Field	Information
Question	Will Starship achieve liftoff before Monday, May 1st, 2023?
Start Date	2023-04-17
End Date	2023-04-30
Resolve Date	2023-04-20
Category	Science & Technology
Platform	Metaculus
Resolution	1.0
URL	`https://www.metaculus.com/api2/questions/15973/`
Background	On April 14th, SpaceX received a launch license for its Starship spacecraft. A launch scheduled for April 17th was scrubbed due to a frozen valve. SpaceX CEO Elon Musk tweeted: "Learned a lot today, now offloading propellant, retrying in a few days . . . "
Resolution Criteria	This question resolves Yes if Starship leaves the launchpad intact and under its own power before 11:59pm ET on Sunday, April 30th.
Community Predictions	(2023-04-17, 0.725), (2023-04-17, 0.793), (2023-04-17, 0.71), (2023-04-17, 0.704), (2023-04-17, 0.722), (2023-04-17, 0.754), (2023-04-18, 0.74), (2023-04-18, 0.726), (2023-04-18, 0.707), (2023-04-18, 0.703), (2023-04-18, 0.701), (2023-04-18, 0.698), (2023-04-18, 0.666), (2023-04-18, 0.665), (2023-04-18, 0.668), (2023-04-18, 0.666), (2023-04-18, 0.63), (2023-04-18, 0.636), (2023-04-18, 0.652), (2023-04-18, 0.659), (2023-04-18, 0.663), (2023-04-18, 0.664), (2023-04-18, 0.678), (2023-04-18, 0.687), (2023-04-18, 0.686), (2023-04-18, 0.686), (2023-04-18, 0.686), (2023-04-18, 0.658), (2023-04-18, 0.658), (2023-04-18, 0.664), (2023-04-18, 0.671), (2023-04-18, 0.677), (2023-04-18, 0.685), (2023-04-18, 0.685), (2023-04-18, 0.69), (2023-04-18, 0.689), (2023-04-18, 0.691), (2023-04-18, 0.698), (2023-04-18, 0.706), (2023-04-18, 0.703), (2023-04-18, 0.704), (2023-04-18, 0.706), (2023-04-18, 0.702), (2023-04-18, 0.703), (2023-04-18, 0.704), (2023-04-18, 0.704), (2023-04-18, 0.702), (2023-04-18, 0.702), (2023-04-18, 0.702), (2023-04-18, 0.701), (2023-04-18, 0.701), (2023-04-18, 0.689), (2023-04-18, 0.689), (2023-04-18, 0.686), (2023-04-18, 0.688), (2023-04-18, 0.686), (2023-04-18, 0.684), (2023-04-19, 0.688), (2023-04-19, 0.688), (2023-04-19, 0.689), (2023-04-19, 0.696), (2023-04-19, 0.695), (2023-04-19, 0.699), (2023-04-19, 0.697), (2023-04-19, 0.699), (2023-04-19, 0.702), (2023-04-19, 0.703), (2023-04-19, 0.703), (2023-04-19, 0.705), (2023-04-19, 0.71), (2023-04-19, 0.712), (2023-04-19, 0.713), (2023-04-19, 0.713), (2023-04-19, 0.714), (2023-04-19, 0.714), (2023-04-19, 0.714), (2023-04-19, 0.717), (2023-04-19, 0.717), (2023-04-19, 0.713), (2023-04-19, 0.713), (2023-04-19, 0.713), (2023-04-19, 0.717), (2023-04-19, 0.717), (2023-04-19, 0.716), (2023-04-19, 0.72), (2023-04-19, 0.721), (2023-04-20, 0.721), (2023-04-20, 0.717), (2023-04-20, 0.716), (2023-04-20, 0.715), (2023-04-20, 0.719), (2023-04-20, 0.723), (2023-04-20, 0.725), (2023-04-20, 0.725), (2023-04-20, 0.726), (2023-04-20, 0.726), (2023-04-20, 0.73), (2023-04-20, 0.73), (2023-04-20, 0.728), (2023-04-20, 0.733), (2023-04-20, 0.734)]
Extracted URLs	`https://www.youtube.com/live/-1wcilQ58hI`, `https://twitter.com/nextspaceflight/status/1648797064183128065`, `https://twitter.com/SciGuySpace/status/1648498635355865089`, `https://twitter.com/nextspaceflight/status/1648425030018293760`, `https://twitter.com/SpaceX/status/1648092752893313024`

Table 12: **A sample question from our dataset with all its fields** (full version of Table 1). Each data point consists of the following fields: question, start date, end date, resolve date, the final resolution, question category, platform, URL, background, resolution criteria, community predictions, and extracted URLs (from the background and comment section). The resolution is not presented to the model. We do not use the URLs that are extracted from the comment section, since certain comments may be made after the resolution.

Sample Question	**Category**	**Start Date**	**Close Date**	**Resolution Date**	**25%**	**50%**	**90%**	**Answer**
Will AI doctors replace human doctors by the end of 2023?	Science & Tech	2023-07-27	2023-12-31	2023-12-30	0.073	0.087	0.102	No
Will US CDC classify a SARS-CoV-2 variant as "high consequence" by August 1, 2022?	Healthcare & Biology	2021-07-31	2021-11-01	2022-08-02	0.39	0.408	0.412	No
Will Coinbase file for bankruptcy in 2022?	Economics & Business	2022-05-12	2022-12-31	2023-01-01	0.08	0.079	0.072	No
Will COP26 finalize the "Paris Rulebook" by November 16, 2021?	Environment & Energy	2021-08-26	2021-11-13	2021-11-14	0.063	0.074	0.13	Yes
Will Bongbong Marcos win the 2022 Philippine Presidential Election?	Politics & Governance	2021-12-20	2022-05-08	2022-05-26	0.759	0.752	0.759	Yes
Will UC Berkeley be primarily in-person for Fall 2021?	Education & Research	2021-01-22	2021-08-01	2021-08-26	0.723	0.74	0.765	Yes
Will Trump issue another NFT Collection before the 2024 Presidential Election?	Arts & Recreation	2023-11-01	2023-12-12	2023-12-12	0.484	0.585	0.556	Yes
Will a nuclear weapon be detonated in 2023 (including tests and accidents)?	Security & Defense	2022-12-09	2023-12-31	2024-01-01	0.28	0.32	0.304	No
Will Charlotte Hornets beat Detroit Pistons on Oct 27, 2023, in the NBA?	Sports	2023-10-16	2023-10-28	2023-10-28	0.46	0.513	0.337	No
Will flight 1111 from Munich to Zurich on 2023-08-29 arrive on time or with more than 30 mins delay?	Other	2023-08-27	2023-08-29	2023-08-29	0.617	0.734	0.809	Yes

Table 13: **One sample question from each category** along with the community's predictions at different prediction dates (25%, 50%, and 90% from the start date to resolve date). As the questions approach their resolution dates, the crowd's confidence in the outcome generally increases, reflecting the influence of new information.

D Details about Our System

We provide details about our system, described at a high-level in Section 4. We specify the hyperparameters used in our (optimized) settings. Some of them are discovered by the hyperparameter sweep (Section 5.2).

D.1 Retrieval System

Our retrieval system consists of 4 steps. We provide further details on each below.

Step 1: Search query generation. We identify two good prompts to generate search queries in our hyperparameter sweep procedure, listed in Figure 12. Given a question, we ask GPT-4-Preview-1106 to generate 6 search queries using both prompts (at 0 temperature). We take the union of all the resulting search queries along with the question itself to query the news API's.

Step 2: News retrieval. On each news API and each search query, our system is set to retrieve the top 10 articles published within a given retrieval date range. We use the default ranking of each API and only retrieve English-language articles.

In cases where the background description of a question contains links to webpages, our system scrapes them, parses the clean texts, and presents the summaries to the reasoning model. We take measures to ensure that this leaks no information beyond the retrieval range. First, we maintain a whitelist of news websites that publish timestamped articles and only retrieve from the whitelist. Second, our system checks the publish date of each article and discard it if the date is not available or outside the retrieval range.

Step 3: Relevance ranking. We use GPT-3.5-Turbo to rate the relevance of an article with respect to a question at 0 temperature. The prompt is given in Figure 14.

Our retrieval system can retrieve a large number of texts (say, > 50 articles) at the initial stage prior to relevance filtering. To improve the run-time and save cost, we only present the article's title and its first 250 words to the model in context for relevance rating. In Section E.3, we test that this well approximates the result from giving the full texts.

The system rates the relevance of each retrieved article at the scale of 1–6. Any article that receives a rating of ≤ 3 is discarded. We do not make an attempt to optimize this threshold or the prompt choice here.

Step 4: Summarization. We use GPT-3.5-Turbo to summarize the relevant articles. The temperature is set to be 0.2. In cases where the article length exceeds the context window, we simply truncate it to fit the window size. We remark that our prompt (Figure 13) also contains the question and its background description, and the model is instructed to keep any information in the article that is relevant to answering the question. Figure 13 shows the best prompt found via hyperparameter sweep on the validation set (Section 5.2).

D.2 Reasoning System

We use both GPT-4-1106-Preview and our fine-tuned GPT-4 to generate forecasts. We prompt the former with our top 3 reasoning prompts, including Figure 15. The other prompts also conform to the basic template as shown in Figure 16, though with different scratchpad reasoning instructions following the retrieved information section. The fine-tuned model does not require detailed scratchpad instructions (Section 5.1). Thus, Figure 16 is the entire prompt structure to the fine-tuned model to elicit its reasonings.

In addition, as we remarked in Section 5.1, Claude-2.1 was prompted to generate reasoning-prediction pairs for fine-tuning. However, it is not directly used for reasoning in our system.

Question: {question}

Background: {background}

Resolution criteria: {criteria}

Today's date: {date_retrieval}
Question close date: {date_end}

We have retrieved the following information:
{retrieved_info}

Figure 16: **All scratchpad prompts begin with a question's basic information, followed by retrieval.** The fine-tuned model only takes this information and requires no further instructions.