	[0.129, 0.170]	<0.001	23%
GPT-4-Turbo-2024-04-09	OpenAI	Freeze values	Scratchpad	0.176	0.123	0.150	[0.126, 0.173]	<0.001	26%
GPT-4	OpenAI	Freeze values	Scratchpad	0.174	0.126	0.150	[0.130, 0.170]	<0.001	21%
Gemini-1.5-Pro	Google	Freeze values	Scratchpad	0.162	0.138	0.150	[0.132, 0.169]	<0.001	22%
Claude-3-5-Sonnet-20240620	Anthropic	–	Zero shot	0.145	0.155	0.150	[0.123, 0.177]	<0.001	24%
Gemini-1.5-Pro	Google	News	Scratchpad	0.163	0.141	0.152	[0.133, 0.171]	<0.001	22%
Claude-3-Opus-20240229	Anthropic	News	Superforecaster 1	0.157	0.147	0.152	[0.131, 0.174]	<0.001	22%
GPT-4-Turbo-2024-04-09	OpenAI	News with freeze values	Scratchpad	0.178	0.128	0.153	[0.130, 0.175]	<0.001	25%
GPT-4o	OpenAI	Freeze values	Zero shot	0.197	0.109	0.153	[0.128, 0.178]	<0.001	26%
Llama-3-70b-Chat-Hf	Meta	Freeze values	Scratchpad	0.191	0.116	0.153	[0.135, 0.171]	<0.001	25%
Mixtral-8x22B-Instruct-V0.1	Mistral AI	Freeze values	Scratchpad	0.181	0.127	0.154	[0.137, 0.172]	<0.001	21%
GPT-4-Turbo-2024-04-09	OpenAI	–	Scratchpad	0.176	0.133	0.154	[0.137, 0.171]	<0.001	23%
Gemini-1.5-Pro	Google	Freeze values	Zero shot	0.185	0.124	0.155	[0.127, 0.182]	<0.001	25%
Claude-3-Opus-20240229	Anthropic	–	Scratchpad	0.160	0.149	0.155	[0.136, 0.173]	<0.001	22%
GPT-4	OpenAI	–	Scratchpad	0.174	0.137	0.155	[0.140, 0.171]	<0.001	18%
Qwen1.5-110B-Chat	Qwen	Freeze values	Zero shot	0.197	0.115	0.156	[0.134, 0.179]	<0.001	20%
Claude-2.1	Anthropic	Freeze values	Scratchpad	0.217	0.095	0.156	[0.138, 0.175]	<0.001	24%
Gemini-1.5-Flash	Google	Freeze values	Zero shot	0.191	0.125	0.158	[0.130, 0.186]	<0.001	23%
Mixtral-8x22B-Instruct-V0.1	Mistral AI	Freeze values	Zero shot	0.185	0.131	0.158	[0.131, 0.185]	<0.001	24%
Llama-3-70b-Chat-Hf	Meta	Freeze values	Zero shot	0.194	0.124	0.159	[0.133, 0.185]	<0.001	25%
GPT-4-Turbo-2024-04-09	OpenAI	News	Scratchpad	0.178	0.144	0.161	[0.140, 0.182]	<0.001	24%
Imputed Forecaster	ForecastBench	–	–	0.250	0.073	0.162	[0.142, 0.181]	<0.001	26%
Qwen1.5-110B-Chat	Qwen	Freeze values	Scratchpad	0.183	0.141	0.162	[0.144, 0.180]	<0.001	18%
Claude-2.1	Anthropic	–	Scratchpad	0.217	0.109	0.163	[0.143, 0.183]	<0.001	22%
Qwen1.5-110B-Chat	Qwen	News with freeze values	Scratchpad	0.179	0.148	0.163	[0.144, 0.183]	<0.001	22%
Claude-2.1	Anthropic	Freeze values	Zero shot	0.221	0.108	0.164	[0.141, 0.188]	<0.001	27%
Mistral-Large-Latest	Mistral AI	–	Scratchpad	0.161	0.168	0.165	[0.146, 0.183]	<0.001	22%
Claude-3-5-Sonnet-20240620	Anthropic	News	Scratchpad	0.197	0.133	0.165	[0.144, 0.186]	<0.001	19%
Claude-3-Opus-20240229	Anthropic	–	Zero shot	0.163	0.167	0.165	[0.140, 0.191]	<0.001	18%

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. 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.

Information Retrieval We use the same information retrieval system from Halawi et al. (2024). The pipeline consists of four steps: search query generation, news retrieval, relevance filtering and re-ranking, and text summarization. One must acquire a Newscatcher API key to implement the same retrieval method.

Information Retrieval Hyperparameters The hyperparameters were selected following the results in Section E.1 of Halawi et al. (2024), in which they used a greedy search approach to identify the optimal hyperparameters. We display the hyperparameters below:

`NUM_SEARCH_QUERY_KEYWORDS`: The number of keywords used in the search query. For our system, this is set to 6.

`MAX_WORDS_NEWSCATCHER`: The maximum number of words allowed in search queries for the NewsCatcher API. This is set to 5.

`MAX_WORDS_GNEWS`: The maximum number of words allowed in search queries for the Google News API. This is set to 8.

`SEARCH_QUERY_MODEL_NAME`: The name of the model used to generate search queries. We use `gpt-4-1106-preview`.

Table 19: Leaderboard: LLM question set (top 50)

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%
GPT-4-Turbo-2024-04-09	OpenAI	–	Scratchpad	0.172	0.115	0.143	[0.138, 0.149]	<0.001	31%
GPT-4-Turbo-2024-04-09	OpenAI	Freeze values	Zero shot	0.204	0.084	0.144	[0.137, 0.150]	<0.001	42%
Claude-3-5-Sonnet-20240620	Anthropic	–	Scratchpad	0.169	0.120	0.144	[0.139, 0.150]	<0.001	10%
Gemini-1.5-Pro	Google	–	Scratchpad	0.162	0.128	0.145	[0.139, 0.151]	<0.001	32%
GPT-4	OpenAI	Freeze values	Scratchpad	0.194	0.100	0.147	[0.141, 0.154]	<0.001	36%
Gemini-1.5-Pro	Google	News	Scratchpad	0.166	0.129	0.147	[0.141, 0.153]	<0.001	32%
Imputed Forecaster	ForecastBench	–	–	0.250	0.048	0.149	[0.145, 0.153]	<0.001	46%
GPT-4o	OpenAI	–	Scratchpad	0.186	0.114	0.150	[0.144, 0.156]	<0.001	31%
Gemini-1.5-Pro	Google	Freeze values	Zero shot	0.217	0.083	0.150	[0.144, 0.157]	<0.001	39%
GPT-4-Turbo-2024-04-09	OpenAI	News with freeze values	Scratchpad	0.211	0.091	0.151	[0.145, 0.157]	<0.001	35%
GPT-4o	OpenAI	News	Scratchpad	0.190	0.114	0.152	[0.146, 0.158]	<0.001	31%
Claude-3-5-Sonnet-20240620	Anthropic	News	Scratchpad	0.184	0.124	0.154	[0.148, 0.160]	<0.001	30%
GPT-4	OpenAI	Freeze values	Zero shot	0.222	0.087	0.154	[0.148, 0.161]	<0.001	38%
Qwen1.5-110B-Chat	Qwen	–	Scratchpad	0.176	0.134	0.155	[0.150, 0.160]	<0.001	28%
LLM Crowd	ForecastBench	News	–	0.242	0.068	0.155	[0.151, 0.159]	<0.001	38%
Claude-3-Opus-20240229	Anthropic	Freeze values	Scratchpad	0.201	0.112	0.156	[0.150, 0.162]	<0.001	27%
Mistral-Large-Latest	Mistral AI	Freeze values	Scratchpad	0.199	0.115	0.157	[0.151, 0.162]	<0.001	26%
Gemini-1.5-Pro	Google	–	Zero shot	0.217	0.097	0.157	[0.151, 0.163]	<0.001	37%
LLM Crowd	ForecastBench	News	–	0.243	0.071	0.157	[0.153, 0.161]	<0.001	37%
LLM Crowd	ForecastBench	News	–	0.244	0.071	0.157	[0.153, 0.161]	<0.001	37%
GPT-4	OpenAI	–	Scratchpad	0.194	0.121	0.158	[0.153, 0.162]	<0.001	28%
Llama-3-70b-Chat-Hf	Meta	Freeze values	Zero shot	0.215	0.101	0.158	[0.151, 0.164]	<0.001	33%
Gemini-1.5-Pro	Google	News	Superforecaster 1	0.186	0.131	0.159	[0.153, 0.165]	<0.001	31%
Claude-3-5-Sonnet-20240620	Anthropic	News	Superforecaster 2	0.190	0.129	0.159	[0.153, 0.165]	<0.001	29%
GPT-4-Turbo-2024-04-09	OpenAI	–	Zero shot	0.204	0.117	0.160	[0.154, 0.167]	<0.001	32%
Gemini-1.5-Flash	Google	Freeze values	Scratchpad	0.194	0.128	0.161	[0.154, 0.168]	<0.001	32%
Claude-3-Opus-20240229	Anthropic	–	Zero shot	0.186	0.136	0.161	[0.154, 0.168]	<0.001	35%
GPT-4-Turbo-2024-04-09	OpenAI	News	Superforecaster 2	0.208	0.116	0.162	[0.156, 0.167]	<0.001	28%
GPT-4-Turbo-2024-04-09	OpenAI	News	Scratchpad	0.211	0.114	0.163	[0.157, 0.168]	<0.001	27%
Claude-3-5-Sonnet-20240620	Anthropic	–	Zero shot	0.192	0.134	0.163	[0.156, 0.170]	<0.001	34%
Qwen1.5-110B-Chat	Qwen	News with freeze values	Scratchpad	0.205	0.122	0.164	[0.158, 0.170]	<0.001	26%
Llama-3-70b-Chat-Hf	Meta	Freeze values	Scratchpad	0.221	0.108	0.164	[0.159, 0.170]	<0.001	25%
Mistral-Large-Latest	Mistral AI	Freeze values	Zero shot	0.208	0.122	0.165	[0.157, 0.172]	<0.001	31%
Gemini-1.5-Flash	Google	Freeze values	Zero shot	0.232	0.098	0.165	[0.158, 0.173]	<0.001	40%
Mixtral-8x22B-Instruct-V0.1	Mistral AI	Freeze values	Scratchpad	0.210	0.121	0.165	[0.159, 0.172]	<0.001	30%
GPT-4o	OpenAI	News	Superforecaster 3	0.211	0.124	0.168	[0.162, 0.173]	<0.001	28%
Claude-2.1	Anthropic	–	Scratchpad	0.237	0.100	0.168	[0.163, 0.174]	<0.001	38%
Gemini-1.5-Flash	Google	–	Scratchpad	0.194	0.146	0.170	[0.164, 0.176]	<0.001	28%
GPT-4o	OpenAI	Freeze values	Zero shot	0.225	0.116	0.171	[0.163, 0.178]	<0.001	37%
Claude-3-Opus-20240229	Anthropic	–	Scratchpad	0.201	0.141	0.171	[0.165, 0.177]	<0.001	26%

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. 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.

SEARCH_QUERY_TEMPERATURE: The temperature setting for the search query model, which controls the randomness of the output. We set this to 0.0 for deterministic outputs.

SEARCH_QUERY_PROMPT_TEMPLATES: The templates used to generate search queries. In our configuration, we use `PROMPT_DICT["search_query"]["0"]` and `PROMPT_DICT["search_query"]["1"]`. The exact search query can be found in search_query.py.

NUM_ARTICLES_PER_QUERY: The number of articles retrieved per search query. This is set to 10.

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

SUMMARIZATION_TEMPERATURE: The temperature setting for the summarization model, which controls the randomness of the output. We set this to 0.2.

SUMMARIZATION_PROMPT_TEMPLATE: The template used for summarizing articles. In our configuration, we use `PROMPT_DICT["summarization"]["9"]`. The exact search query can be found in summarization.py.