Approaching Human-Level Forecasting with Language Models

Danny Halawi[*]   **Fred Zhang**[*]   **Chen Yueh-Han**[*]

Jacob Steinhardt
UC Berkeley
{dhalawi, z0, john0922ucb, jsteinhardt}@berkeley.edu

Abstract

Forecasting future events is important for policy and decision making. In this work, we study whether language models (LMs) can forecast at the level of competitive human forecasters. Towards this goal, we develop a retrieval-augmented LM system designed to automatically search for relevant information, generate forecasts, and aggregate predictions. To facilitate our study, we collect a large dataset of questions from competitive forecasting platforms. Under a test set published after the knowledge cut-offs of our LMs, we evaluate the end-to-end performance of our system against the aggregates of human forecasts. On average, the system nears the crowd aggregate of competitive forecasters, and in some settings surpasses it. Our work suggests that using LMs to forecast the future could provide accurate predictions at scale and help to inform institutional decision making.

1 Introduction

Forecasting events is important in the modern world. Governments rely on economic and geopolitical forecasts for decision-making. Companies hire and invest based on forecasts of market conditions (Armstrong, 2001). In 2020, epidemiological forecasts for COVID-19 prompted national lockdowns across the globe (Adam, 2020).

There are two main approaches to forecasting. *Statistical forecasting* primarily uses tools from time-series modeling. This methodology typically excels when data are abundant and under minimal distributional shift. By contrast, in *judgmental forecasting*, human forecasters assign probabilities to future events based on their own judgments, making use of historical data, domain knowledge, Fermi estimates, and intuition. They draw information from diverse sources and reason based on detailed contexts of the task. This enables accurate forecasts even with scarce past observations or under significant distributional shift (Tetlock and Gardner, 2015). We will refer to judgmental forecasting simply as "forecasting".

Since forecasting relies on human effort and expertise, it can be expensive, delayed, or applicable only in specific domains. Moreover, most human forecasts contain little or no explanatory reasoning. These limitations motivate using language models (LMs) to automate forecasting (Hendrycks et al., 2021). Because they can parse and produce texts rapidly, LMs can provide cheap and timely forecasts. Because they are pre-trained on web-scale data, they are endowed with massive, cross-domain knowledge. And because we can elicit their reasonings through prompts, we can examine them to (partially) understand the final forecast.

In this work, we build a LM pipeline for automated forecasting, with a focus on predicting binary outcomes. Our system implements and automates three key components in the traditional forecasting

---

[*]Equal Contribution.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

(a) **Our retrieval system**. The LM takes in the question and generates search queries to retrieve articles from historical news APIs. Then the LM ranks the articles on relevancy and summarizes the top k articles.

(b) **Our reasoning system**. The system takes in the question and summarized articles and prompts LMs to generate forecasts. The forecasts are then aggregated into a final forecast using the trimmed mean.

Figure 1: **Overview of our retrieval and reasoning systems**. Our retrieval system retrieves summarized new articles and feeds them into the reasoning system, which prompts LMs for reasonings and predictions that are aggregated into a final forecast.

process: (1) retrieval, which gathers relevant information from news sources; (2) reasoning, which weighs available data and makes a forecast; and (3) aggregation, which ensembles individual forecasts into an aggregated prediction. Each step makes use of an LM or a collection of LMs which is either prompted or fine-tuned (Figure 1).

To optimize and evaluate our system, we collect a large dataset of forecasting questions from 5 competitive forecasting platforms. The test set consists only of (binary) questions published after June 1st, 2023. Since this is after the knowledge cutoff date of our models, this prevents leakage from pre-training. The train set contains questions before June 1st, 2023, which we use for hyperparameter search and fine-tuning our system.

We use a self-supervised approach to fine-tune a LM to make accurate predictions and explanatory reasonings. We first prompt a base LM with various scratchpads to elicit forecasts to questions in our training set. We then fine-tune a new LM on the outputs that outperformed the crowd, which teaches the model what reasoning method to apply in a given context and improves forecasting performance. For hyperparameter search, we identify system configurations, including retrieval and LM prompting strategies, that lead to the best end-to-end performance.

Our optimized system approaches the performance of aggregated human forecasts over the test set, as measured by Brier score, a standard metric in forecasting. To our knowledge, this is the first automated system with forecasting capability that nears the human crowd level, which is generally stronger than individual human forecasters (Section 3.1). We also consider a selective setting where our system uses heuristics, based on the LM's strengths, to decide whether to submit a forecast for a given question and date. In this setting, our system outperforms the human crowd.

2 Related Work

Event forecasting. Machine learning systems that make accurate, automated forecasts can help inform human decision-making (Hendrycks et al., 2021). Jin et al. (2021) provided ForecastQA, the first dataset for this task, which contains questions created by crowdworkers based on events from news articles. Zou et al. (2022) introduced Autocast, a benchmark dataset compiled from forecasting competition questions up to 2022. In a competition with a large prize pool, no machine learning system was able to approach the performance of human forecasters on Autocast (Zou et al., 2022). The knowledge cut-offs of the latest LMs have moved past 2022, necessitating more recent data. In this work, we source questions in 2023–2024, enabling us to apply recent LMs.

Yan et al. (2024) built a retrieval system that led to improved accuracy on Autocast. They trained a Fusion-in-Decoder model to directly predict the final (binary) resolution (Izacard and Grave, 2021) and reported accuracy, whereas we elicit both explanatory reasonings and probability forecasts from LMs and measure performance with the standard Brier score metric.

2

Field	Content
Question	Will Starship achieve liftoff before Monday, May 1st, 2023?
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.
Key Dates	Begin Date: 2023-04-17 | Close Date: 2023-04-30 | Resolve Date: 2023-04-20

Table 1: **A sample question** with its background, resolution criteria, and key dates. The question resolved early (with a final resolution of Yes). See Table 10 for the complete sample point.

Schoenegger and Park (2023) and Abolghasemi et al. (2023) evaluated GPT-4 and other LLMs on forecasting tournaments and found that they underperform the human crowd. This observation is in line with ours in Section 3.3. Unlike us, they make little efforts to improve these LMs on forecasting.

Finally, there has been recent work on using transformer models or LMs for statistical time-series forecasting (Nie et al., 2023; Gruver et al., 2023; Dooley et al., 2023; Rasul et al., 2023; Jin et al., 2024; Das et al., 2024; Woo et al., 2024), but this is distinct from our focus on judgmental forecasting.

Information retrieval (IR). IR can improve question-answering capabilities of LMs (Lewis et al., 2020; Shuster et al., 2021; Nakano et al., 2021). Access to diverse, up-to-date information is crucial in forecasting (Tetlock and Gardner, 2015). Thus, a key component of our system is an IR architecture that furnishes the reasoning model with news articles, using LMs for query writing, ranking and summarization. Beyond our setting, using LMs for IR is an active research topic (Zhu et al., 2024).

Calibration. Calibration is important for accurate forecasting (Tetlock and Gardner, 2015). On competitive platforms, forecasters are evaluated by proper scoring rules, such as Brier score (Brier, 1950), which incentivize calibration (Gneiting and Raftery, 2007). There is a vast literature on calibration in deep learning; see Gawlikowski et al. (2021) and Wang (2023) for surveys.

3 Preliminaries: Data, Models and Baseline

3.1 Dataset

Data format. Forecasting platforms such as Metaculus, Good Judgment Open, INFER, Polymarket, and Manifold invite participants to predict future events by assigning probabilities to outcomes of a question. Each question consists of a *background description*, *resolution criterion*, and 3 timestamps: a *begin date* when the question was published, a *close date* when no further forecasts can be submitted, and (eventually) a *resolve date* when the outcome is determined (Table 1). A forecast can be submitted between the begin date and min(resolve date, close date).

Crowd prediction. On any given question, as individual forecasts are submitted, forecasting platforms continuously aggregate them into a crowd prediction (see Section A.4 for details about the aggregation mechanisms on different platforms). The crowd prediction is a strong benchmark to compete with. For example, Metaculus (2023) shows that an ensemble of all forecasters consistently outperforms using just the top 5, 10, ..., 30 best forecasters (based on past scores). In this work, we compare our system performance to the crowd aggregates.

Raw data. We source forecasting questions from the 5 above-mentioned platforms. This yields a total of 50,343 questions and 6,534,042 user forecasts spanning from 2015 to 2024. The dataset includes 33,664 binary questions, 9,725 multiple-choice questions, 4,019 numerical questions, and 1,346 questions of other types. The questions cover a wide range of topics across the globe (Figure 8).

The raw dataset contains questions that are ill-defined, overly personal, or of niche interests. Furthermore, recent questions are highly unbalanced, with over 80% of questions since June 1, 2023 coming from Manifold and Polymarket.

Data curation. To address the above issues, we curate a subset by filtering ill-defined questions and removing questions that received few forecasts or trading volume on Manifold and Polymarket. We focus on predicting binary questions and split multiple-choice questions into binary ones.

3

Platform	Train	Validation	Test		Model	Zero-shot	Scratchpad
Metaculus	1,576	230	275	GPT-4-1106-Preview	$\mathbf{.208}_{.006}$	$\mathbf{.209}_{.006}$
GJOpen	806	161	38	Llama-2-13B	$.226_{.004}$	$.268_{.004}$
INFER	52	50	4	Mistral-8x7B-Instruct	$.238_{.009}$	$.238_{.005}$
Polymarket	70	229	300	Claude-2.1	$.220_{.006}$	$.215_{.007}$
Manifold	1,258	170	297	Gemini-Pro	$.243_{.009}$	$.230_{.003}$
All Platforms	**3,762**	**840**	**914**	Trimmed mean	$.208_{.006}$	$.224_{.003}$

(a) Dataset distribution across platforms.	(b) Baseline performance of pre-trained models.

Table 2: **(a) Distribution of our train, validation, and test sets across all 5 forecasting platforms.** **(b) Baseline performance** of pre-trained models on the test set (see full results in Table 6). Subscript numbers denote 1 standard error. Random baseline: 0.25; human crowd: 0.149.

To guard potential leakage from LMs' pre-training, we only include questions in the test set that appear after the knowledge cut-off for the models we use (June 1, 2023). All test set questions were opened after the date, and all train and validation questions were resolved before. Questions that span across the date are discarded. This yields a set of 5,516 binary questions, including 3,762 for training, 840 for validation, and 914 for testing (Table 2a). We show a sample data point in Table 10 and provide all details about the curation process in Appendix C.

3.2 Evaluation

Retrieval schedule. We can simulate forecasting the future by using the fact that models are only trained up to a cut-off date (Zou et al., 2022). To simulate a forecast for a resolved question, we query a historical news corpus to retrieve articles between the question begin date and a specified *retrieval date* (Zou et al., 2022; Yan et al., 2024). The retrieval date can be viewed as the "simulated date" of the forecast, as we are mimicking the information the model would have had access to on that date.

To create a set of retrieval dates for each question, we use geometrically increasing time points between the open and close dates (see Section A.1 further motivations). Specifically, we take $n = 5$ retrieval dates per question, where the kth date is given by

$$\text{retrieval\_date}_k = \text{date}_{\text{begin}} + (\text{date}_{\text{close}} - \text{date}_{\text{begin}} - 1)^{k/n}. \quad (1)$$

For questions that resolve before they close, we exclude all retrieval dates occurring after the question has been resolved. Under this geometric retrieval schedule, we retain 86% of retrieval dates on average across all questions. The average question window in our corpus is approximately 70 days, and the average time until resolution is 42 days.

Metric. Our work focuses on binary questions and uses the Brier score as the performance metric, defined as $(f - o)^2$, where $f \in [0, 1]$ is the probabilistic forecast and $o \in \{0, 1\}$ is the outcome. The Brier score is a strictly proper scoring rule: assuming the true probability that $o = 1$ is p, the optimal strategy is to report $f = p$. This is desirable, since improper scoring rules would incentivize reporting distorted probabilities. As a baseline, an (unskilled) forecast of .5 attains a Brier score of .25.

To compute the final Brier score, we first average the Brier scores across retrieval dates for each question, then average across questions. We also report standard errors; however, note that the computation of standard errors assumes the data are i.i.d., while our data are in fact time-series, so this likely underestimates the true error. We also report root mean square (RMS) calibration error.

Models. We evaluate 14 instruction-tuned LMs (Section A.2). Our main system relies on GPT-3.5, GPT-4-0613, and GPT-4-1106-Preview, all of which have a knowledge cut-off of April 2023 or before. To ensure that the models have no knowledge after this date, we submit them to a knowledge test where we prompt models to answer questions about major events after April 2023. In this test, we find no evidence of leakage (see Section A.3 for detailed methodology and results).

3.3 Models are not naturally good at forecasting

As a baseline, we evaluate all 14 LMs with no additional information retrieval. We use zero-shot prompts and scratchpad prompts (Nye et al., 2021). For each prompting strategy, we craft candidate

4

Criteria	Brier Score ↓			% Accuracy ↑			% Data Retained ↑	
	Ours	**Crowd**	**Aggregate**	**Ours**	**Crowd**	**Aggregate**	**Forecasts**	**Questions**
All Questions	$.179_{.003}$	$.149_{.003}$	$\mathbf{.146}_{\mathbf{.002}}$	$71.5_{.7}$	$77.0_{.7}$	$\mathbf{77.8}_{\mathbf{.6}}$	100%	100%
Crowd Uncertain	$\mathbf{.238}_{\mathbf{.004}}$	$.240_{.003}$	$\underline{\mathbf{.233}}_{\mathbf{.002}}$	$58.1_{1.3}$	$58.3_{1.3}$	$\underline{\mathbf{60.2}}_{\mathbf{1.2}}$	51%	56%
Early Retrieval	$.186_{.003}$	$.162_{.004}$	$\underline{\mathbf{.159}}_{\mathbf{.003}}$	$70.0_{.9}$	$74.4_{.9}$	$\underline{75.0}_{.8}$	84%	100%
5+ Articles	$.175_{.003}$	$.142_{.003}$	$\underline{\mathbf{.140}}_{\mathbf{.002}}$	$72.3_{.8}$	$77.7_{.7}$	$\underline{78.7}_{.7}$	84%	94%
All Criteria	$\mathbf{.240}_{\mathbf{.005}}$	$.247_{.004}$	$\underline{\mathbf{.237}}_{\mathbf{.003}}$	$\underline{\mathbf{58.0}}_{\mathbf{1.7}}$	$54.2_{1.7}$	$\mathbf{56.6}_{\mathbf{1.7}}$	22%	43%

Table 3: **System performance** on the test set. "All Questions": Brier score on full test set. Other rows: selective evaluation when criteria are met. "Crowd Uncertain": crowd predictions 0.3-0.7. "Early Retrieval": first 3 retrieval dates. "5+ Articles": ≥5 relevant articles. "All Criteria": all 3 criteria met. System-crowd aggregate performs best in all settings. Subscripts: 1 standard error. Bold: outperforms crowd aggregate. Underlined: best in category.

prompts, pick the best prompt on the validation set, and report its Brier scores on the test set. The prompt choices appear in Figure 4 and Figure 5 and further details are in Appendix B.

None of the models are naturally good at forecasting (Table 2b, Table 6). Most models' scores are around or worse than random guessing (.25). Only the GPT-4 and Claude-2 series beat the unskilled baseline by a large margin ($>$ .02). Moreover, while GPT-4-1106-Preview achieves the lowest Brier score of .208, it trails significantly behind the human crowd performance of .149.

4 Our System

LMs perform poorly in the baseline setting (Table 2b). Forecasting requires models to have access to detailed, up-to-date contextual information and the ability to effectively utilize this information to generate accurate probabilities. Our system addresses this challenge via news retrieval, and elicits better reasoning via optimized prompting and fine-tuning.

4.1 Retrieval

Our retrieval system consists of 4 steps: search query generation, news retrieval, relevance filtering and re-ranking, and text summarization (Figure 1a).

To generate search queries, we initially tried a simple query expansion prompt (Figure 12a), instructing the model to create queries based on the question and its background. However, we find that this overlooks sub-considerations that often contribute to accurate forecasting. To achieve broader coverage, we prompt the model to decompose the forecasting question into sub-questions (Figure 12b) and use each to generate a search query (Min et al., 2019). For instance, when forecasting election outcomes, the first approach searches directly for polling data, while the latter creates sub-questions that cover campaign finances, economic indicators, and geopolitical events. We combine both approaches for comprehensive coverage.

We next retrieve articles from news APIs using LM-generated search queries. We evaluate 5 APIs on the relevance of retrieved articles and select NewsCatcher and Google News (Section F.3).

Our initial retrieval provides wide coverage at the cost of obtaining some irrelevant articles. To ensure that they do not mislead the model at the reasoning step, we prompt GPT-3.5-Turbo to rate the relevancy of all articles (Figure 13) and filter out low-scoring ones. Since the procedure is costly in run-time and budget, we only present the article's title and first 250 words to the model in context. We validate that this approach achieves high recall and precision while saving 70% cost (see Section F.4 for alternative methods and results).

Since LMs are limited by their context window, we summarize the articles. In particular, we prompt GPT-3.5-Turbo to distill the most relevant details from each article with respect to the forecasting question (Figure 14). Finally, we present the top k article summaries to the LM, ordered by their relevancy. We choose the ranking criterion, article count k, and summarization prompt based on end-to-end Brier scores over the validation set (Section 5.2).

5

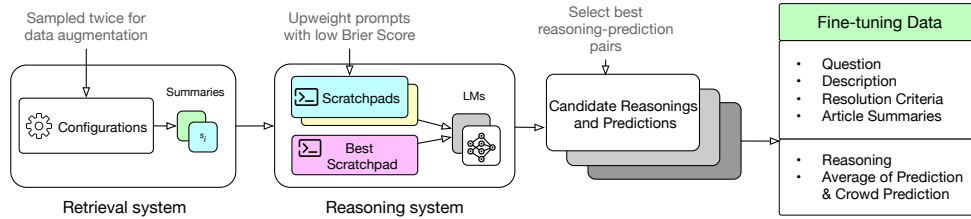



Figure 2: **Our method of self-supervised training.** For each question, the method produces multiple candidate reasoning-predictions and selects those that outperform human aggregates for fine-tuning.

4.2 Reasoning

Prior work in forecasting has focused on eliciting predictions from models without requiring rationales (Zou et al., 2022; Yan et al., 2024). However, accurately predicting the future is a difficult task that often requires computation beyond a single forward pass. Having the model externalize its reasoning also allows us to understand the explanation for the forecast and improve it accordingly.

We use open-ended scratchpad to structure model's reasoning paths. Our prompt begins with posing the question, providing a description, and specifying resolution criteria and key dates, followed by the top k relevant summaries from our retrieval system (see Figure 10 for the basic template).

Qualitatively, we observe several failure modes from our baseline LM evaluations (Section 3.3), including misunderstanding of the question and overconfidence in its output. The optimal reasoning prompt (Figure 11), as identified by hyperparameter sweep (Section 5.2), incorporates four additional components inspired by the literature on forecasting and LM reasoning to fix these issues:

- First, to ensure that the model comprehends the question, we prompt it to rephrase the question. It is also instructed to expand the question with its own knowledge. Intuitively, a more detailed and precise phrasing of the question elicits better responses (Deng et al., 2023).
- Forecasting requires a holistic consideration of the possibilities (Tetlock and Gardner, 2015). We next prompt the model to leverage the retrieved information and its pre-training knowledge to produce arguments for why the outcome may or may not occur.
- The model can potentially generate weak arguments. To avoid treating all considerations as equal, it is instructed to weigh them by importance and aggregate them into an initial forecast.
- Finally, to prevent potential bias and miscalibration, the model is asked to check if it is over- or under-confident (Tian et al., 2023) and to consider historical base rates (Tetlock and Gardner, 2015), prompting it to calibrate and amend the prediction accordingly.

Base model. We prompt GPT-4-1106-Preview with the best scratchpads, since it consistently gives the lowest Brier scores among the LMs we test (Section 5.2).

Fine-tuned model. We also prompt a fine-tuned version of GPT-4 that we trained to generate reasonings with accurate predictions (Section 5.1). We prompt it with only the question's basic information (no scratchpad instructions) since our fine-tuned model is trained to reason without prescriptive instructions.

Ensembling. Since the aggregate of predictions is usually superior to individual forecasts (Tetlock and Gardner, 2015), we elicit multiple predictions from the base and fine-tuned models. We prompt GPT-4-1106-Preview with the optimal scratchpad prompt (Figure 11), along with the 2 next best scratchpad prompts identified in Section 5.2. For our fine-tuned model, we set temperature $T = 0.5$ and prompt it 3 times to sample 3 additional forecasts. This gives us 6 forecasts in total: 3 from the base model, and 3 from the fine-tuned model.

Given these forecasts, the system ensembles them into a final prediction by taking a trimmed mean, as it performs best on the validation set among the ensemble methods we test (Section F.1). We provide further details about our system in Appendix D, including hyperparameters and prompt designs.

5 Optimizing the System

We now describe the procedure to optimize our retrieval and reasoning system and the results obtained.

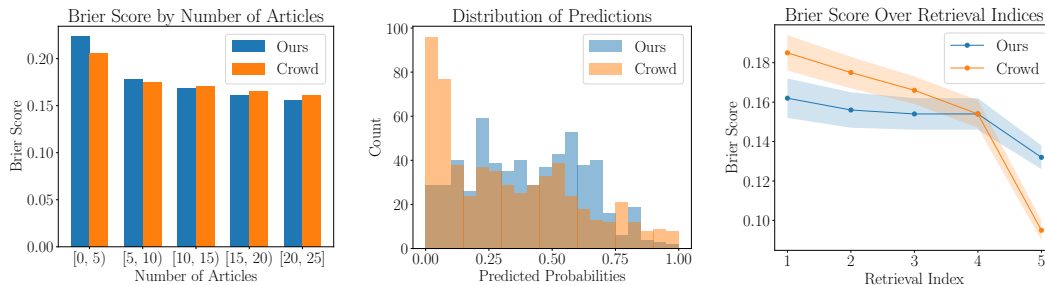
6

(a) Brier Score w.r.t. # of Articles   (b) Distribution of predictions.   (c) Brier score over retrieval indices.

Figure 3: **System strengths**. Our system outperforms the crowd on the validation set when: **(a)** given sufficient relevant articles, **(b)** the crowd is uncertain (predictions between .3 and .7), where we achieve a better Brier score (.199 vs. .246), and **(c)** at earlier retrieval dates.

5.1 Fine-tuning a Reasoning Model

Fine-tuning language models for judgmental forecasting is challenging due to the scarcity of high-quality human-written reasonings. Moreover, directly fine-tuning on outcomes can lead to miscalibration. To address these issues, we propose a self-supervised procedure inspired by recent work on distillation and self-training (Zelikman et al., 2022; Gulcehre et al., 2023; Singh et al., 2024; Agarwal et al., 2024; Gu et al., 2024). We generate data for fine-tuning through a two-step process: (1) collecting a large set of LM-produced reasonings on the train set, and (2) selecting a subset where the model outperforms the human crowd within a fixed margin to avoid overconfidence.

Collecting fine-tuning data. To generate the preliminary data, we run our system at each retrieval date in the retrieval schedule and on each question in the train set. To elicit a diverse set of reasonings for each question, we vary the prompts, retrieval configurations, and other hyperparameters. See Appendix E for details.

Selecting fine-tuning data. We seek to fine-tune our model on strong forecasts. To select the data, we only keep outputs that give a lower Brier score than the crowd's. However, this can inadvertently cause overconfidence in our fine-tuned model. To mitigate this, we discard pairs where the prediction deviates by more than 0.15 from the crowd prediction, and we also average our prediction with the crowd prediction when constructing the target output.

The resulting fine-tuning data has the following structure (Figure 2, right): The **input** to the model consists of the question, description, and resolution criteria, followed by summarized articles; and the **target output** consists of a reasoning and a prediction.

Importantly, the fine-tuning input excludes the scratchpad instructions. By doing so, we directly teach the model which reasoning to apply in a given context. In total, 73,632 reasonings are generated from which 13,253 meet the above desiderata. Finally, we fine-tune GPT-4-0613[1] only on the 6,000 most recent points for 2 epochs, due to budget constraint (Figure 2, right).

5.2 Hyperparameter Sweep

Our hyperparameter sweep optimizes an (intermediate) metric over a discrete set of choices, such as LM prompts and the number of articles presented to the reasoning model in context.

We divide the hyperparameters into groups of 1-2 and optimize them iteratively. For each group, we select the best configuration based on the average Brier score on the validation set, except for search query generation where we use proxy metrics for efficiency. See Appendix F for details and results.

6 Evaluations

We evaluate our optimized system on the test set and find that it comes close to human crowd performance (Section 6.1). We then analyze its strengths and weaknesses (Section 6.2). Motivated by the results, we introduce a relaxed setting, where the system may make forecasts selectively (given its

---

[1]While the latest GPT-4 has 2 years of more recent knowledge, its fine-tuning access is not available.

7

identified strengths), and find that our system surpasses the crowd aggregate (Section 6.3). Finally, we demonstrate how our system can be used to complement aggregated human forecasts (Section 6.4).

6.1 System Nears Human Performance

We first evaluate the Brier score of our end-to-end system on the test set. Note that all hyperparameters were chosen based on the validation set and all test set questions appear temporally after the validation questions, mirroring the setting of a real-time forecasting competition. In addition to the Brier score, we also report accuracy to compare with past work (Zou et al., 2022; Yan et al., 2024).

As the main result, our averaged Brier score is .179, while the crowd achieves .149, resulting in a difference of .03. In comparison with the baseline evaluation (Section 3.3), our system's Brier score (.179) significantly outperforms the best baseline model (.208 with GPT-4-1106-Preview). Further detailed results across different platforms and categories can be found in Table 14.

In prior work, Zou et al. (2022) evaluated their system on the forecasting dataset Autocast, which consists of questions from 3 of the platforms we use: Metaculus, INFER, and GJOpen. They achieved an accuracy of 65.4% compared to a community baseline of 92.8%. Yan et al. (2024) later improved this to 67.9%. We achieve a better accuracy of 71.5% even though the questions we consider are harder, with a significantly lower crowd accuracy of 77.0%.

Finally, on the test set, we observe that our system is well calibrated (Figure 15c) with RMS calibration error .42 (human crowd: .38). Interestingly, this is not the case in the baseline evaluations (Section 3.3), where the models are *not* well calibrated in the zero-shot setting (Figure 15b). Through fine-tuning and ensembling, our system improves the calibration of the base models, without undergoing specific training for calibration.

6.2 System Strengths and Weaknesses

We next seek to understand our system's strengths and weaknesses on the validation set, and later use these insights to improve performance on the test set (Section 6.3). We find that our system performs best relative to the crowd when (1) the crowd is less confident, (2) at earlier retrieval dates, and (3) when it retrieves many articles.

First, our system outperforms the crowd when their prediction ranges between .3 and .7 (Figure 3b). However, it underperforms when the crowd is highly certain, possibly due to the model's tendency to hedge predictions as a result of its safety training (Schulman (2023); see Figure 16 for a qualitative example). For retrieval dates, our system outperforms the crowd on earlier dates but not on later ones, with its Brier score improving slowly relative to the crowd as questions approach resolution (Figure 3c). This may be attributed to the model's hedging even when evidence becomes decisive. The system's performance nears and eventually surpasses the crowd's when there are at least 5 relevant articles, indicating that it relies on high-quality retrieval for better performance (Figure 15a).

6.3 System Beats Crowd in the Selective Setting

In real-word forecasting competitions, forecasters do not have to make predictions on every question in the platform at every possible date. Instead, they typically make predictions on questions where they have expertise or interest in and at times that they choose. Thus, it is natural to use our system's strengths and decide accordingly if we should forecast on a retrieval date k for a question q.

Leveraging the insights from Section 6.2, we make selective forecasts under the following conditions: (1) the crowd's probability estimate is between 0.3 and 0.7, (2) the retrieval date is within the first 3 out of 5 total retrieval dates, and (3) at least 5 relevant articles are retrieved. The gap in Brier score between our system and the crowd shrinks under each heuristic except the third one (Table 3).

Under the first heuristic, we outperform the crowd by a small margin (.238 vs. .240). This is valuable as our system can be used to complement the crowd's prediction when there is greater uncertainty. When all three conditions are jointly met, our system beats the crowd significantly (by more than 1.5 standard errors in both Brier score and accuracy).

6.4 System Complements the Crowd

Finally, aggregates of our system with the crowd forecasts outperform either one in isolation. By combining the system's predictions with the crowd using a weighted average — 4x weight for the

8

Criteria	Brier Score ↓		% Accuracy ↑	
	Ours	**Aggregate**	**Ours**	**Aggregate**
Full System	$.179_{.003}$	$\mathbf{.146}_{.002}$	$71.5_{.7}$	$\mathbf{77.8}_{.6}$
Fine-tuned GPT-4-0613	$.182_{.002}$	$\mathbf{.146}_{.002}$	$70.7_{.7}$	$\mathbf{77.4}_{.6}$
Fine-tuned GPT-3.5	$.183_{.002}$	$\mathbf{.146}_{.002}$	$71.5_{.7}$	$\mathbf{77.4}_{.6}$
Base GPT-4-1106-Preview	$.186_{.002}$	$\mathbf{.148}_{.002}$	$70.6_{.7}$	$\mathbf{77.1}_{.6}$
Base GPT-4-1106-Preview; no IR	$.206_{.002}$	$.150_{.002}$	$66.6_{.7}$	$76.9_{.6}$

Table 4: **Ablation results:** Fine-tuning GPT-3.5 has similar performance to fine-tuning GPT-4-0613 **(rows 2–3)**. Our system degrades without fine-tuning **(row 4)** or retrieval **(row 5)**, as expected. "Aggregate" is the weighted average with crowd prediction. Subscripts are standard errors; bold entries beat the human crowd.

crowd, which we find optimal on the validation set — we improve the overall Brier score from .149 to .146 on the full test set (Table 3, top row).

Moreover, our system excels under certain criteria (Section 6.2). It is especially useful in these cases to supplement the crowd prediction. We report these results in Table 3, using an unweighted average. This outperforms the crowd prediction in all cases: For example, the crowd Brier score is .24 for questions that have a crowd prediction between .3 and .7, but when averaging with our system's prediction the Brier score decreases to .233.

7 Ablations

We conduct three ablation studies to investigate the factors contributing to our system's performance. The first study validates that our performance is not solely due to the power of our base model, GPT-4. The second and third studies demonstrate the effectiveness of our retrieval and fine-tuning methods.

Fine-tuning a less capable model. To demonstrate that our system's performance does not depend entirely on the capabilities of the base model (i.e., GPT-4), we fine-tune GPT-3.5 on all our fine-tuning data and compare its performance to that of fine-tuned GPT-4-0613. We find that the Brier score of the system using fine-tuned GPT-3.5 is only slightly worse at .183, compared to .182 achieved by the system using fine-tuned GPT-4-0613 (Table 4). This suggests that our fine-tuning procedure is effective even with a less capable base model.

No fine-tuning. To demonstrate the benefit of fine-tuning (Section 5.1), we evaluate our optimal system using only the base GPT-4-Preview-1106 as the reasoning model. In this setup, the ablated system achieves a Brier score of 0.186, an increase of 0.007 compared to the original score. This indicates that fine-tuning provides a significant boost to our system's performance.

No fine-tuning and no retrieval. We evaluate our optimal system without any news retrieval and using the base GPT-4-1106-Preview model. The ablated system attains a Brier score of 0.206.

Recall that in our baseline evaluation (Section 3.3), the lowest Brier score attained by any model is 0.208. Our ablated system essentially deteriorates to this baseline level. Without any fine-tuning or retrieval, the only expected advantage of our system over the baseline evaluation setup is its reasoning prompt, which was found through searching a set of candidate prompts (Section 5). The experiment suggests that the reasoning prompt alone provides only a minor improvement in performance.

8 Conclusion

Our work presents the first ML system that can forecast at near human levels. We develop a novel retrieval mechanism that uses a LM to determine which information to source and how to evaluate its relevance. We also give a self-supervised fine-tuning method to generate reasonings with accurate predictions. In addition, our system can potentially aid human forecasters by providing effective news retrieval and novel perspectives in reasoning drawn from LM pre-training knowledge.

To facilitate further research, we release our dataset: the largest and most recent forecasting dataset compiled from 5 real-world forecasting competitions. We discuss future directions (e.g., domain-adaptive training and data augmentation) to improve our system in Appendix H.

9

At a high level, our results suggest that in the near future, LM-based systems may be able to generate accurate forecasts at the level of competitive human forecasters. We hope that our work paves the way for automated, scalable forecasting that can help to inform institutional decision making.

References

Abolghasemi, M., Ganbold, O., and Rotaru, K. (2023). Humans vs large language models: Judgmental forecasting in an era of advanced AI. *arXiv preprint arXiv:2312.06941*.

Adam, D. (2020). Special report: The simulations driving the world's response to COVID-19. *Nature*, 580(7802):316–319.

Agarwal, R., Vieillard, N., Zhou, Y., Stanczyk, P., Garea, S. R., Geist, M., and Bachem, O. (2024). On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations (ICLR)*.

Anthropic (2023). Model card and evaluations for Claude models. https://www-cdn.anthropic.com/files/4zrzovbb/website/5c49cc247484cecf107c699baf29250302e5da70.pdf.

Armstrong, J. S. (2001). *Principles of Forecasting: a Handbook for Researchers and Practitioners*. Springer.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language models are few-shot learners. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Chen, X., Aksitov, R., Alon, U., Ren, J., Xiao, K., Yin, P., Prakash, S., Sutton, C., Wang, X., and Zhou, D. (2023). Universal self-consistency for large language model generation. *arXiv preprint arXiv:2311.17311*.

Das, A., Kong, W., Sen, R., and Zhou, Y. (2024). A decoder-only foundation model for time-series forecasting. *arXiv preprint arXiv:2310.10688*.

Deng, Y., Zhang, W., Chen, Z., and Gu, Q. (2023). Rephrase and respond: Let large language models ask better questions for themselves. *arXiv preprint arXiv:2311.04205*.

Dooley, S., Khurana, G. S., Mohapatra, C., Naidu, S. V., and White, C. (2023). ForecastPFN: Synthetically-trained zero-shot forecasting. In *Advanced in Neural Information Processing Systems (NeurIPS)*.

Gawlikowski, J., Tassi, C. R. N., Ali, M., Lee, J., Humt, M., Feng, J., Kruspe, A., Triebel, R., Jung, P., Roscher, R., Shahzad, M., Yang, W., Bamler, R., and Zhu, X. X. (2021). A survey of uncertainty in deep neural networks. *arXiv preprint arXiv:2107.03342*.

Gemini Team (2023). Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378.

Gruver, N., Finzi, M. A., Qiu, S., and Wilson, A. G. (2023). Large language models are zero-shot time series forecasters. In *Advanced in Neural Information Processing Systems (NeurIPS)*.

Gu, Y., Dong, L., Wei, F., and Huang, M. (2024). MiniLLM: Knowledge distillation of large language models. In *International Conference on Learning Representations (ICLR)*.

Gulcehre, C., Paine, T. L., Srinivasan, S., Konyushkova, K., Weerts, L., Sharma, A., Siddhant, A., Ahern, A., Wang, M., Gu, C., et al. (2023). Reinforced self-training (ReST) for language modeling. *arXiv preprint arXiv:2308.08998*.

10

Hanson, R. (2007). Logarithmic markets coring rules for modular combinatorial information aggregation. *The Journal of Prediction Markets*, 1(1):3–15.

Hendrycks, D., Carlini, N., Schulman, J., and Steinhardt, J. (2021). Unsolved problems in ML safety. *arXiv preprint arXiv:2109.13916*.

Izacard, G. and Grave, É. (2021). Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*.

Jiang, A. Q., Sablayrolles, A., Roux, A., Mensch, A., Savary, B., Bamford, C., Chaplot, D. S., de las Casas, D., Hanna, E. B., Bressand, F., Lengyel, G., Bour, G., Lample, G., Lavaud, L. R., Saulnier, L., Lachaux, M.-A., Stock, P., Subramanian, S., Yang, S., Antoniak, S., Scao, T. L., Gervet, T., Lavril, T., Wang, T., Lacroix, T., and Sayed, W. E. (2024). Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Jin, M., Wang, S., Ma, L., Chu, Z., Zhang, J. Y., Shi, X., Chen, P.-Y., Liang, Y., Li, Y.-F., Pan, S., and Wen, Q. (2024). Time-LLM: Time series forecasting by reprogramming large language models. In *International Conference on Learning Representations (ICLR)*.

Jin, W., Khanna, R., Kim, S., Lee, D.-H., Morstatter, F., Galstyan, A., and Ren, X. (2021). ForecastQA: A question answering challenge for event forecasting with temporal text data. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL)*.

Lewis, P. S. H., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., and Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Manifold (2022). Maniswap. https://manifoldmarkets.notion.site/manifoldmarkets/Maniswap-ce406e1e897d417cbd491071ea8a0c39.

Metaculus (2023). Wisdom of the crowd vs. the best of the best of the best. https://www.metaculus.com/notebooks/15760/wisdom-of-the-crowd-vs-the-best-of-the-best-of-the-best.

Min, S., Zhong, V., Zettlemoyer, L., and Hajishirzi, H. (2019). Multi-hop reading comprehension through question decomposition and rescoring. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Nakano, R., Hilton, J., Balaji, S., Wu, J., Ouyang, L., Kim, C., Hesse, C., Jain, S., Kosaraju, V., Saunders, W., Jiang, X., Cobbe, K., Eloundou, T., Krueger, G., Button, K., Knight, M., Chess, B., and Schulman, J. (2021). WebGPT: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*.

Nie, Y., Nguyen, N. H., Sinthong, P., and Kalagnanam, J. (2023). A time series is worth 64 words: Long-term forecasting with transformers. In *International Conference on Learning Representations (ICLR)*.

Nye, M., Andreassen, A. J., Gur-Ari, G., Michalewski, H., Austin, J., Bieber, D., Dohan, D., Lewkowycz, A., Bosma, M., Luan, D., Sutton, C., and Odena, A. (2021). Show your work: Scratchpads for intermediate computation with language models. *arXiv preprint arXiv:2112.00114*.

OpenAI (2023). GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

Polymarket (2023). Polymarket/poly-market-maker: Market Maker Keeper for the polymarket CLOB. https://github.com/Polymarket/poly-market-maker.

Rasul, K., Ashok, A., Williams, A. R., Khorasani, A., Adamopoulos, G., Bhagwatkar, R., Biloš, M., Ghonia, H., Hassen, N. V., Schneider, A., Garg, S., Drouin, A., Chapados, N., Nevmyvaka, Y., and Rish, I. (2023). Lag-Llama: Towards foundation models for time series forecasting. *arXiv preprint arXiv:2310.08278*.

11

Schoenegger, P. and Park, P. S. (2023). Large language model prediction capabilities: Evidence from a real-world forecasting tournament. *arXiv preprint arXiv:2310.13014*.

Schulman, J. (2023). Proxy objectives in reinforcement learning from human feedback. `https://icml.cc/virtual/2023/invited-talk/21549`. Invited talk at International Conference on Machine Learning (ICML).

Shuster, K., Poff, S., Chen, M., Kiela, D., and Weston, J. (2021). Retrieval augmentation reduces hallucination in conversation. In *Findings of the Association for Computational Linguistics (Findings of EMNLP)*.

Singh, A., Co-Reyes, J. D., Agarwal, R., Anand, A., Patil, P., Garcia, X., Liu, P. J., Harrison, J., Lee, J., Xu, K., Parisi, A. T., Kumar, A., Alemi, A. A., Rizkowsky, A., Nova, A., Adlam, B., Bohnet, B., Elsayed, G. F., Sedghi, H., Mordatch, I., Simpson, I., Gur, I., Snoek, J., Pennington, J., Hron, J., Kenealy, K., Swersky, K., Mahajan, K., Culp, L. A., Xiao, L., Bileschi, M., Constant, N., Novak, R., Liu, R., Warkentin, T., Bansal, Y., Dyer, E., Neyshabur, B., Sohl-Dickstein, J., and Fiedel, N. (2024). Beyond human data: Scaling self-training for problem-solving with language models. *Transactions on Machine Learning Research (TMLR)*.

Tetlock, P. E. and Gardner, D. (2015). *Superforecasting: The Art and Science of Prediction*. Crown.

Tian, K., Mitchell, E., Zhou, A., Sharma, A., Rafailov, R., Yao, H., Finn, C., and Manning, C. D. (2023). Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., Fuller, B., Gao, C., Goswami, V., Goyal, N., Hartshorn, A., Hosseini, S., Hou, R., Inan, H., Kardas, M., Kerkez, V., Khabsa, M., Kloumann, I., Korenev, A., Koura, P. S., Lachaux, M.-A., Lavril, T., Lee, J., Liskovich, D., Lu, Y., Mao, Y., Martinet, X., Mihaylov, T., Mishra, P., Molybog, I., Nie, Y., Poulton, A., Reizenstein, J., Rungta, R., Saladi, K., Schelten, A., Silva, R., Smith, E. M., Subramanian, R., Tan, X. E., Tang, B., Taylor, R., Williams, A., Kuan, J. X., Xu, P., Yan, Z., Zarov, I., Zhang, Y., Fan, A., Kambadur, M., Narang, S., Rodriguez, A., Stojnic, R., Edunov, S., and Scialom, T. (2023). Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Wang, C. (2023). Calibration in deep learning: A survey of the state-of-the-art. *arXiv preprint arXiv:2308.01222*.

Woo, G., Liu, C., Kumar, A., Xiong, C., Savarese, S., and Sahoo, D. (2024). Unified training of universal time series forecasting transformers. *arXiv preprint arXiv:2402.02592*.

Yan, Q., Seraj, R., He, J., Meng, L., and Sylvain, T. (2024). Autocast++: Enhancing world event prediction with zero-shot ranking-based context retrieval. In *International Conference on Learning Representations (ICLR)*.

Zelikman, E., Wu, Y., Mu, J., and Goodman, N. (2022). STaR: Bootstrapping reasoning with reasoning. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Zhang, Y., Chen, X., and Park, D. (2018). Formal specification of constant product (xy= k) market maker model and implementation. *White paper*.

Zhu, Y., Yuan, H., Wang, S., Liu, J., Liu, W., Deng, C., Dou, Z., and Wen, J.-R. (2024). Large language models for information retrieval: A survey. *arXiv preprint arXiv:2308.07107*.

Zou, A., Xiao, T., Jia, R., Kwon, J., Mazeika, M., Li, R., Song, D., Steinhardt, J., Evans, O., and Hendrycks, D. (2022). Forecasting future world events with neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*.

12

A Details about Retrieval Schedule, Models and Knowledge Accuracy

A.1 Retrieval Schedule

To create a set of retrieval dates for each question, we use geometrically increasing time points between the open and close dates. We choose this schedule for two reasons: (1) questions are often most active shortly after they open, and (2) some questions have overly conservative close dates that are long after the question resolves. We use $n = 5$ retrieval dates per question; the kth retrieval date is given by

$$\text{retrieval\_date}_k = \text{date}_{\text{begin}} + (\text{date}_{\text{close}} - \text{date}_{\text{begin}} - 1)^{k/n}. \qquad (2)$$

For questions that resolve before they close, we exclude all retrieval dates occurring after the question has been resolved. Under this geometric retrieval schedule, we retain 86% of retrieval dates on average across all questions. The average question window in our corpus is approximately 70 days, and the average time until resolution is 42 days.

In our dataset, questions can get resolved long before their official close date. This occurs for questions like "Will $\langle event \rangle$ happen by $\langle date \rangle$", where resolving early indicates that the event did occur (see Table 1 for an example). It is tempting to choose retrieval dates with respect to the resolve date so that each question can receive the same number of retrieval dates, e.g. by retrieving at evenly spaced intervals between the open and resolve date. However, this would leak information, since the retrieval date would now depend on the resolve date, which correlates with the resolution.

A.2 Models

We evaluate 14 instruction-tuned LMs: GPT-3.5-Turbo-1106, GPT-3.5-Turbo-Instruct (Brown et al., 2020); GPT-4-0613, GPT-4-1106-Preview (OpenAI, 2023); Llama-2-7B, Llama-2-13B, Llama-2-70B (Touvron et al., 2023); Mistral-7B-Instruct, Mistral-8x7B-Instruct (Jiang et al., 2024), Nous Hermes 2 Mixtral-8x7B-DPO, Yi-34B-Chat, Claude-2, Claude-2.1 (Anthropic, 2023), and Gemini-Pro (Gemini Team, 2023).

We give a list of detailed information of the models we use below. The weights of the open models are available publicly on Hugging Face, and we primarily use Together AI's serving API to access them. All cut-offs are based on official statements.

Model	Source	Open Weights	Knowledge Cutoff	Evaluation Cost
GPT-4-0613	OpenAI	No	Sep 2021	$0.03/1K tokens
GPT-4-1106-Preview	OpenAI	No	Apr 2023	$0.01/1K tokens
GPT-3.5-Turbo-1106	OpenAI	No	Sep 2021	$0.001/1K tokens
GPT-3.5-Turbo-Instruct	OpenAI	No	Sep 2021	$0.0015/1K tokens
Claude-2	Anthropic	No	Dec 2022	$0.024/1K tokens
Claude-2.1	Anthropic	No	Dec 2022	$0.024/1K tokens
Gemini-Pro	Google	No	Early 2023	$0.0005/1K characters
Mistral-7B-Instruct	Mistral AI	Yes	*Unknown*	$0.0002/1K tokens
Mistral-8x7B-Instruct	Mistral AI	Yes	*Unknown*	$0.0002/1K tokens
Mixtral-8x7B-DPO	NousResearch	Yes	*Unknown*	$0.0002/1K tokens
YI-34B-Chat	01.AI	Yes	June 2023	$0.000776/1K tokens
Llama-2-7B-Chat	Meta	Yes	Sep 2022	$0.0002/1K tokens
Llama-2-13B-Chat	Meta	Yes	Sep 2022	$0.00025/1K tokens
Llama-2-70B-Chat	Meta	Yes	Sep 2022	$0.0009/1K tokens

Table 5: **Language models (LMs) evaluated in our study**, including their sources, availability of weights, knowledge cutoff dates, and evaluation costs. The evaluation costs for open-weight models are based on Together AI's pricing. Gemini-Pro's knowledge cutoff is claimed to be early 2023. The exact knowledge cutoffs for the Mistral series are not publicly reported.

A.3 Testing Potential Leakage from Post-training

The three models we use in our system are GPT-3.5-Turbo-1106, fine-tuned GPT-4-0613, and GPT-4-1106-Preview. The first two models have a knowledge cutoff of September 2021, while GPT-4-1106-Preview has a knowledge cutoff of December 2023. We find no evidence that the

13

post-training phase leaks further information after their knowledge cutoffs. As a test, we manually query the model on 20 major events[2] between June 2023 and September 2023, such as "Who won the 2023 Turkish presidential election?". For all 20 questions, both models either claim no knowledge or simply hallucinate.

As a sanity check, we also prompt GPT-4-1106-Preview to answer another 20 questions about events during November, 2022–January, 2023, prior to its knowledge cutoff, such as "Which team won the 2022 FIFA World Cup Final?". The model answers all of them correctly.

A.4 Crowd Predictions

On any given question, each platform computes a community prediction that aggregates all individual forecasts. The prediction is dynamically updated and recorded as the forecasts are made. We source the records directly from the platforms (instead of computing them from scratch using the individual forecasts). For binary questions, we provide more details on the aggregation mechanisms as follows.

- On Metaculus, for a given question, each prediction of a forecaster is marked by t (starting at 1), from their earliest prediction to the latest. The platform computes the crowd prediction of the question by weighted median. The weight of the tth forecast from an individual forecaster is $e^{\sqrt{t}}$, so the more recent forecasts receive higher weights. We remark that the platform also publishes another aggregated forecast called "Metaculus prediction," which we do not use or compare with in this paper. This differs from the crowd prediction described above and is computed via a proprietary algorithm.
- GJOpen computes the crowd predictions by the mean of the most recent 40% of forecasts from each forecaster.
- INFER initializes the crowd prediction to be the mean of all individual forecasts. As the question progresses, it reweights the forecasts, for example, by "putting more weight on the forecasts of individuals with the best track record."[3] Exact details on the aggregation mechanisms are not found on their website.
- Manifold and Polymarket are prediction markets, where the community predictions are the prices (between 0 and 1). The prices are adjusted by their automated market makers, as bets are made. The mechanisms are variants of constant-product market makers (Hanson, 2007; Zhang et al., 2018). See Polymarket (2023) and Manifold (2022) for more details.

---

[2]sourced from https://www.onthisday.com/events/date/2023/.
[3]https://www.infer-pub.com/frequently-asked-questions

14

B Details about Base Evaluations

In this section, we provide experimental details on our baseline evaluations (Section 3.3).

B.1 Evaluation Method

For both zero-shot and scratchpad prompting, we conduct basic prompt optimization by crafting 5 candidate zero-shot prompts and 4 candidate scratchpad prompts. We evaluate each prompt on the validation set by comparing Brier scores. Specifically, we randomly select 200 questions from the validation set and calculate the mean Brier scores across the 14 LMs under consideration.

- The best zero-shot prompt achieves an average Brier score of 0.246, outperforming the others, which score 0.261, 0.276, 0.279, and 0.252, respectively.
- For scratchpad, all prompts yield similar Brier scores. We observe that potentially due to safety training, models can sometimes refuse to answer forecasting questions by simply claiming "I don't know". Therefore, we use the number of "refuse to answer" responses as the deciding metric. The winning scratchpad prompt averages 88 "refuse to answer" responses, while the others average 106, 93, and 94, respectively.

The best zero-shot and scratchpad prompts are shown in Figure 4 and Figure 5. In both prompting styles, models are only provided with the question, background, resolution criterion, and question's open and close dates (`date_begin` and `date_end`). All the data are sourced from the forecasting platforms and publicly available on the question page to human forecasters. We do no additional news retrieval.

Finally, we use the best prompt of each prompting strategy to forecast on each question in the test set. In Section 3.3, we find that none of the models are naturally good at forecasting. We provide the full results next in Section B.2.

You are an expert superforecaster, familiar with the work of Tetlock and others. Make a prediction of the probability that the question will be resolved as true. You MUST give a probability estimate between 0 and 1 UNDER ALL CIRCUMSTANCES. If for some reason you can't answer, pick the base rate, but return a number between 0 and 1.

Question: {question}
Question Background: {background}
Resolution Criteria: {resolution_criteria}

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

Output your answer (a number between 0 and 1) with an asterisk at the beginning and end of the decimal. Do not output anything else.
Answer: {{ Insert answer here }}

Figure 4: **The simple zero-shot prompt used for baseline evaluations**. No retrieval is performed. The prompt simply asks the model to make a prediction on a given question from the test set. We add the directive "You MUST ... UNDER ALL CIRCUMSTANCES" to push the model to answer the question, which in some cases it refuses to, potentially due to safety training. See Section 3.3 for results and Appendix B for more details.

15

```
    Question: {question}
    Question Background:{background}
    Resolution Criteria:{resolution_criteria}

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

    Instructions:
    1. Provide reasons why the answer might be no.
    {{ Insert your thoughts }}

    2. Provide reasons why the answer might be yes.
    {{ Insert your thoughts }}

    3. Aggregate your considerations.
    {{ Insert your aggregated considerations }}

    4. Output your answer (a number between 0 and 1) with an asterisk at the beginning and end of
    the decimal.
    {{ Insert your answer }}
```

Figure 5: **The scratchpad prompt used for baseline evaluations.** No retrieval is performed. The prompt asks the model to make a prediction on a given question from the test set, after making considerations for yes and no. See Section 3.3 for results and Appendix B for more details.

B.2   Baseline Evaluation Results

We now give the full results of our baseline evaluation (Section 3.3) in Table 6.

Model	Zero-shot	Scratchpad
GPT-4-0613	$.219_{.013}$	$.222_{.009}$
GPT-4-1106-Preview	$\mathbf{.208}_{.013}$	$\mathbf{.209}_{.012}$
GPT-3.5-Turbo-1106	$.274_{.016}$	$.261_{.010}$
GPT-3.5-Turbo-Instruct	$.237_{.014}$	$.257_{.009}$
Claude-2	$.220_{.013}$	$.219_{.014}$
Claude-2.1	$.220_{.013}$	$.215_{.014}$
Gemini-Pro	$.243_{.019}$	$.230_{.007}$
Mistral-7B-Instruct	$.237_{.018}$	$.243_{.008}$
Mistral-8x7B-Instruct	$.238_{.018}$	$.238_{.010}$
Mixtral-8x7B-DPO	$.260_{.022}$	$.248_{.010}$
Yi-34B-Chat	$.238_{.012}$	$.241_{.009}$
Llama-2-7B	$.353_{.020}$	$.264_{.011}$
Llama-2-13B	$.226_{.009}$	$.268_{.008}$
Llama-2-70B	$.283_{.014}$	$.282_{.011}$

Table 6: **Zero-shot and scratchpad Brier scores** on the test set. All models fall significantly far from human aggregate. Subscript numbers denote 2 standard errors. Random baseline: 0.25; human crowd: 0.149.

B.3   Knowledge Evaluation by Category

We present an evaluation of model's knowledge about resolved questions on past events and notice variations in performance across categories. To investigate further, we analyzed each model's zero-shot Brier score on the test set by category. This analysis showed a correlation between models' knowledge on the training and validation sets and their Brier scores on the test set across categories. This suggests that domain-adaptive training could be used to improve model performance in categories where its existing knowledge is limited.

16

First, we assessed pre-trained language model knowledge across categories by evaluating their ability to give the correct resolutions to resolved questions from the train and validation sets (Table 7).

We noticed variations in knowledge accuracy across categories. To dig deeper, we analyze the zero-shot Brier score on the test set in Table 8 and assess if there is a correlation between knowledge accuracy on the training and validation sets and zero-shot Brier score on the test set in Table 9.

The potential for domain-adaptive training. We calculate the correlation between the models' knowledge accuracy and their Brier scores of the zero-shot evaluation. Notably, in the Politics & Governance, Arts & Recreation, and Education & Research categories, there exists a strong negative correlation (Table 9). This negative correlation is expected because a higher knowledge accuracy should intuitively correspond to a lower Brier score. As a direction for future research, we propose that domain-adaptive training could be employed to enhance forecasting performance in specific categories.

Model	Arts & Recreation	Economics & Business	Education Research	Environment Energy	Healthcare & Biology	Politics & Governance	Science & Technology	Security & Defense	Social Sciences	Sports	Other
GPT-4-0613	.083	.114	.125	.220	.157	.349	.196	.279	.077	.040	.071
GPT-4-1106-Preview	.094	.142	.125	.153	.144	.391	.227	.207	0	.234	0
GPT-3.5-Turbo-1106	.196	.195	.375	.229	.262	.278	.247	.286	.154	.166	.250
GPT-3.5-Turbo-Instruct	.411	.323	.250	.314	.419	.328	.387	.314	.462	.365	.107
Claude-2	.140	.205	.219	.254	.245	.446	.296	.134	.154	.392	.071
Claude-2.1	.136	.205	.219	.246	.249	.446	.294	.136	.077	.395	.071
Gemini-Pro	.155	.425	.188	.415	.314	.425	.356	.545	.077	.350	.250
Mistral-7B-Instruct	.011	.024	0	.034	.022	.050	.054	.007	0	.018	0
Mistral-8x7B-Instruct	.019	.045	.094	.051	.066	.071	.049	.040	0	.027	0
Mixtral-8x7B-DPO	.004	.084	.031	.051	.087	.041	.054	.014	0	.010	0
YI-34B-Chat	.423	.552	.625	.593	.555	.630	.588	.738	.538	.624	.536
Llama-2-7B	.042	.069	.156	.203	.284	.046	.067	.033	0	.050	.071
Llama-2-13B	.102	.181	.156	.288	.288	.210	.247	.163	.231	.189	.036
Llama-2-70B	.143	.175	.344	.322	.266	.243	.384	.115	.077	.075	.107

Table 7: **Comparison of knowledge accuracy across categories and models** on the train and validation sets. We list the knowledge accuracy of all base models with respect to all categories in the train and validation set.

B.4 Other Baselines

We evaluated three other baselines to compare our model's performance: overall base rate, platform-level base rate, and majority class baselines.

A model predicting the overall base rate (.37) achieved a Brier score of .234, while a platform-level base rate model achieved a Brier score of .22. In comparison, our model significantly outperformed both, with a Brier score of .179.

Furthermore, we evaluated our model against the majority class baseline. Our dataset is divided into training, validation, and test sets with class distributions of 33%/67%, 36%/64%, and 38%/62%, respectively, for resolving questions as yes (1) or false (0). Predicting all questions as false (0) yielded a Brier score of .37.

17

Model	Arts & Recreation	Economics & Business	Education Research	Environment Energy	Healthcare & Biology	Politics & Governance	Science & Technology	Security & Defense	Sports
GPT-4-0613	.278	.260	.437	.201	.203	.228	.200	.224	.178
GPT-4-1106-Preview	.240	.244	.394	.222	.122	.218	.178	.207	.177
GPT-3.5-Turbo-1106	.309	.294	.336	.239	.336	.343	.225	.250	.214
GPT-3.5-Turbo-Instruct	.292	.281	.270	.245	.388	.244	.178	.235	.205
Claude-2	.293	.239	.326	.199	.226	.214	.175	.244	.194
Claude-2.1	.293	.242	.316	.199	.226	.213	.183	.244	.194
Gemini-Pro	.301	.303	.432	.227	.210	.263	.175	.255	.189
Mistral-7B-Instruct	.291	.265	.295	.228	.238	.271	.184	.236	.191
Mistral-8x7B-Instruct	.354	.272	.452	.256	.335	.252	.176	.227	.189
Mixtral-8x7B-DPO	.367	.315	.543	.213	.217	.287	.184	.265	.194
YI-34B-Chat	.263	.240	.332	.196	.208	.265	.196	.236	.212
Llama-2-7B	.381	.356	.331	.359	.351	.399	.351	.288	.327
Llama-2-13B	.260	.247	.263	.218	.230	.245	.197	.222	.199
Llama-2-70B	.318	.329	.319	.299	.498	.329	.308	.264	.212

Table 8: **Comparison of zero-shot Brier scores across categories and models** on the test set. This table lists the Brier scores of all base models with respect to the specified categories.

Category	Score
Arts & Recreation	**-0.417103**
Economics & Business	-0.228040
Education & Research	**-0.359102**
Environment & Energy	-0.135552
Healthcare & Biology	0.162110
Politics & Governance	**-0.487266**
Science & Tech	-0.091878
Security & Defense	-0.183253
Sports	-0.136017

Table 9: **Correlation between knowledge accuracy and zero-shot prompt Brier score by category**. Categories with an absolute correlation of 0.3 or greater, shown in bold, indicate a high correlation between accuracy on the training and validation set and forecasting performance on the test set. This highlights that in certain domains model's forecasting capabilities are correlated with its pre-training knowledge.

18

C   Dataset: Curation and Further Analysis

C.1   Data Collection and Curation

Scraping. To compile our dataset from the forecasting platforms, we query their APIs or scrape the questions' webpages for initial data gathering. For Metaculus, we first extract basic information via the API and scrape the resolution criteria from webpage. INFER (CSET) and Good Judgment Open data are gathered via web scraping, since no API provides the full data we need. Polymarket's data, except for community predictions, is obtained from their API. Manifold's data is fully scraped via API.

Assigning categories. There is no standard, uniform categorization of the forecast questions across the platforms. We prompt GPT-3.5-Turbo to assign one of the 11 categories to each question. See Figure 6 for the category set and the prompt we use.

Question: {question}
Background: {background}
Options:

- Science & Tech
- Healthcare & Biology
- Economics & Business
- Environment & Energy
- Politics & Governance
- Education & Research
- Arts & Recreation
- Security & Defense
- Social Sciences
- Sports
- Other

Instruction: Assign a category for the given question.
Rules:

1. Make sure you only return one of the options from the option list.
2. Only output the category, and do not output any other words in your response.
3. You have to pick a string from the above categories.

Answer: {{ Insert answer here }}

Figure 6: **Prompt for categorizing questions based on the provided options**. The prompt presents the forecasting question, along with 11 candidate category choices, and prompts the model to classify the question into one of the categories.

Screening and curation. From manual examination, we notice that the initial dataset contains questions that are ambiguously formulated or overly personal. In a preliminary screening phase, we prompt GPT-3.5 to identify and exclude these questions. See Figure 7 for a prompt to detect ill-defined questions, where we provide several few-shot examples.

We then manually examine to eliminate all questions of low quality. This includes those with few community forecasts or trading engagement on platforms such as Manifold and Polymarket, as well as any ill-defined questions that GPT-3.5 is unable to identify during the initial screening.

C.2   Further Statistics and Samples

We give a list of detailed statistics and plots on our data:

- Figure 8 visualizes the location mentions in all the questions from our full dataset.
- Table 11 gives the distribution of questions and forecasts across platforms in our full dataset.
- Table 10 showcases a complete data sample in our curated set.

19

I'm trying to assess the quality of an old forecasting dataset.
Here is a forecasting question from the dataset: {question}.
Please flag questions that don't sound like binary forecasting questions by outputting "flag". If it sounds like a reasonable question, output "ok".

Examples of strings that should be flagged:

- Will I finish my homework tonight?
- Metaculus party 2023
- Will Hell freeze over?
- Heads or tails
- Will this video reach 100k views by the EOD?

Examples of strings 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?

If a question is already resolved, that doesn't mean it should be flagged. When in doubt, mark it as "ok".
Your response should take the following structure:
Insert thinking: {{ insert your concise thoughts here }}
Classification: {{ insert "flag" or "ok" }}

Figure 7: **The prompt for flagging ill-defined forecasting questions** in our dataset. The prompt contains several few-shot examples where the questions are ill-defined. A LM is prompted to filter out any questions of similar nature.

- Table 12 shows a list of questions with how community predictions shift over time.
- Figure 9 shows the opening dates of the questions in the full dataset.

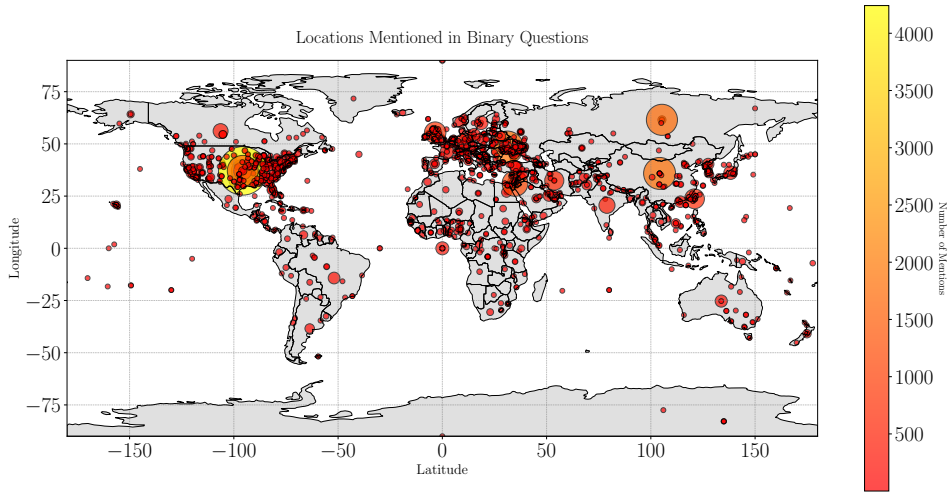



Figure 8: **Location mentions in all binary questions in our full dataset**. We visualize all location mentions in our full dataset on a world map. This shows that the dataset provides a diverse coverage of topics across the globe.

20

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, 0.701), (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, 0.715), (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 10: **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.

Platform	Questions (All)	Predictions (All)	Questions (Binary)	Predictions (Binary)	Brier Score (Binary)
Metaculus	8,881	638,590	4,862	387,488	.130
INFER	308	73,778	192	47,918	.079
GJOpen	2,592	743,671	1,168	342,216	.134
Manifold	24,284	1,997,928	20,319	1,387,668	.155
Polymarket	12,689	3,720,640	7,123	1,879,035	.158

Table 11: **Raw dataset statistics across platforms**. The Brier scores are calculated by averaging over all time points where the platforms provide crow aggregates.

21

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 12: **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.

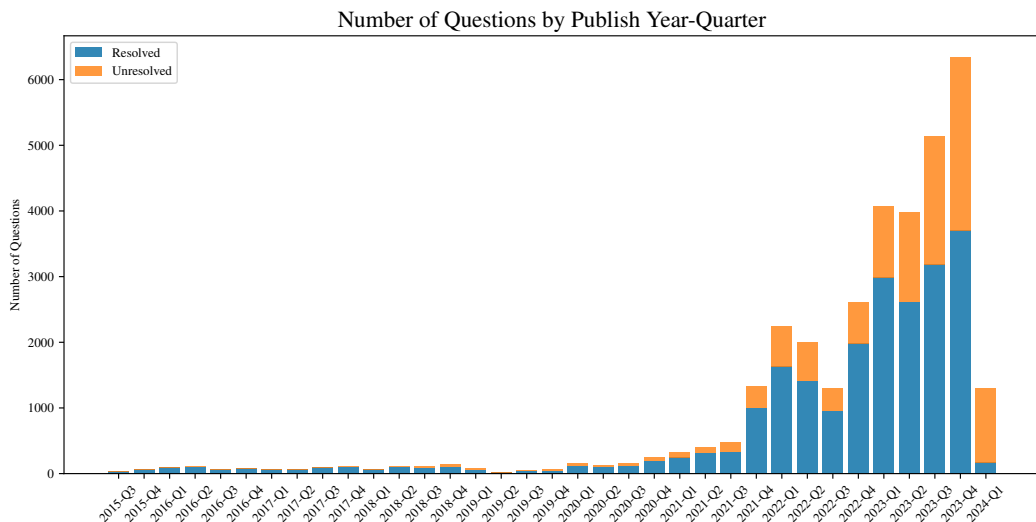



Figure 9: **Distribution of the opening dates of the questions in our full datasets**, ordered by year-quarter. Activity on these platforms has sharply increased over the past two years.

22

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 13.

Our retrieval system can retrieve a large number of texts (e.g. > 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 F.4, 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 14) 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 14 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 11. The other prompts also conform to the basic template as shown in Figure 10, 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 10 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 10: **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.

23

Question: {question}

Question Background: {background}

Resolution Criteria: {resolution_criteria}

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

We have retrieved the following information for this question:
{retrieved_info}

Instructions:
1. Given the above question, rephrase and expand it to help you do better answering. Maintain all information in the original question.
{{ Insert rephrased and expanded question. }}

2. Using your knowledge of the world and topic, as well as the information provided, provide a few reasons why the answer might be no. Rate the strength of each reason.
{{ Insert your thoughts }}

3. Using your knowledge of the world and topic, as well as the information provided, provide a few reasons why the answer might be yes. Rate the strength of each reason.
{{ Insert your thoughts }}

4. Aggregate your considerations. Think like a superforecaster (e.g. Nate Silver).
{{ Insert your aggregated considerations }}

5. Output an initial probability (prediction) given steps 1-4.
{{ Insert initial probability }}

6. Evaluate whether your calculated probability is excessively confident or not confident enough. Also, consider anything else that might affect the forecast that you did not before consider (e.g. base rate of the event).
{{ Insert your thoughts }}

7. Output your final prediction (a number between 0 and 1) with an asterisk at the beginning and end of the decimal.
{{ Insert your answer }}

Figure 11: **The scratchpad reasoning prompt that gets lowest Brier score on validation set**. The prompt first provides the basic information about the question, along with retrieved article summaries. Then it gives instructions to guide the model's reasoning path (Section 4).

24

I will provide you with a forecasting question and the background information for the question.

Question: {question}
Background: {background}

Task:
- Generate brief search queries (up to {max_words} words each) to gather information on Google that could influence the forecast.

You must generate this exact amount of queries: {num_keywords}

Your response should take the following structure: Thoughts: {{ Insert your thinking here. }}
Search Queries: {{ Insert the queries here. Use semicolons to separate the queries. }}

(a) A straighforward search query expansion prompt

I will provide you with a forecasting question and the background information for the question. I will then ask you to generate short search queries (up to {max_words} words each) that I'll use to find articles on Google News to help answer the question.

Question: {question}
Background: {background}

You must generate this exact amount of queries: {num_keywords}

Start off by writing down sub-questions. Then use your sub-questions to help steer the search queries you produce.

Your response should take the following structure: Thoughts: {{ Insert your thinking here. }}
Search Queries: {{ Insert the queries here. Use semicolons to separate the queries. }}

(b) The second search query prompt we use. It first asks the model to consider sub-questions and use that to steer the outputs.

Figure 12: **Prompts to generate search queries** based on the question's data.

25

Please consider the following forecasting question and its background information. After that, I will give you a news article and ask you to rate its relevance with respect to the forecasting question.

Question: {question}
Question Background:{background}
Resolution Criteria:{resolution_criteria}

Article: {article}

Please rate the relevance of the article to the question, at the scale of 1-6
1 – irrelevant
2 – slightly relevant
3 – somewhat relevant
4 – relevant
5 – highly relevant
6 – most relevant

Guidelines:
- You don't need to access any external sources. Just consider the information provided.
- Focus on the content of the article, not the title.
- If the text content is an error message about JavaScript, paywall, cookies or other technical issues, output a score of 1.

Your response should look like the following:
Thought: {{ Insert your thinking }}
Rating: {{ Insert answer here }}

Figure 13: **Prompt used to rate the relevance of an article with respect to a question**. The prompt asks a LM to rate the relevance of an article with respect to a question at the scale of 1–6. We extract the numerical value following "Rating:".

I want to make the following article shorter (condense it to no more than 100 words).

Article: {article}
When doing this task for me, please do not remove any details that would be helpful for making considerations about the following forecasting question.

Forecasting Question: {question}
Question Background: {background}

Figure 14: **The summarization prompt we use in our retrieval system**. The prompt provides a question, its background, and a relevant article. It asks the LM to condense the article without removing any information relevant to the forecasting question.

26

E Details on Fine-tuning

To generate the preliminary data, we run our system at each retrieval date in the retrieval schedule and on each question in the train set. We generate these 16 reasonings on different prompts and retrieval configurations. Details follow.

First, as a form of data augmentation, we retrieve 2 sets of articles for each question by sampling 2 distinct retrieval configurations (Figure 2, left). Specifically, we sample twice the retrieval prompt, number of queries, and articles per query, with relevancy filtering and summarization following the process described in Section 4.1. This results in 2 inputs to the reasoning model per question, each with the same question but a different set of articles.

To increase the chance of attaining a prediction that outperforms the crowd, we generate 4 candidate outputs per input (8 total per question) by trying different scratchpad prompts. The first uses the optimal prompt (Figure 11) found in Section 5.2. We then sample 3 other scratchpad prompts, with probability inversely proportional to their Brier score on the validation set. We prompt both Claude-2.1 and GPT-4-Preview, since we find that Claude-2.1 is better on some questions. In total, this gives 16 candidate forecasts per question.

F Details on Hyperparameter Optimization

F.1 Hyperparameter Sweep Procedure

Methodology We divide the hyperparameters into groups of 1-2 and optimize them iteratively. For each group, we select the best configuration based on the average Brier score on the validation set, except for search query generation where we use proxy metrics for efficiency.

We optimize the groups sequentially, fixing the optimal configurations from previous groups while sweeping the current one. The hyperparameters yet to be swept are randomized independently for each input question.

Retrieval. Our retrieval uses LMs for search query generation, relevance rating, and summarization. We independently optimize the prompt choices for search query generation and summarization. The relevance rating prompt is fixed in our system (Figure 13).

For search query generation, we evaluate the prompts by retrieving articles with the generated queries and examining two metrics: (1) the average relevance score across all retrieved articles, and (2) the average relevance score of articles exceeding a relevance threshold of 4 on a 6-point scale. The 2 high-scoring prompts perform similarly under both metrics and generate queries with little overlap. As a result, we use both prompts (Figure 12) to generate queries and take the union.

For summarization, we run our system end-to-end and pick the top 1 prompt (Figure 14) with respect to the Brier score.

Reasoning. The reasoning system takes a ranked list of article summaries and prompts LMs to make forecasts. We optimize: (1) the ordering criterion of the summaries (by relevance or recency); (2) the number k of article summaries presented to LMs; and (3) the choice of scratchpad instructions to elicit the forecasts.

For efficiency, we optimize them in 2 independent stages:

- In the first stage, we jointly optimize (1) and (2). Ranking by relevance and setting $k = 15$ achieve the lowest average Brier score.
- In the second stage, we optimize (3) the reasoning prompt. We identify the top 3 prompts out of 15 candidates to elicit 3 predictions from our base model in our system; see Figure 11 for the best one.

In optimizing the reasoning system, we test both Claude-2.1 and GPT-4-1106-Preview as candidate models for generating forecasts. GPT-4-1106-Preview consistently yields a 0.01-0.03 lower Brier score. Therefore, our final system elicits predictions from GPT-4-1106-Preview and the fine-tuned GPT-4-0613.

Ensembling. We implement 5 ensembling methods, including mean, median, geometric mean, trimmed mean, and a variant of universal self-consistency (USC; Chen et al. (2023)). Trimmed mean performs the best in our evaluation; see Section F.2 for details.

27

F.2 Further Details and Results on Hyperparameter Sweep

Throughout the experiment, we set the retrieval date to be the midpoint between a question's open and resolve date. At this time point, the crowd aggregates achieve 0.160 Brier score, averaged over all questions in our validation set.

All the hyperparameter sweeps below evaluate all questions in the validation set.

Search query prompt. We sweep over 6 candidate prompts for generating search queries. The top 2 prompts lead to retrieved articles that have average relevance rating of 3.08 and 3.09, while other prompts below 3.04. Among all articles with rating at least 4, the average rating is 4.37 and 4.38 via the top 2 search query prompts, which is also the highest among all candidates.

Summarization prompt. We sweep over 5 candidate prompts for summarization and evaluate the resulting Brier scores. The best summarization prompt gives a Brier score of 0.193 and the second gives 0.201. In this step of hyperparameter search, the ordering of the summaries, article count and reasoning prompt are randomly chosen for each question to avoid confounding.

Article ordering and count. In this step, we sweep over both orderings of articles (by recency or relevance), and over 5 candidate choices of $k \in [5, 10, 15, 20, 30]$. We run our full system on all questions in the validation set. Presenting 15 article summaries and ordering them by relevance gives the lowest Brier score of 0.177 on GPT-4-1106-Preview. Similar performance can be achieved by presenting 20 articles summaries.

Reasoning prompt. We hand-craft 15 prompts for eliciting forecasts and reasonings. The best prompt (Figure 11) achieves a Brier score of 0.167 on the validation set (while fixing the optimal hyperparameter choices found by the optimization stages above). Two other top prompts get 0.170 and 0.174. The best prompt is given in Figure 11.

Ensembling. We implement 5 ensembling methods, including mean, median, geometric mean, trimmed mean, and a variant of universal self-consistency (USC) (Chen et al., 2023). The last two approaches are defined as follows:

- For the trimmed mean, we assign uniform weights over the input forecasts, reduce the weight of the forecast furthest from the median by half, redistribute the half-weight uniformly to the other forecasts, and finally output the weighted average. We remark that this is not a standard implementation of the trimmed mean, and it is set this way since we only aggregate a small number (i.e., 6) of forecasts in our system.
- For USC, we present the (external) reasoning-prediction pairs to a final LM, which is then prompted to form an aggregated forecast. In this hyperparameter sweep, we use GPT-4-1106-Preview as the aggregator model.

Ensemble Method	Brier Score
Mean	0.1656
Median	0.1651
Geometric Mean	0.1655
Trimmed Mean	**0.1649**
USC (Chen et al., 2023)	0.1691
Baseline (No Ensemble)	0.1676
Human Crowd	0.1600

Table 13: **Brier scores across different ensembling methods on the validation set.** "Baseline" refers to the average Brier score of the base predictions (i.e., the inputs to ensembling).

We evaluate all these methods on the validation set by generating 6 base reasonings for ensembling, using our optimal system setup. Trimmed mean achieves the lowest Brier score; see Table 13 for the results. The USC method, in contrast, does not demonstrate improvement over the baseline.

F.3 News API Evaluations

We justify our choice of the news API's. To begin with, we experiment with 5 eligible APIs to news corpus that accept retrieval date ranges, which, for our purpose, must be specified to prevent

28

leakage. In particular, we assess Google News (accessed via Python open source package `gnews`), NewsCatcher, Newsdata.io, Aylien, and NewsAPI.org.

To assess the quality of their retrievals, we first take 24 unresolved forecasting questions. Next, we prompt GPT-4-1106-Preview to generate two search queries for each of these 24 questions, similar to the first stage of our retrieval system (Section 4). We use these queries to search for articles via all 5 APIs, restricting the retrieval range to the last 24 hours.

Finally, we prompt GPT-4 to rate the relevance of the articles with respect to the original questions. Higher scores indicate greater relevance. We compute the sum of scores of all retrieved articles for each API option. As a result, NewsCatcher and Google News achieve the highest scores of 35 and 39, respectively. The other three APIs, Newsdata.io, Aylien, and NewsAPI.org, score 16.5, 30.5, and 23.5.

F.4 Relevance Rating Approximations

We prompt GPT-3.5 Turbo to score the relevance of all retrieved articles (from Google News and NewsCatcher) with respect to the question. Our prompt is given by Figure 13, where the question's metadata along with an article text is provided to the model in context. The prompt asks the LM to rate the relevance of an article to the given question at the scale of 1–6, where 6 is considered "most relevant" and 1 "irrelevant". In our system, we filter out articles with ratings below 4.

Methods. Due to cost constraints, we cannot afford to evaluate relevance scores using the full article texts. We experimentally explore 3 cost-saving alternatives to approximate full text-based ratings:

1. Title-only. We only give the article title to the model for relevance rating. Unfortunately, via manual inspection, we find the web scraper[4] sometimes fails to to load the full text of an article page, usually due to rendering error or paywall. In such cases, the article content may be incomplete or simply a error message, whereas the title is retrieved and appears relevant. Consequently, the model may be misled by the title. Therefore, we believe that this approach is unviable.
2. Title + first 250 words. We give the article title and the first 250 words to the model for rating.
3. Embedding similarity. We embed the article text and question text metadata and compute their cosine similarities. We threshold by the similarities to filter articles.

Experiment. We experiment with approach (ii) and (iii) above. For (ii), we prompt GPT-3.5-Turbo and Mixtral-8x7B-DPO for relevance rating with the same prompt template (Figure 13). For (iii), we use OpenAI's `text-embedding-3-large` as the embedding model.

To generate raw articles, we randomly sample 15 questions from the validation set, run our retrieval system up to relevant rating, and collect 169 articles in total.

We evaluate the relevance scores of the full texts via GPT-4 (considered as gold labels) and compare with the approximations described above. An article is said be relevant if its rating is at least 4 from the full text query via GPT-4. We compute the recall and precision of the approximate methods as follows.

- At a threshold of 4, title + first 250 words via GPT-3.5-Turbo gives recall of 0.73 and precision of 0.65.
- At a threshold of 3, title + first 250 words via Mixtral-8x7B-DPO gives recall of 0.70 and precision of 0.63.
- At a threshold of 0.48, the cosine similarity method gives recall of 0.73 and precision of 0.54.

In sum, fixing recall at above 70%, querying GPT-3.5-Turbo with title + first 250 words attains the highest precision. The average token length of the articles is 1087.6. Hence, the method of using first 250 words (or roughly 330 tokens) saves about 70% of the cost.

---

[4]https://pypi.org/project/newspaper4k/

29

G System Performance and Calibration

G.1 Performance Across Categories and Platforms

In Table 14 we show performance across different platforms and categories. Across categories, our system exhibits noticeable variations: on Sports, our system nearly matches the crowd aggregate, and on Environment & Energy, it falls much behind. However, we caution against drawing strong conclusions from subcategories, since the sample size is smaller and variation could be due to noise. .

Category	Brier Score $\downarrow$		Accuracy $\uparrow$	
	Ours	Crowd	Ours	Crowd
Science & Tech	$.143_{.011}$	$.114_{.011}$	$82.2_{2.7}$	$84.3_{2.6}$
Healthcare & Biology	$.074_{.015}$	$.063_{.020}$	$93.8_{4.3}$	$90.6_{5.2}$
Economics & Business	$.198_{.007}$	$.147_{.009}$	$68.8_{2.1}$	$78.3_{1.9}$
Politics & Governance	$.172_{.006}$	$.145_{.007}$	$72.6_{1.4}$	$78.2_{1.3}$
Education & Research	$.163_{.024}$	$.129_{.024}$	$80.6_{6.7}$	$77.8_{7.0}$
Arts & Recreation	$.221_{.010}$	$.146_{.010}$	$62.4_{2.5}$	$76.9_{2.2}$
Security & Defenses	$.174_{.008}$	$.129_{.009}$	$71.0_{2.1}$	$78.4_{1.9}$
Sports	$.175_{.004}$	$.171_{.005}$	$73.0_{1.3}$	$73.1_{1.3}$
All Categories	$\mathbf{.179}_{.003}$	$\mathbf{.149}_{.003}$	$\mathbf{71.5}_{.7}$	$\mathbf{77.0}_{.7}$

Platform	Brier Score $\downarrow$		Accuracy $\uparrow$	
	Ours	Crowd	Ours	Crowd
Metaculus	$.134_{.005}$	$.104_{.005}$	$80.3_{1.2}$	$86.6_{1.1}$
GJOpen	$.193_{.011}$	$.157_{.013}$	$67.9_{3.4}$	$72.6_{3.2}$
INFER	$.247_{.053}$	$.310_{.086}$	$60.0_{13.1}$	$53.3_{13.3}$
Polymarket	$.172_{.005}$	$.127_{.006}$	$73.6_{1.3}$	$79.9_{1.1}$
Manifold	$.219_{.004}$	$.200_{.005}$	$63.6_{1.3}$	$67.9_{1.3}$
All Platforms	$\mathbf{.179}_{.003}$	$\mathbf{.149}_{.003}$	$\mathbf{71.5}_{.7}$	$\mathbf{77.0}_{.7}$

Table 14: **Results of system evaluation** by category **(left)** and by platform **(right)**. Our system achieves .179 Brier score (human crowd: .149) and accuracy .715 (human crowd: .770). Brier score and accuracy are averaged across retrieval dates and then questions. Subscript numbers are 1 standard error.

G.2 Calibration

We observe that our system is well calibrated on the validation set and test set (Figure 15a, Figure 15c). Interestingly, this is not the case in the baseline evaluations, where the models are not well calibrated in the zero-shot setting (Figure 15b). Through fine-tuning and ensembling, our system improves the calibration of the base models, without undergoing specific training for calibration.

(a) Calibration of System on Val.   (b) Calibration of Base LMs on Test   (c) Calibration of System on Test

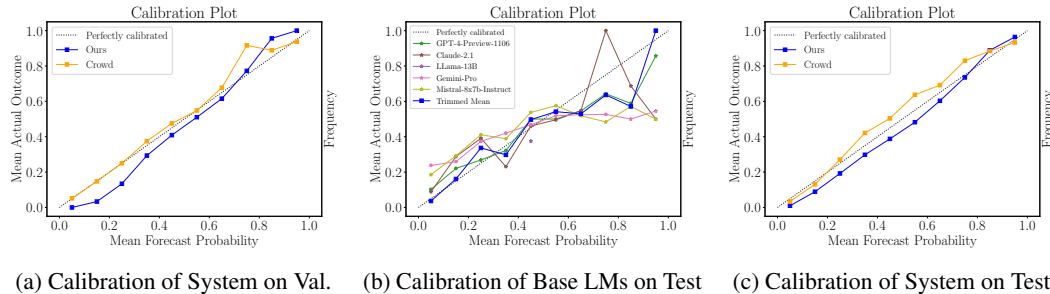



Figure 15: **Our system is naturally well calibrated** on both (a) validation and (c) test. The crowd is also well calibrated, consistent with Zou et al. (2022)'s findings. In contrast, the base models in the zero-shot setting (b) are less calibrated (Section 3.3) on the test set.

30

H Future Directions and Limitations

We discuss a few opportunities to improve these systems further.

Iterative self-supervision. With a larger training corpus, our self-supervised fine-tuning approach can be used for iterative self-improvement. Specifically, after fine-tuning a model on its previous optimal predictions and reasonings, we can generate more fine-tuning data by using the same model again, which can be repeated until training data is exhausted.

Data. While our forecasting benchmark is a good initial corpus to train a system, we believe that it is possible to use LMs with later training cutoffs to teach an earlier LM. This could be done by using later LMs to generate questions it knows the answer to but an earlier LM does not (postdiction). In addition, while we source questions from forecasting platforms, it is possible to collect historical data in the wild and re-formulate them as forecasting questions, leading to a larger training set.

Domain-adaptive training. In Section B.3, we observe that in the baseline evaluations, the Brier scores across categories are correlated with models' pre-training knowledge. This suggests that we may be able to specialize models to areas of particular interests by fine-tuning them on domain knowledge.

LMs get better at forecasting naturally. We observe that as LMs improve, they naturally also become better at forecasting. In particular, in Section 3.3, we see that newer generations of models forecast better than older ones. For example, GPT-4-1106, released in 2023, outperforms GPT-4-0613, released in 2021, by .02 with respect to the Brier score. If we were to have fine-tuned the more recent model, we would expect better performance.

At a high level, our results suggest that in the near future, LM-based systems may be able to generate accurate forecasts at the level of competitive human forecasters. We hope that our work paves the way for automated, scalable forecasting that can help to inform institutional decision making.

Limitations. Our work attains near human-level results on event forecasting. It comes with a few limitations.

- We source our forecasting questions from 5 platforms as discussed in Section 3. While we believe they are representative, we have not tried to exhaustively gather questions from all forecasting platforms and prediction markets. Notably, we do not scrape questions on Kalshi or PredictIt.
- In Section 6, we hypothesize that our system is hurt by the LM's tendency to hedge its answers. Such tendency may arise from safety training. Future work should investigate our hypothesis and propose potential fix.
- Our system's performance is influenced by the quality of news retrieval. Improving that part of the system is an important future direction.
- Our system requires extensive use of LLMs, which use significant computational costs.

31

I Broader Impact

This paper presents work whose goal is to advance the field of automated forecasting. We believe that our work can help to inform human forecasters and institutional decision makers, and will generate positive social impact.

I.1 Positive Impacts

Enhanced Forecasting Accuracy The proposed system approaches the performance of competitive human forecasters and even surpasses them in some settings. This indicates that language models (LMs) can generate accurate forecasts, which can significantly improve decision-making processes in various fields such as economics, geopolitics, and epidemiology .

Scalability and Efficiency Automating the forecasting process with LMs can provide timely and cost-effective forecasts. Unlike human forecasters who may require extensive time and resources, LMs can quickly parse and produce forecasts based on large datasets, making the forecasting process more scalable.

Transparency and Understanding The system elicits explanatory reasonings from LMs through prompts, allowing users to understand the rationale behind the forecasts. This transparency can enhance trust in the automated forecasts and facilitate better-informed decisions.

Improved Decision Making With the ability to provide accurate and timely forecasts, institutions can make more informed decisions. For example, governments can rely on these forecasts for policy-making, while businesses can use them for strategic planning.

I.2 Negative Impacts and Mitigation Solutions

Over-reliance on Automated Systems There is a risk that decision-makers might over-rely on automated forecasts, potentially overlooking critical insights that human judgment could provide. **Mitigation**: Encourage a hybrid approach where automated forecasts complement human judgment.

Data Quality and Bias The accuracy of forecasts depends on the quality of the reasoning data used. If the data is biased or incomplete, the forecasts may also be biased or inaccurate. **Mitigation**: Implement robust data validation and cleaning processes. Additionally, use diverse data sources to minimize bias and ensure comprehensive coverage.

32

J   Example Outputs

We provide 5 qualitative example outputs from our fine-tuned GPT-4 model on the test set.

---

1. Rephrased and Expanded Question: Is it likely that Kevin McCarthy, the former House Speaker, will make a public announcement or have his resignation reported by credible sources, indicating his departure from the United States House of Representatives, before the end of November 30, 2023? This question takes into account the current political climate, McCarthy's recent loss of the Speaker position, his consideration of not seeking re-election, and the potential impact of his decision on the Republican Party and the 20th District of California. The resolution of this market hinges on whether McCarthy will officially declare his intention to vacate his seat within the specified timeframe.

2. Reasons why the answer might be no:
- McCarthy has not made a final decision and has expressed uncertainty about his future in Congress, which could mean he is still weighing his options and may not choose to resign within the given timeframe. (Strength: Strong)
- The articles suggest that while there is speculation and consideration of resignation, there is no definitive statement or action indicating an imminent resignation. (Strength: Strong)
- McCarthy may be waiting to see how the political landscape shifts or to make a decision after discussing with his family, which could extend beyond the November 30 deadline. (Strength: Moderate)
- Given his history of strong financial support and political influence, McCarthy may feel compelled to stay and influence the political process, especially with the upcoming 2024 elections. (Strength: Moderate)

3. Reasons why the answer might be yes:
- There are reports and hints from McCarthy himself that he is considering resignation and may want to "get the hell out," which could indicate a potential for an early departure. (Strength: Moderate)
- The pressure from political opponents and the recent loss of the Speaker position may contribute to a decision to resign, especially if he feels his influence is waning. (Strength: Moderate)
- The possibility of a special election being called could provide an opportunity for a controlled transition of power, which might be appealing to McCarthy. (Strength: Weak)

4. Aggregate Considerations:
While there is speculation and some suggestive comments from McCarthy about the possibility of resignation, there is no concrete evidence or announcement that he will resign by the specified date. His history of fundraising and political influence, as well as the lack of a definitive statement about his future, suggest that a resignation by November 30 is not certain.

5. Initial Probability:
Given the considerations above, the initial probability of McCarthy resigning by November 30 seems low.

6. Adjusted Probability Considerations:
- The calculated probability should not be excessively confident in either direction due to the uncertainty expressed by McCarthy himself.
- The base rate for such high-profile resignations in the face of uncertainty is not well-established, which adds to the difficulty in predicting this event.
- Intuitively, politicians often weigh their options carefully, and a decision to resign is not typically rushed, especially without a clear strategic reason.

7. Final Answer: *0.20*

Figure 16: A response of the fine-tuned model on the question "Will Kevin McCarthy resign from the House by Nov 30?" (resolved NO) at the last retrieval index. Inspecting the original prompt (not shown here), we notice that our retrieval system presents articles, published on November 30, that clearly state that no official announcement had yet been made about Kevin McCarthy's potential resignation. On the day of November, 30, the community prediction had dropped to 1%, but our system hedges by claiming 20%.

33

1. Will Reddit make any official announcements prior to the 1st of July, 2023, regarding alterations to the pricing structure or a postponement of the implementation of fees for its Application Programming Interface (API), which is currently set to charge developers for higher usage? This question will be resolved affirmatively if Reddit communicates any of the following changes: a different pricing model than the current "$0.24 per 1K API calls," adjustments to the free API usage thresholds, a delay in the implementation date beyond July 1, 2023, or any negotiations for specific rates with individual organizations do not count. The resolution of this question will be determined by Metaculus, and the question may be annulled if the announcement is unclear or does not meet the specified criteria.

2. Reasons why the answer might be no:
- Reddit's CEO, Steve Huffman, has consistently shown a firm stance on the API changes, emphasizing the necessity for the company to move forward with the new pricing structure to ensure profitability and sustainability. This resolute position suggests that a change in the announced API fee pricing is unlikely.
- Despite the significant backlash and coordinated protests from thousands of Reddit communities, there has been no indication from Reddit's internal communications or public statements that they are considering altering the API fee pricing or its implementation timeline.
- The internal memo from Reddit's CEO, where he mentioned that the protest "will pass" and that there has been no significant revenue impact from the protests, indicates that the company may be prepared to weather the discontent and proceed with the changes as planned.

3. Reasons why the answer might be yes:
- The widespread protest and the large number of communities going dark could potentially lead to a significant user and developer exodus, which might force Reddit to reconsider its position to avoid a substantial negative impact on its platform's activity and content.
- The specific mention of negotiations with individual organizations for different rates could be a sign that Reddit is open to adjusting the pricing for certain parties, which might lead to broader changes if the protests continue or if key developers threaten to shut down their apps.
- The negative publicity and user dissatisfaction could reach a tipping point where the company decides that the long-term damage to its brand and community is not worth the short-term financial gain, prompting a change in the API fee pricing policy.

4. Rating the strength of each reason:
- The strength of the resolute stance of Reddit's CEO is high, as it is based on consistent public statements and the company's broader financial strategy.
- The lack of any indication of a potential change in the API fee pricing is also a strong reason, given the proximity to the implementation date and the current lack of communication suggesting otherwise.
- The internal memo's dismissive tone about the protests and lack of revenue impact is a moderate reason, as it reflects the current mindset of the company but does not preclude a change in response to escalating pressure.
- The potential for a user and developer exodus is a moderate reason, as it is speculative and depends on the actual impact of the protests, which has not been fully realized yet.
- The mention of negotiations with individual organizations is a weak reason for a broader change, as it does not necessarily indicate a willingness to change the policy for all developers.
- The possibility of a tipping point due to negative publicity is a moderate reason, as it is based on the assumption that Reddit will prioritize its long-term brand and community over short-term financial gains, which is uncertain.

5. Aggregate your considerations:
Considering the strong and consistent stance of Reddit's leadership, the lack of any public indications of a change, and the dismissal of the protests' impact on revenue, it seems unlikely that Reddit will announce changes to its API fee pricing before July 1, 2023. However, the potential for negative publicity and the impact of the protests on the platform's activity cannot be entirely dismissed, which introduces some uncertainty into the forecast.

6. Output your answer: Given the information available and the considerations above, the likelihood of Reddit announcing changes or a delay to its proposed API fee pricing before July 1, 2023, seems low. My answer, therefore, is: *0.2*

Figure 17: A response from our fine-tuned model on the question "Will Reddit announce changes or a delay to its proposed API fee pricing before July 1, 2023?" (resolved NO). The forecast was made at the retrieval date June, 14, 2023, one day after the question was opened. At the time, the community gives a forecast of 40%, where this response suggests 20% which turned out to be closer to the resolution.

34

1. Rephrased and Expanded Question:
Considering the current geopolitical tensions and recent historical events, is it likely that the State of Israel will engage in a military operation that results in the death of at least five Iranian nationals within Iran's borders, and subsequently either publicly claim responsibility for this action or execute the operation in such a clear and overt manner that credible sources can unambiguously attribute the attack to Israel, all before the end of the year 2023?

2. Decision Tree Outline:
- Yes Outcome:
    - Israel conducts a military operation against Iran.
        - The operation is overt and clearly attributable to Israel.
        - Israel publicly acknowledges and takes credit for the operation.
    - The operation results in at least five Iranian deaths within a 24-hour period.
- No Outcome:
    - Israel does not conduct a military operation against Iran.
    - Israel conducts an operation, but it is covert and not attributable.
    - Israel conducts an operation, but does not publicly acknowledge it.
    - The operation does not result in at least five Iranian deaths within a 24-hour period.

3. Branch Probability Analysis:
- Yes Outcome:
    - Given the recent escalation in tensions and historical precedents, there is a non-negligible chance that Israel could conduct a military operation against Iran.
    - The likelihood of Israel publicly acknowledging such an operation is lower, as they often maintain plausible deniability.
    - The probability of an operation resulting in at least five Iranian deaths is uncertain and would depend on the scale and nature of the operation.
- No Outcome:
    - The probability of no military operation or a covert operation is higher, given Israel's historical preference for covert actions and the potential for international backlash.
    - The likelihood of Israel not acknowledging an operation is also high, as this is a common practice to avoid direct conflict and international pressure.
    - The chance of an operation not meeting the death threshold is also possible, especially if the operation is designed to be limited or surgical.

4. Discussion on Wildcards:
- A significant escalation in the conflict between Israel and Iran, such as a direct attack on Israeli soil attributed to Iran, could increase the probability of a 'Yes' outcome.
- International diplomatic efforts or significant pressure on Israel to de-escalate could decrease the probability of a 'Yes' outcome.
- Unpredictable events, such as a major terrorist attack or a sudden change in leadership in either country, could also impact the forecast.

5. Initial Probability (Prediction):
Given the current information and the outlined decision tree, the initial probability of a 'Yes' outcome seems low but not impossible. I would estimate it to be around 0.20.

6. Adjusted Probability Thoughts:
- The calculated probability may be too confident given the volatile nature of the Middle East and the unpredictability of military engagements.
- The recent death of an Iranian general in an Israeli airstrike could escalate tensions, potentially increasing the likelihood of a 'Yes' outcome.
- The lack of direct evidence linking Iran to the October 7 Hamas attack and the potential for covert operations could decrease the likelihood of a 'Yes' outcome.

7. Final Answer: *0.15*

Figure 18: A response from our fine-tuned model on the question "Will Israel carry out and explicitly acknowledge a deadly attack on Iran before 2024?" (resolved NO). In our fine-tuning phase, one of the prompts we use to generate strong reasonings asks the model to build a decision tree and assign probabilities. The fine-tuned model learns the reasoning path (without being explicitly prompted to do so).

35

1. Rephrased and Expanded Question:
Is it confirmed that former President Donald J. Trump will be present and actively participate in the inaugural Republican National Committee (RNC) presidential primary debate for the 2024 election cycle, which is scheduled to take place on August 23, 2023? The criteria for his participation include him speaking at least once during the debate and being recognized by the debate's moderator. This applies even if he joins the debate remotely. The resolution of this question will consider all segments of the debate if it is spread over multiple days. However, if the debate is postponed beyond September 30, 2023, the market will resolve to "No." The resolution will be based primarily on the actual debate footage, supplemented by a consensus from credible news reports.
2. Reasons why the answer might be no: - Trump may choose to skip the debate to maintain a strategic advantage or to avoid early confrontations with other candidates.
- He might not want to share the stage with candidates he perceives as less significant, which could dilute his brand or message.
- There could be concerns about the debate format or the choice of moderators, which might lead to a decision not to participate.
3. Reasons why the answer might be yes:
- Participating in the debate would give Trump a platform to outline his policies and vision, which could be beneficial for his campaign.
- It would allow him to directly engage with and challenge his opponents, potentially strengthening his position.
- Trump might want to demonstrate leadership and confidence by being present at the first major event of the primary season.
4. Rating the strength of each reason:
- The strategic advantage of skipping the debate: Moderate. Trump has a history of unconventional campaign strategies, but skipping the first debate is a risky move.
- Avoiding sharing the stage: Moderate. Trump's brand is strong, but he may not want to risk being overshadowed or challenged by multiple opponents.
- Concerns about format or moderators: Moderate to High. Trump has previously expressed dissatisfaction with debate arrangements, which could influence his decision.
- Benefiting from outlining policies: High. The debate is a significant opportunity to reach a large audience and make a case for his candidacy.
- Engaging with opponents: High. Trump is known for his confrontational style, which could play out to his advantage in a debate setting.
- Demonstrating leadership: High. As a former president, Trump might want to assert his dominance in the Republican field from the outset.
5. Aggregate your considerations:
Considering the reasons and their strengths, it seems there is a balanced set of factors that could influence Trump's decision to participate or not. His history of unconventional tactics and concerns about debate arrangements could lead to a "No," while the potential benefits of participating in terms of policy promotion, engaging opponents, and demonstrating leadership suggest a "Yes." The decision will likely depend on the specific circumstances closer to the debate date and Trump's assessment of the strategic value of his participation.
6. Output your answer: Given the balanced considerations and the unpredictable nature of Trump's campaign strategies, I would assign a probability that leans slightly towards participation due to the high benefits it offers, but not overwhelmingly so due to the potential strategic considerations against it.
0.55

Figure 19: A response from our fine-tuned model on the question "Will Trump attend the first RNC debate?" (resolved NO on August, 23, 2023). The forecast was made at the retrieval index 1 on June 15, 2023, 2 days after the question was opened. At the time, the community predicted 60%, close to our model's forecast of 55% from this response.

36

1. Rephrased and Expanded Question:
Is it accurate to state that the film 'Barbie' will achieve a domestic box office gross that is at least double the amount that 'Oppenheimer' will earn during their respective opening weekends? This comparison will be based on the final reported figures for the three-day opening weekend (July 21-23) in the United States, as documented on Box Office Mojo. The outcome will be determined by comparing the "Domestic Weekend" grosses of 'Barbie' and 'Oppenheimer,' and the market will resolve in favor of 'Yes' if 'Barbie's gross is at least twice that of 'Oppenheimer.' If no final data is available by July 31, 2023, an alternative credible source will be used to resolve the market.
2. Reasons why the answer might be no:
- 'Oppenheimer' could have a stronger than expected performance, reducing the ratio between the two films' grosses. (Strength: Weak, given the current projections and reports.)
- There could be a significant underestimation of 'Oppenheimer's appeal to a specific demographic that turns out in large numbers. (Strength: Weak, as current data does not support this.)
- Unforeseen events could negatively impact 'Barbie's attendance, such as controversies or widespread technical issues in theaters. (Strength: Weak, as there is no indication of such events.)
3. Reasons why the answer might be yes:
- 'Barbie' has already been reported to have a significantly higher opening weekend gross than 'Oppenheimer,' with figures exceeding the 2x mark. (Strength: Strong, based on multiple reports.)
- 'Barbie' is playing in more theaters than 'Oppenheimer,' which could contribute to a higher gross. (Strength: Moderate, as screen count can influence box office performance.)
- 'Barbie' has broader appeal and targets a wider demographic, including families and younger audiences, which could lead to higher attendance. (Strength: Moderate, as demographic appeal is a significant factor in box office performance.)
4. Aggregate your considerations:
Considering the information provided, 'Barbie' has a strong lead over 'Oppenheimer' in terms of opening weekend gross, with multiple reports indicating that it has already achieved more than double the gross of 'Oppenheimer.' The wider release and demographic appeal of 'Barbie' further support the likelihood of it maintaining at least a 2x lead in gross over 'Oppenheimer.' There is little evidence to suggest that 'Oppenheimer' will close the gap significantly enough to prevent 'Barbie' from doubling its gross.
5. Initial probability (prediction):
Given the strong evidence in favor of 'Barbie' doubling 'Oppenheimer's gross, the initial probability is high.
6. Evaluate whether your calculated probability is excessively confident or not confident enough:
The calculated probability seems confident, but it is based on concrete data and reports from reliable sources. However, it is important to consider the possibility of unforeseen factors that could affect the final outcome. While the probability is high, it should not be absolute to account for any potential uncertainties.
7. Output your final prediction: *0.95*

Figure 20: A response from our fine-tuned model on the question "Will 'Barbie' gross 2x more than 'Oppenheimer' on opening weekend?". The question resolved NO on July 24, 2023. On July 20, 2023, the community gave 73% and the response above gives 95%. In the original prompt (not shown here), our news retrieval provides projections that Barbie will likely outperform Oppenheimer at the box office. However, the model hallucinates them as facts ("[...] with figures exceeding the 2x mark"), resulting in overconfidence.

37

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We have made the introduction and abstract summarize the main findings of the paper. The introduction also provides an overview of our methodology.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We have discussed multiple future directions as well as limitations in Appendix H.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: Our work is not theoretical in nature.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.

38

- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We have discussed our methodology, model usage and algorithm in details throughout the paper. We also attach the code and data in our submission.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We have submitted our code and data along with paper.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.

39

- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: All experimental settings are discused in detail in the relevant sections.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: We have provided standard errors of all the evaluation results.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

40

Answer: [Yes]

Justification: We have given detailed methodology of our work throughout the paper. In particular, our method only requires API access to the LMs, which we specify in Section A.2.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

Answer: [Yes]

Justification: Our research conform with the NeurIPS code of ethics.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: We discuss social impact of our work in Appendix I.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

41

Answer: [NA]

Justification: Our work builds a scaffold on top of existing LMs. The result does not yet supersede the forecasts available on prediction markets. We believe that there is not high risk for misuse.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: On the data side, we only used existing questions on public forecasting platforms. We have acknowledged them in Section 3. Our code is fully original. Finally, we use LMs developed by other institutions which we specify in Section A.2.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: Our code and data are well documented.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: Our work involves no human subjects.

42

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: Our work involves no human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

43