PROPHET: An Inferable Future Forecasting Benchmark with Causal Intervened Likelihood Estimation

Zhengwei Tao[12] **Zhi Jin**[12]✉ **Bincheng Li**[3] **Xiaoying Bai**[4]✉ **Haiyan Zhao**[12]
Chengfeng Dou[12] **Xiancai Chen**[12] **Jia Li**[12] **Linyu Li**[12] **Chongyang Tao**[5]
[1]Key Laboratory of High Confidence Software Technologies (PKU), MOE, China
[2]School of Computer Science, Peking University
[3]Guangzhou University, [4]Advanced Institute of Big Data [5]SKLSDE Lab, Beihang University
{tttzw,xiancaich}@stu.pku.edu.cn, {zhijin,zhhy.sei,chengfengdou,lijiaa}@pku.edu.cn
baixy@aibd.ac.cn, chongyang@buaa.edu.cn

Abstract

Predicting future events stands as one of the ultimate aspirations of artificial intelligence. Recent advances in large language model (LLM)-based systems have shown remarkable potential in forecasting future events, thereby garnering significant interest in the research community. Currently, several benchmarks have been established to evaluate the forecasting capabilities by formalizing the event prediction as a retrieval-augmented generation (RAG)-and-reasoning task. In these benchmarks, each prediction question is answered with relevant retrieved news articles. However, because there is no consideration on whether the questions can be supported by valid or sufficient supporting rationales, some of the questions in these benchmarks may be inherently noninferable. To address this issue, we introduce a new benchmark, PROPHET, which comprises inferable forecasting questions paired with relevant news for retrieval. To ensure the inferability of the benchmark, we propose Causal Intervened Likelihood (CIL), a statistical measure that assesses inferability through causal inference. In constructing this benchmark, we first collected recent trend forecasting questions, and then filtered the data using CIL resulting in an inferable benchmark for event prediction. Through extensive experiments, we first demonstrate the validity of CIL and in-depth investigations into event prediction with the aid of CIL. Subsequently, we evaluate several representative prediction systems on PROPHET, drawing valuable insights for future directions. The dataset is available on https://github.com/TZWwww/PROPHET.

1 Introduction

The quest to predict future events has long been a central pursuit in the field of artificial intelligence (AI). The ability to foresee outcomes and

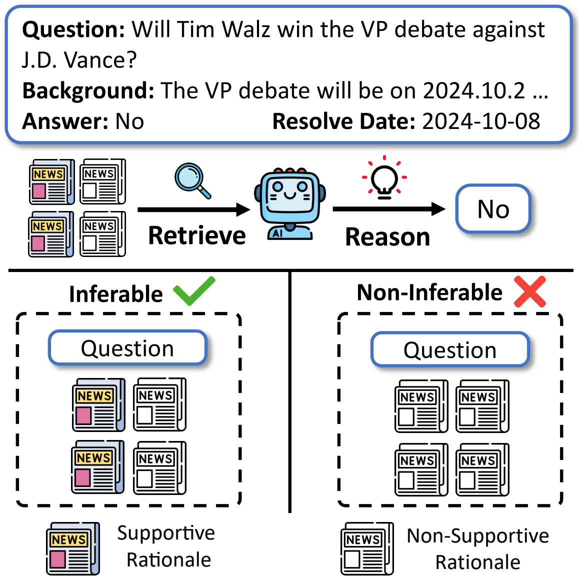

Figure 1: The upper Figure demonstrates the task of future forecasting. The lower half shows both inferable and non-inferable scenarios.

trends holds the promise of revolutionizing numerous sectors covering finance (Li et al., 2024), climate science (Wang and Karimi, 2024), and social policy (Rotaru et al., 2022). Recent years have witnessed a surge in interest and progress, particularly with the advent of large language model (LLM)-based systems. These systems, leveraging the power of deep learning and vast amounts of data, have demonstrated an unprecedented capacity for forecasting, capturing the imagination and focus of the research community (Halawi et al., 2024; Hsieh et al., 2024; Pratt et al., 2024).

To evaluate the abilities of these LLM-based future forecasting systems, pilot works construct several benchmarks based on real-world forecasting questions (Halawi et al., 2024; Guan et al., 2024; Karger et al., 2024). These benchmarks have successfully framed future forecasting as a retrieval-augmented generation (RAG)-and-reasoning task. Within this framework, systems should first search the Web or databases for news articles related to the prediction question in the benchmarks to gain

✉ Corresponding author.

knowledge base, then reason based on the retrieved knowledge base. Nevertheless, in order to truly evaluate the abilities of the LLM-based future forecasting, the prediction questions in the benchmarks need to be inferable, meaning that the supporting knowledge base must contain sufficient information to substantiate the answers. In traditional RAG tasks, the answer can definitely be found within the knowledge base. However, future forecasting tasks do not inherently satisfy this characteristic compared to traditional RAG benchmarks such as HotpotQA (Yang et al., 2018) and 2WikiMultiHopQA (Ho et al., 2020). That is, future forecasting needs to be inferred by rationales, i.e. facts and reasoning clues, but the knowledge base may only provide partially supportive rationales for the prediction questions (Zhao et al., 2024). Collecting real-world prediction questions as the benchmark without nuanced validation, the knowledge base may not be able to provide sufficient supportive facts which makes some of the prediction questions non-inferable (Birur et al., 2024).

To overcome this challenge and advance the field, we introduce an inferable future forecasting benchmark, PROPHET, designed to provide a more accurate evaluation. To ensure reproducibility, PROPHET is an RAG task where each prediction question pairs with relevant downloaded news articles for retrieval. We are next motivated to select prediction questions that are inferable, based on their related articles. The most challenging part is to estimate the inferability of each question since we cannot observe the completed real-world event evolution process. Even if we can, it is difficult to determine as well, due to the lack of expert knowledge of a wide spectrum of domains. A key innovation in our approach is the introduction of Causal Intervened Likelihood (CIL), a statistical measure that assesses the inferability of prediction questions through causal inference. CIL is calculated via principles of causal inference where we measure the supporting degree of each article for the answer to the question. We regard each article as an event and compute the effect of intervening in the event from happening to not happening. CIL provides a robust estimate of whether a question can be answered. We then filter the prediction questions using CILto ensure the inferability of the benchmark, providing a fair and accurate evaluation of the systems' forecasting ability. Assisted by CIL, PROPHET performs as a more well-formulated RAG-and-reasoning task with hidden rationale (Zhao et al., 2024).

To validate the effectiveness of CIL , we conducted a series of extensive experiments. These experiments were designed to rigorously test how this estimation can represent the inferability of prediction questions. The results of the experiments were highly encouraging, demonstrating a strong correlation between CIL scores and the actual performance of the systems in terms of both retrieval and prediction accuracy. Further, CIL enables us to conduct in-depth investigations into future forecasting, drawing out innate properties of this complicated task. Finally, we evaluated several state-of-the-art prediction systems on the PROPHET benchmark. This evaluation provided effective measurements of the strengths and weaknesses of each system, highlighting areas for improvement and potential directions for future research. We will also regularly update the dataset to ensure its timeliness and to minimize the risk of data leakage due to model evolution. To summarize our contribution:

- We are the first to introduce CIL for inferability estimation of the future forecasting questions and provide a feasible method for calculating this metric.
- Assisted by CIL, we establish an automatic pipeline to construct the future forecasting benchmark PROPHET where the prediction questions are insufficiently inferable based on their related articles.
- We evaluate several baselines for future forecasting. The results show the pros and cons of these systems and present great potential and development directions for this task.

2 Related Work

2.1 Future Forecasting and Benchmarks

Previous research on future forecasting benchmarks has evolved in different paradigms, each addressing different aspects of the task. Early benchmarks, such as MCNC (Granroth-Wilding, 2016), SCT (Mostafazadeh et al., 2017), and CoScript (Yuan et al., 2023), focused on script learning and common sense reasoning in synthetic scenarios. Although these data sets facilitated structured reasoning, they lacked real-world applicability and grounding in factual news. Time series datasets such as GDELT (Leetaru and Schrodt, 2013) and ICEWS (Schrodt et al., 2012) introduced real-world event tracking but did not formalize prediction as a retrieval-augmented reasoning task or ensure answerability. Later works, such as ECARE (Du et al., 2022) and EV2 (Tao et al., 2024), advanced event reasoning but remained confined to settings without real-world grounding.

With the rise of LLMs, recent benchmarks such as Halawi et al. (2024), OpenEP (Guan et al., 2024), and ForecastBench (Karger et al., 2024) shifted the focus to real-world questions and news-based search. However, these datasets suffer from two critical limitations: (1) they lack explicit validation of inferability, allowing questions with insufficient supporting evidence to persist, and (2) they prioritize dynamic data sources over reproducibility, risking inconsistent evaluations due to evolving news archives. `PROPHET` addresses these gaps by filtering via the introduced Causal Intervened Likelihood estimation. We show the benchmark comparison in Table 4.

2.2 RAG and Benchmarks

Foundational QA Datasets for RAG: Traditional QA datasets, including MMLU (Hendrycks et al., 2021), StrategyQA (Geva et al., 2021), ASQA (Stelmakh et al., 2022), Multi-HopQA (Lin et al., 2020), and 2WikiMultiHopQA (Lin et al., 2020), are adapted to evaluate RAG systems. These datasets, grounded in knowledge bases like Wikipedia, form the basis for RAG evaluation.

Domain-Agnostic: RAGBench (Friel et al., 2024) is a multi-domain benchmark across biomedical, legal, customer support, and finance domains. CRAG (Wang et al., 2024a) provides a factual QA benchmark across five domains, simulating web and knowledge graph search.

Domain-Specific: Domain-specific benchmarks include LegalBench-RAG (Wang et al., 2024b), WeQA (Meyur et al., 2024), PubHealth (Zhang et al., 2023), and MTRAG (Tang and Yang, 2024). These benchmarks address niche applications and improve evaluation precision in domains.

Capability-Oriented: RGB (Liu et al., 2024) evaluates four RAG capabilities: noise robustness, negative rejection, information integration, and counterfactual robustness. TRIAD (Zong et al., 2024) assesses retrieval quality, fidelity, and task-specific utility through a three-dimensional framework.

In this work, we focus on the inferability of RAG benchmarks, a key property for domain-specific and real-world scenarios. Our method can be generalized to other domains.

3 Preliminaries

3.1 Future Forecasting

Future forecasting stands for predicting whether a certain event will happen in the future based on the events that occurred. We now formalize the task as a binary question-answering task. Given a prediction question $\mathcal{Q}$ which can be "Will Tim Walz win the VP debate against J.D. Vance?" or "Will Bitcoin rise to $100,000 by December 2024?". There would be background information $\mathcal{B}$ that describes the context of $\mathcal{Q}$ and resolution criteria $\mathcal{R}$ explaining how the question can be regarded as answered. A large set of documents $\mathbb{X}$ serves as a knowledge base to retrieve. The forecasting system must answer the question as:

$$\mathcal{Y} = \mathrm{Reason}(\mathcal{Q}, \mathcal{B}, \mathcal{R}, \mathrm{Retrieve}(\mathcal{Q}, \mathbb{X})), \quad (1)$$

where $\mathcal{Y} \in [0, 1]$ is the predicted probability of how likely the event in $\mathcal{Q}$ would occur. A ground truth answer $\hat{\mathcal{Y}} \in \{0, 1\}$ paired with a resolved date $\mathcal{D}$ represents whether the event in $\mathcal{Q}$ finally occurs and the date the question resolves. As the same in previous works (Halawi et al., 2024; Karger et al., 2024), we use Brier Score (Brier, 1950) as the metric for evaluation:

$$\texttt{Brier Score} = \frac{1}{N}\sum_{n}^{N}(\mathcal{Y}_n - \hat{\mathcal{Y}}_n)^2, \quad (2)$$

N is the number of the questions in the dataset.

We formalize future forecasting as an RAG task. As an RAG, it features distinctly compared with traditional dataset such as HotpotQA (Yang et al., 2018) and 2WikiMultiHopQA (Ho et al., 2020). The knowledge base $\mathbb{X}$ stores the rationales and clues for answering $\mathcal{Q}$ (Zhao et al., 2024). Future forecasting mainly detects two core entangled abilities of the systems: retrieval and reasoning.

Current future forecasting benchmarks are constructed by harvesting real-world prediction questions and paired with news articles before the resolved date $\mathcal{D}$ (Halawi et al., 2024; Guan et al., 2024; Karger et al., 2024) without nuanced validation of the inferability of the questions. It is possible that there is a lack of sufficient supportive information in $\mathbb{X}$ for the question. Methods need to be established to ensure that the prediction questions in the benchmarks are sufficiently inferable.

3.2 Causal Inference

Causal inference is a vital statistical method to determine causal relationships between variables (Pearl, 2010). In real-world scenarios, a mere correlation between two variables may be due to chance or hidden factors. Causal inference aims to establish direct causality. For example, the increase in ice cream sales and drowning incidents is not a causal link, although both are affected by hot weather. Causal inference uses concepts such as structural causal models, interventions, and counterfactual inferences. These are applied in medicine, economics, and social sciences.

Structural causal model (SCM) It is a framework designed to represent and analyze causal relationships between variables using a combination of causal graphs and structural equations. At its core, SCM relies on a directed graph where nodes represent variables $\mathcal{X}$, and edges denote direct causal influences, forming a network that captures dependencies and pathways of causation. Each variable in the model is determined by its direct causes (parent nodes). SCM enables the identification of causal effects, and exploration of intervention questions (e.g., "What would happen if we intervened on X?"). This has been widely applied in fields like epidemiology, economics, and machine learning to disentangle complex causal mechanisms and validate hypotheses (Stolfo et al., 2023).

Interventional distribution An SCM allows the study of interventions. An atomic intervention $\text{do}(\mathcal{X}_i = x)$ fixes $\mathcal{X}_i$ with a fixed value x. For example, in a medical trial, the dose of a new drug is set at a specific value for a group. In the view of structural causal model, interventions can be understand as changing of the original structure and variable distributions. After $\text{do}(\mathcal{X}_i = x)$, the resulting distribution is $P(\cdot|\text{do}(\mathcal{X}_i = x)) \doteq P_m(\cdot|\mathcal{X}_i = x)$, which shows how other variables respond.

4 PROPHET Benchmark

In this section, we introduce PROPHET which is an future forecasting benchmark with inferability estimation and selection. We first describe the data collection process in Section 4.1. Then we introduce the Causal Intervened Likelihood (CIL) metric in Section 4.2. We finally describe the benchmark construction in Section 4.3.

4.1 Data Collection

Our objective is to gather a dataset that encompasses recent and prominent prediction questions. To achieve this, we have sourced questions from two well-known platforms: Metaculas[1] and Manifold[2]. The choice of these source websites, Metaculas and Manifold, is well justified for constructing the benchmark. The domains covered by the questions on these platforms are highly diverse, ranging from scientific breakthroughs to social and economic trends. This diversity ensures that the benchmark is representative of a wide spectrum of forecasting tasks. Moreover, the questions are trending and among the most attention-attracting ones on these platforms. This indicates that they are not only relevant in the current context but also likely to be of interest to the broader forecasting community. As such, the data collected from these sources provides a robust foundation for evaluating and developing practical forecasting models.

To avoid model leakage, we carefully selected questions. From Metaculas, we chose questions resolved in August 2024 along with meta-information. Since there were few pre-August 2024 questions on Metaculas, we added questions resolved before August from Manifold, ensuring both the latest trends and a historical perspective. We filtered out meaningless questions, such as personal inquiries or those with little community interest, to focus on realistic forecasting scenarios.

After collecting questions, we collected relevant news articles. Using GPT4o-mini[3], we generated three types of news search queries per question: entities in the question, resolving steps, and similar historical events using prompts in the Appendix A.7 (a-c). Then we searched on the MediaCloud open-source platform[4] with these queries. MediaCloud's vast news repository helped us gather comprehensive information. However, many retrieved articles were irrelevant. To address this, we used GPT4o-mini again to filter the articles, retaining 100 relevant ones per question by prompt in the Appendix A.7 (d). That reduces noise and mimics real-world prediction analysis.

4.2 Causal Intervened Likelihood

To measure the sufficiency of the supportive rationales of each question and construct an inferable benchmark, we introduce a statistic estimation named Causal Intervened Likelihood (CIL) via causal inference. CIL estimates the supportivity of each news article to the question. We use Bernoulli variables to model the occurrence of events. Specifically, let $Y \in \{0, 1\}$ indicate whether the event asked by the question happens or not, and let $X_i \in \{0, 1\}$ indicate whether the

[1] https://www.metaculus.com
[2] https://manifold.markets
[3] https://openai.com
[4] https://www.mediacloud.org

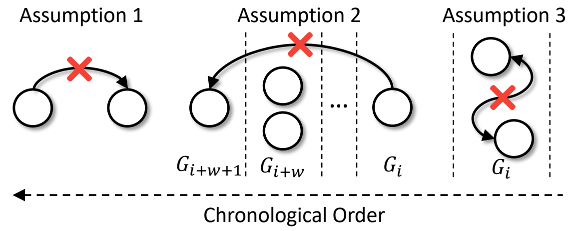

Figure 2: Illustration of assumptions. Nodes represent news variables that are in chronological order corresponding to their $\mathcal{T}$.

situation described in the i-th news happens or not. Each variable $\mathcal{X}_i$ is associated with a date $\mathcal{T}_i$ since each news also has the occurrence date. We use the notation $\mathcal{T}_i \prec \mathcal{T}_j$ to represent that the occurrence of the i^{th} news is before that of the j^{th}. Note that the date of $\mathcal{Y}$ is later than any date of $\mathcal{X}$.

Intuitively, if the i^{th} news article's occurrence $(X_i = 1)$ constitutes a necessary condition for $Y = \hat{Y}$ (ground-truth answer), then the intervention $\mathrm{do}(X_i = 0)$ would significantly increase the probability of $Y \neq \hat{Y}$. With this intuition, we define the CIL of the i^{th} news article as:

$$\begin{aligned}\texttt{CIL}_i =& P(\mathcal{Y} = \hat{\mathcal{Y}}|do(\mathcal{X}_i = 1)) \\ &- P(\mathcal{Y} = \hat{\mathcal{Y}}|do(\mathcal{X}_i = 0)),\end{aligned} \tag{3}$$

where *do* is the intervention operation in causal inference standing for $\mathcal{X}$ is intervened to happen or not as stated in Section 3.2.

To compute this estimation, we model all $\mathcal{X}_i$ and $\mathcal{Y}$ as a structural causal model (SCM). For this SCM, we treat all $\mathcal{X}_i$ and $\mathcal{Y}$ as nodes and causal relationships between them as edges. However, it is extremely hard to extract causal edges in our case due to incomplete knowledge base and intensive dependency on experts. It is difficult to calculate CIL via methods relying on the complete SCM.

To fill this gap, we introduce three assumptions. We illustrate these assumptions in Figure 2. Firstly, the causal relations between the news should be aligned with temporality. This assumption is consistent with common sense and eliminates circle paths in the SCM. Notice that $\mathcal{Y}$ is the variable in this SCM with the latest date.

Assumption 1. Temporality *For any two occurrences of news, the one that occurs later in date cannot have an effect on the one earlier:*

$$\begin{aligned}&\forall\, i, j, \quad if\ \ \mathcal{T}_i \prec \mathcal{T}_j, \\ &\qquad then\ \ P(\mathcal{X}_i|\mathcal{X}_j) = P(\mathcal{X}_i).\end{aligned} \tag{4}$$

Second, causal relationships between events that are widely separated in time should be mediated by events that occur between them. We group all the news in chronological order, with a group size representing 10 days passing. $\mathrm{G}(\mathcal{X}_i)$ stands for the index of the group in which $\mathcal{X}_i$ is in. In our case, if $\mathrm{G}(\mathcal{X}_i) < \mathrm{G}(\mathcal{X}_j)$ indicates $\mathcal{T}_i \prec \mathcal{T}_j$, namely the i^{th} news happens before the j^{th}.

# News	# Token	Max TS	Mean TS
100	853.95	31	16.54

Table 1: Statistics of the grounding news. TS stands for time span between the oldest and latest news of a question. The unit is a month.

Assumption 2. w-window Dependency *Variables in the i^{th} group can only be directly influenced by variables within the previous w groups (i.e., groups $i-1, i-2, \ldots, i-w$). Consequently, there exist no direct edges between $\mathcal{X}_i$ and $\mathcal{X}_j$ for any j outside this window:*

$$\begin{aligned}&\forall\, i, j, \quad if\ \ \mathrm{G}(\mathcal{X}_j) - \mathrm{G}(\mathcal{X}_i) > w, \\ &\qquad then\ \ (\mathcal{X}_i, \mathcal{X}_j) \notin \text{edges of SCM}.\end{aligned} \tag{5}$$

Lastly, news in the same group should have no causal relation in between.

Assumption 3. Concurrent Independency *Any two pieces of news that occurred in the same group are independent:*

$$\begin{aligned}&\forall\, i, j, \quad if\ \ \mathrm{G}(\mathcal{X}_j) = \mathrm{G}(\mathcal{X}_i), \\ &\qquad then\ \ (\mathcal{X}_i, \mathcal{X}_j) \notin \text{edges of SCM}.\end{aligned} \tag{6}$$

With these assumptions, we can derive CIL estimation. We show the calculation of $P(\mathcal{Y} = \hat{\mathcal{Y}}|do(\mathcal{X}_i = 1))$, then $P(\mathcal{Y} = \hat{\mathcal{Y}}|do(\mathcal{X}_i = 0))$ can be computed similarly.

Proposition. *The intervened probability $P(\mathcal{Y} = \hat{\mathcal{Y}}|do(\mathcal{X}_i = 1))$ can be convert into observation probability:*

$$\begin{aligned}&P(\mathcal{Y} = \hat{\mathcal{Y}}|do(\mathcal{X}_i = 1)) \doteq P_m(\mathcal{Y}|\mathcal{X}_i = 1) \\ &= \sum_{n_1,} \cdots P(\mathcal{Y} = \hat{\mathcal{Y}}|\mathcal{X}_i = 1, \mathcal{X}_{n_1}, \cdots) P(\mathcal{X}_{n_1}, \cdots) \\ &\ 0 < \mathrm{G}(\mathcal{X}_i) - \mathrm{G}(\mathcal{X}_{n_j}) \leqq w, \forall n_j.\end{aligned} \tag{7}$$

We leave the proof in the Appendix A.1. The remaining things are to compute $P(\mathcal{Y} = \hat{\mathcal{Y}}|\mathcal{X}_i = 1, \mathcal{X}_{n_1}, \cdots)$ and $P(\mathcal{X}_{n_1}, \cdots)$. Enlightened by Bynum and Cho (2024), we use LLMs to calculate the probabilities. For $P(\mathcal{Y} = \hat{\mathcal{Y}}|\mathcal{X}_i = 1, \mathcal{X}_{n_1}, \cdots)$, note that all $\mathcal{X}_{n_1}$ have two possible values, namely 0 or 1. We need to sum over all the permutations. We take $P(\mathcal{Y} = \hat{\mathcal{Y}}|\mathcal{X}_i = 1, \mathcal{X}_{n_1} = 1, \mathcal{X}_{N-2} = 0)$ as an example, and derive the prompt from Halawi et al. (2024). We show the prompts in the Appendix A.7 (e). Similar to $P(\mathcal{X}_{n_1}, \cdots)$, we take

Models	Retrieval	Reasoning	L1		L2	
			Brier Score ↓	CIL ↑	Brier Score ↓	CIL ↑
GPT-4o	w.o. RAG	ScratchPAD	25.42 ± 0.09	-	23.09 ± 1.38	-
	Naive RAG		21.22 ± 0.30 (+4.20)	0.07 ± 0.00	22.79 ± 0.64 (+0.30)	-4.60 ± 0.00
	APP		20.02 ± 0.26 (+5.40)	1.47 ± 0.16	24.25 ± 0.69 (-1.16)	-4.68 ± 0.21
Claude	w.o. RAG	ScratchPAD	26.19 ± 1.31	-	26.09 ± 0.17	-
	Naive RAG		23.46 ± 0.85 (+2.73)	0.07 ± 0.00	24.93 ± 0.20 (+1.16)	-4.60 ± 0.00
	APP		22.75 ± 0.96 (+3.44)	1.53 ± 0.02	28.16 ± 0.17 (-2.07)	-4.69 ± 0.01
Gemini	w.o. RAG	ScratchPAD	25.39 ± 0.41	-	20.82 ± 0.01	-
	Naive RAG		22.18 ± 0.39 (+3.21)	0.07 ± 0.00	23.25 ± 0.29 (-2.43)	-4.60 ± 0.00
	APP		19.78 ± 0.24 (+5.61)	1.66 ± 0.09	26.07 ± 0.05 (-5.24)	-4.95 ± 0.04

Table 2: Validation of CIL estimation. Retrieval number $N = 10$. We report mean and std values on twice runs.

$P(\mathcal{X}_{n_1} = 1, \mathcal{X}_{n_2} = 0,)$ for example and use the prompt in the Appendix A.7 (f) to compute. We use window size $w = 3$.

Note that LLMs cannot be used to calculate the intervened probability directly since they are trained to be a world model with observation probability (Bynum and Cho, 2024). We now finish calculating CIL each news article by Equation 3.

4.3 Construction

After calculating the CIL for all pieces of news, we construct the benchmark with them. For each question, we count the number of pieces of news where their CIL are above a threshold. If the number is large enough, we add the question to the chosen set *L1*, otherwise to *L2*. We consider *L1* to be the main part of our benchmark because answering the questions can be sufficient supported by *L1*. It can serve as an RAG benchmark. While *L2* lacks sufficient support to answer the questions, it also provides valuable information for prediction questions, but needs to be supplemented with additional information beyond the news. We currently create 99 questions for *L1* and 53 for *L2*. We make several discussions about our benchmark:

Data volume. There is not a large volume of valuable prediction questions in total. To ensure the validity of PROPHET, we apply filtering operations during construction by CIL estimation. As a result, the volume of PROPHET is smaller than that of datasets where data collection without inferability validation. This is also the case for other future forecasting datasets with question filtering (Karger et al., 2024). We'll address this issue by using automatic pipelines to regularly collect and add new questions to update the benchmark.

Causality Assumptions. Our assumptions are rooted in general commonsense and aim to capture the dominant patterns in news-event relationships. We don't attempt to model global causality; instead, it suffices to model the causality required for the task with appropriate parameters.

Probability Computing. In pilot experiments, different LLMs provided slightly different scores when computing probabilities in CIL. Thus, we use a single LLM multiple times for reliable estimation. Later experiments showed that CIL is model-agnostic: different models reach the same conclusions, validating this estimation method.

4.4 Statistics and Properties of PROPHET

We do basic statistics of PROPHET. Assisted by CIL, we also explore key properties of future forecasting task and the benchmark. We currently harvest 99 data in L1 and 53 data in L2. The statistics of news articles we crawled are shown in Table 1. During the construction process, we only discard obviously irrelevant news. Therefore, we did not significantly alter the data distribution of the valid news. News we remain can reflect the real distribution of situations about certain queried events.

We retain 100 top relevant news for each question. The average news tokens are 853.95 leading to context problem if a method longs for simply adding all news in the prompt. We calculate the time span between the oldest and the newest news. The average time span is 16.54 months which is large enough for the method to retrieve similar events in the history for answering.

We conduct in-depth analysis and draw findings of PROPHET assisted by CIL:1) As the resolved date approaches, both high and low CIL news articles increase. It poses a challenge for models to resist forecasting bias. 2) Two main volume distributions of news articles were identified: one with few articles early on and a sudden surge near the end, and another with a uniform distribution over time. We leave details in the Appendix A.3.

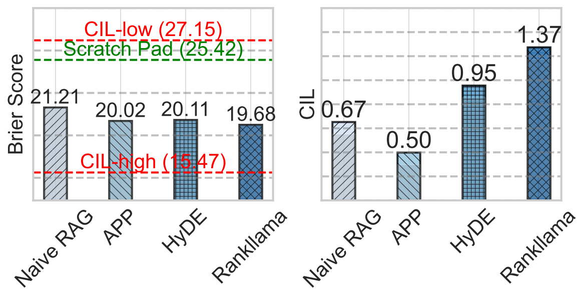

Figure 3: Retrieval evaluation.

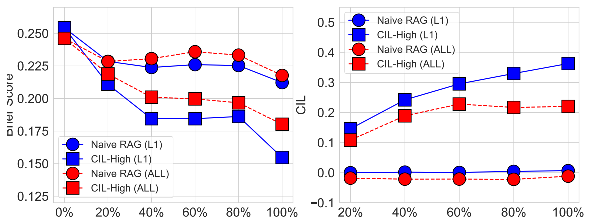

Figure 4: Temporal analysis. The horizontal axis represents the entire prediction process.

5 Experiments

We first conduct experiments to show the validity of CIL estimation and our benchmark in Section 5.2. Then we evaluate the current retrieval and reasoning baselines on PROPHET in Section 5.3. Lastly, assisted by CIL, we conduct a temporal analysis on PROPHET to provide insights into future forecasting systems in Section 5.4. We use **the cases to show the effectiveness of CIL** in the Appendix A.6.

5.1 Evaluated Methods

For retrieval methods, we evaluate Naive RAG, APP (Halawi et al., 2024), Rankllama (Ma et al., 2024), HyDE (Gao et al., 2023). For reasoning methods, we include ScrathPAD (Halawi et al., 2024), CoT (Wei et al., 2022), Long-CoT (OpenAI, 2024). Details are in the Appendix A.4. Since the news would be long, we pre-summarize each news and all methods use the same summarization in RAG.

5.2 Validity of CIL and PROPHET

To validate the estimation of CIL, we conduct branches of experiments. We test numerous methods and LLMs on both L1 and L2 parts of data. The results are shown in Figure 2. To ensure comparability, all methods are on ScratchPAD reasoning prompting. Native RAG and APP are two RAG methods. We also report the differences between w.o. RAG and each RAG method.

As shown, all RAG methods applied to various LLMs perform better than w.o. RAG on L1 while showing little or no improvement on L2. These results strongly suggest that CIL estimation is effective in identifying inferable data. It can measure the supportiveness of news articles. Questions lacking supportive rationales are difficult to accurately forecast. In addition, the results also show CIL estimation is **model-agnostic**. Although we use GPT-4o to calculate CIL, all models are subjected to these data partitions by CIL. That demonstrates the nature of the intervened causality captured by this robust estimation. Last, we also notice that, in some methods or LLMs, it drops compared to w.o. RAG. It indicates some articles would contribute negatively in prediction. This is consistent with the findings in Section 4.4. Our CIL score is able to measure the negative effects of the news articles.

5.3 Performances on Future Forecasting

In this section, we evaluate current methods in our future forecasting benchmark. We evaluate two branches of methods representing two core abilities of this task, retrieval and reasoning.

5.3.1 Retrieval Performances

We compare between Naive RAG, APP, HyDE, and Rankllama as retrieval evaluation. For all methods, we retrieve 10 news articles and use ScratchPAD reasoning on GPT-4o. We also compare these methods to CIL-high[5] and CIL-low where we directly use the news articles with the highest and lowest CIL scores. The results are in Figure 3.

CIL-high performs the best while CIL-low is the worst. This further demonstrates the validity of CIL estimation. Among other methods, Rankllama performs the best in Brier Score and improves on CIL score. Rankllama can understand the complicated instructions indicating that it requires deep comprehension of retrieval queries for news. This provides insights that training retrieval methods for complicated query instructions are crucial in such RAG task with hidden rationales.

In total, compared to the CIL-high, all methods still have a significant gap on CIL and Brier Score, indicating that there is still much room for improvement in this retrieval task. It requires delicate approaches that excel in real-world knowledge grounding and comprehension.

5.3.2 Reasoning Performances

In this section, we evaluate three reasoning methods on PROPHET:ScratchPad, CoT, and Long-CoT. We use various models and test under two retrieval conditions: (1) using news articles with top

[5]Note that CIL-high and CIL-low are not actual methods, they are only empirical methods for studying the performance bounds.

Reasoning	Model	$N = 5$		$N = 10$	
		CIL-High	Naive RAG	CIL-High	Naive RAG
ScratchPad	GPT-4o	17.02 ± 0.46	$\mathbf{21.53 \pm 0.35}$	16.03 ± 0.21	$\mathbf{21.22 \pm 0.30}$
	GPT-4o-mini	19.37 ± 0.31	23.66 ± 0.24	18.37 ± 0.67	24.03 ± 0.57
	Claude-3-5-sonnet	20.03 ± 0.17	24.64 ± 1.16	15.82 ± 0.53	23.46 ± 0.85
	Gemini-1.5-pro	16.89 ± 0.35	22.51 ± 0.19	17.69 ± 0.54	22.18 ± 0.39
	Qwen2.5-32B	21.38 ± 1.30	25.10 ± 0.70	20.74 ± 1.51	23.89 ± 0.26
	Qwen2.5-7B	26.17 ± 0.69	30.93 ± 1.36	24.86 ± 0.35	26.64 ± 0.76
CoT	GPT-4o	16.70 ± 1.15	22.04 ± 0.37	15.60 ± 0.25	23.75 ± 0.25
	Gemini-1.5-pro	17.68 ± 0.13	26.45 ± 2.87	15.57 ± 1.77	25.34 ± 1.14
	Qwen2.5-32B	17.90 ± 2.51	22.29 ± 0.16	15.89 ± 3.45	26.38 ± 0.72
	Qwen2.5-7B	23.04 ± 1.87	33.13 ± 3.60	23.27 ± 0.42	34.82 ± 1.33
Long-CoT	O1-mini	$\mathbf{15.66 \pm 1.14}$	23.49 ± 2.94	$\mathbf{13.72 \pm 0.38}$	24.19 ± 0.65

Table 3: Reasoning evaluation. We report mean and std values on twice runs.

CIL scores, and (2) using Naive RAG. We also compare retrieval sizes (N=5 vs. N=10). Results are shown in Table3. Key findings include:

① Long-CoT achieves the best results across all methods and models, highlighting its potential for future forecasting tasks. This suggests that event prediction relies heavily on deep, multi-step reasoning based on available information. Specialized post-training in forecasting reasoning is crucial for improving performance.

② Effective information retrieval is fundamental for reasoning. Under Naive RAG, methods show significantly lower performance gains compared to CIL-High. Moreover, models and methods exhibit minimal differences in Naive RAG, while clear distinctions emerge inCIL-High. This underscores the importance of retrieval quality for reasoning. More sophisticated retrieval and reasoning techniques could enhance performance.

③ ScratchPad outperforms CoT under Naive RAG, but the reverse is true for CIL-High. This finding, not previously reported (Halawi et al., 2024), suggests that texttttScratchPad constrains the model's reasoning when useful information is scarce leading to improvements. However, when information is abundant, it may limit the model's reasoning ability. This insight offers potential for developing advanced reasoning methods.

5.4 Temporal Studies

Future forecasting is a continuous process that begins when the question is posed and ends when the question is answered. The earlier the answer can be predicted, the more valuable it is. We investigate the system's forecasting at different times. Similar as in Section 4.4, we compute the progress in the whole forecasting. We represent the progress of each news by the percentage of its date in the forecasting. We show the performances of Naive RAG and CIL-High at different times. These experiments are on both L1 part and the whole benchmarks (L1+L2). (L1+L2) is the real-world forecasting scenario. All results are on GPT-4o and ScratchPAD reasoning. The results are in Figure 4.

① We find significant potential in the early-time future forecasting. The CIL-High at 20% progress performs even better than Naive RAG at 100%. It indicates that if we have a sufficiently powerful retrieval method, we can expect to achieve effective predictions at the early stages of event development. This finding applies to both scenarios where evidence is sufficient and where it is insufficient.

② When the forecasting progress precedes, there would be news that is harmful for prediction. We find that during the progress of forecasting, the performances of some methods fluctuate. And the CIL of the Naive RAG stops increasing at 60%. This is consistent with the conclusions in Section 4.4. It shows a desired prediction system should be aware of negative evidence and can self-correct in the retrieval and reasoning process.

6 Conclusion

We address the challenge of building the inferable RAG benchmark for evaluating future forecasting systems by introducing PROPHET. It is rigorously validated for inferability by our Causal Intervened Likelihood (CIL) estimation. By leveraging causal inference to quantify the inferability of prediction questions based on their associated news articles, PROPHET ensures that questions are answerable through retrieved rationales, thereby providing a more accurate assessment of the model capabilities. Experimental validation confirms the effectiveness of CIL in correlating with system performance, while evaluations of state-of-the-art systems on PROPHET reveal key strengths and limitations, particularly in retrieval and reasoning. This work establishes a basis for the development of more nuanced models. With ongoing updating, PROPHET ensures the inferable evaluation in driving progress towards AI-powered forecasting.

Limitations

In this work, we evaluate methods of retrieval and reasoning disentangling. However, entangled methods could further improve future forecasting. We leave it to future work.

Ethics Statement

This dataset is strictly for non-commercial research purposes under the following conditions: 1) Restricted Application Scope: All narrative scenarios contained herein are intended solely for academic exploration of future forecasting methodologies. Any utilization for purposes involving defamation, harassment, malicious targeting, or other unethical practices is expressly prohibited. 2) Prohibited Misinterpretation: Statistical patterns derived from this resource should not be interpreted as deterministic predictions of real-world events. 3) Accountability Framework: The creators explicitly disclaim liability for consequences arising from dataset misuse, including but not limited to algorithmic bias propagation, privacy infringements, or sociotechnical harms caused by improper application.

References

Nitin Aravind Birur, Tanay Baswa, Divyanshu Kumar, Jatan Loya, Sahil Agarwal, and Prashanth Harshangi. 2024. Vera: Validation and enhancement for retrieval augmented systems. *arXiv preprint arXiv:2409.15364*.

Glenn W Brier. 1950. Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1):1–3.

Lucius EJ Bynum and Kyunghyun Cho. 2024. Language models as causal effect generators. *arXiv preprint arXiv:2411.08019*.

Li Du, Xiao Ding, Kai Xiong, Ting Liu, and Bing Qin. 2022. e-care: a new dataset for exploring explainable causal reasoning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 432–446.

Robert Friel, Masha Belyi, and Atindriyo Sanyal. 2024. Ragbench: Explainable benchmark for retrieval-augmented generation systems. *arXiv preprint arXiv:2407.11005*.

Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. 2023. Precise zero-shot dense retrieval without relevance labels. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1762–1777, Toronto, Canada. Association for Computational Linguistics.

Mor Geva, Daniel Khashabi, Tushar Khot, Ashish Sabharwal, and Dan Roth. 2021. Strategyqa: A question answering benchmark requiring strategy and planning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5835–5847, Online. Association for Computational Linguistics.

Granroth-Wilding. 2016. What happens next? event prediction using a compositional neural network model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30.

Yong Guan, Hao Peng, Xiaozhi Wang, Lei Hou, and Juanzi Li. 2024. Openep: Open-ended future event prediction. *arXiv preprint arXiv:2408.06578*.

Danny Halawi, Fred Zhang, Chen Yueh-Han, and Jacob Steinhardt. 2024. Approaching human-level forecasting with language models. *arXiv preprint arXiv:2402.18563*.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *Transactions of the Association for Computational Linguistics*, 9:479–498.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. 2020. Constructing a multi-hop qa dataset for comprehensive evaluation of reasoning steps. *arXiv preprint arXiv:2011.01060*.

Elvis Hsieh, Preston Fu, and Jonathan Chen. 2024. Reasoning and tools for human-level forecasting. *arXiv preprint arXiv:2408.12036*.

Ezra Karger, Houtan Bastani, Chen Yueh-Han, Zachary Jacobs, Danny Halawi, Fred Zhang, and Philip E Tetlock. 2024. Forecastbench: A dynamic benchmark of ai forecasting capabilities. *arXiv preprint arXiv:2409.19839*.

Kale Leetaru and Philip A Schrodt. 2013. Gdelt: Global data on events, location, and tone, 1979-2012. *The GDELT Project*.

Xiang Li, Zhenyu Li, Chen Shi, Yong Xu, Qing Du, Mingkui Tan, and Jun Huang. 2024. AlphaFin: Benchmarking financial analysis with retrieval-augmented stock-chain framework. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 773–783, Torino, Italia. ELRA and ICCL.

Chin-Yew Lin, Xi Victoria Lin, and Jimmy Lin. 2020. 2wikimultihopqa: A dataset for multi-hop question answering on wikipedia. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7380–7391, Online. Association for Computational Linguistics.

Nianzu Liu, Tianyi Zhang, and Percy Liang. 2024. Benchmarking large language models in retrieval-augmented generation. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*, pages 17754–17762, Washington, DC, USA. AAAI Press.

Xueguang Ma, Liang Wang, Nan Yang, Furu Wei, and Jimmy Lin. 2024. Fine-tuning llama for multi-stage text retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2421–2425.

Rounak Meyur, Hung Phan, Sridevi Wagle, Jan Strube, Mahantesh Halappanavar, Sameera Horawalavithana, Anurag Acharya, and Sai Munikoti. 2024. Weqa: A benchmark for retrieval augmented generation in wind energy domain. *arXiv preprint arXiv:2408.11800*.

Nasrin Mostafazadeh, Michael Roth, Annie Louis, Nathanael Chambers, and James Allen. 2017. Lsdsem 2017 shared task: The story cloze test. In *Proceedings of the 2nd Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics*, pages 46–51.

OpenAI. 2024. Openai o1: Reinforcement learning with chain-of-thought reasoning. Technical report, OpenAI.

Judea Pearl. 2010. An introduction to causal inference. *The international journal of biostatistics*, 6(2).

Sarah Pratt, Seth Blumberg, Pietro Kreitlon Carolino, and Meredith Ringel Morris. 2024. Can language models use forecasting strategies? *arXiv preprint arXiv:2406.04446*.

Victor Rotaru, Yi Huang, Timmy Li, James Evans, and Ishanu Chattopadhyay. 2022. Event-level prediction of urban crime reveals a signature of enforcement bias in us cities. *Nature human behaviour*, 6(8):1056–1068.

Philip A Schrodt, David J Gerner, Peter W Foltz, Moon-Soo Cho, and Young Joon Park. 2012. The integrated crisis early warning system (icews). *Conflict Management and Peace Science*, 29(4):432–450.

Ivan Stelmakh, Yi Luan, Bhuwan Dhingra, and Ming-Wei Chang. 2022. Asqa: Factoid questions meet long-form answers. *arXiv preprint*.

Alessandro Stolfo, Zhijing Jin, Kumar Shridhar, Bernhard Schoelkopf, and Mrinmaya Sachan. 2023. A causal framework to quantify the robustness of mathematical reasoning with language models. In *The 61st Annual Meeting Of The Association For Computational Linguistics*.

Yixuan Tang and Yi Yang. 2024. MTRAG: A multi-turn conversational benchmark for evaluating retrieval-augmented generation systems. *arXiv preprint*.

Zhengwei Tao, Zhi Jin, Yifan Zhang, Xiancai Chen, Haiyan Zhao, Jia Li, Bing Liang, Chongyang Tao, Qun Liu, and Kam-Fai Wong. 2024. A comprehensive evaluation on event reasoning of large language models. *arXiv preprint arXiv:2404.17513*.

Steven H. Wang, Antoine Scardigli, Leonard Tang, Wei Chen, Dimitry Levkin, Anya Chen, Spencer Ball, Thomas Woodside, Oliver Zhang, and Dan Hendrycks. 2024a. CRAG: Corrective retrieval-augmented generation for robust knowledge grounding. *arXiv preprint*.

Steven H. Wang, Antoine Scardigli, Leonard Tang, Wei Chen, Dimitry Levkin, Anya Chen, Spencer Ball, Thomas Woodside, Oliver Zhang, and Dan Hendrycks. 2024b. LegalBench-RAG: A domain-specific benchmark for evaluating retrieval in legal rag systems. *arXiv preprint*.

Yang Wang and Hassan A Karimi. 2024. Exploring large language models for climate forecasting. *arXiv preprint arXiv:2411.13724*.

Jason Wei, Andrew Zou, Denny Zhou, Hattie Kim, Tianyi Chen, and Quoc V. Le. 2022. Chain-of-thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. *arXiv preprint arXiv:1809.09600*.

Siyu Yuan, Jiangjie Chen, Ziquan Fu, Xuyang Ge, Soham Shah, Charles Jankowski, Yanghua Xiao, and Deqing Yang. 2023. Distilling script knowledge from large language models for constrained language planning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4303–4325.

Yuxuan Zhang, Zhiyuan Zhang, Yicheng Wang, Yuxuan Su, Yixuan Su, Yixuan Su, Yixuan Su, Yixuan Su, Yixuan Su, Yixuan Su, Yixuan Su, Yixuan Su, Yixuan Su, and Yixuan Su. 2023. Pubhealth: A benchmark for public health question answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9802–9822, Toronto, Canada. Association for Computational Linguistics.

Siyun Zhao, Yuqing Yang, Zilong Wang, Zhiyuan He, Luna K Qiu, and Lili Qiu. 2024. Retrieval augmented generation (rag) and beyond: A comprehensive survey on how to make your llms use external data more wisely. *arXiv preprint arXiv:2409.14924*.

Chang Zong, Yuchen Yan, Weiming Lu, Jian Shao, Eliot Huang, Heng Chang, and Yueting Zhuang. 2024. Triad: A framework leveraging a multi-role llm-based agent to solve knowledge base question answering. *Preprint*, arXiv:2402.14320.

A Appendix

A.1 Proof of Proposition

We show the proof of Proposition Eq.(7) below. This proof is mainly based on causal inference theory and our assumptions shown in Figure 2.

Proof. By the law of total probability,

$$\begin{aligned}
&P(\mathcal{Y}=\hat{\mathcal{Y}}|do(\mathcal{X}_i=1)) \doteq P_m(\mathcal{Y}|\mathcal{X}_i=1)\\
=&\sum_{n_1,}\cdots\sum_{m_1}\cdots\\
&P_m(\mathcal{Y}=\hat{\mathcal{Y}}|\mathcal{X}_i=1,\mathcal{X}_{n_1},\cdots,\mathcal{X}_{m_1},\cdots)\\
&\times P_m(\mathcal{X}_{n_1},\cdots,\mathcal{X}_{m_1},\cdots|\mathcal{X}_i=1)\\
&0<\forall n_j, \mathrm{G}(\mathcal{X}_i)-\mathrm{G}(\mathcal{X}_{n_j})\leqq w,\\
&\forall m_j, \mathrm{G}(\mathcal{X}_i)-\mathrm{G}(\mathcal{X}_{m_j})>w.
\end{aligned} \tag{8}$$

Since the $\mathcal{Y}$ is the latest variable and happened w window later than $\mathcal{X}_i$, with Assumption 2, we have

$$\begin{aligned}
&P_m(\mathcal{Y}=\hat{\mathcal{Y}}|\mathcal{X}_i=1,\mathcal{X}_{n_1},\cdots,\mathcal{X}_{m_1},\cdots)\\
=&P_m(\mathcal{Y}=\hat{\mathcal{Y}}|\mathcal{X}_i=1,\mathcal{X}_{n_1},\cdots),\\
&\times P_m(\mathcal{X}_{n_1},\cdots,\mathcal{X}_{m_1},\cdots|\mathcal{X}_i=1)\\
=&P_m(\mathcal{X}_{n_1},\cdots|\mathcal{X}_i=1,\mathcal{X}_{m_1},\cdots)\\
&\times P(\mathcal{X}_{m_1},\cdots|\mathcal{X}_i=1)\\
=&P_m(\mathcal{X}_{n_1},\cdots|\mathcal{X}_i=1)P(\mathcal{X}_{m_1},\cdots)\\
&0<\forall n_j, \mathrm{G}(\mathcal{X}_i)-\mathrm{G}(\mathcal{X}_{n_j})\leqq w,\\
&\forall m_j, \mathrm{G}(\mathcal{X}_i)-\mathrm{G}(\mathcal{X}_{m_j})>w.
\end{aligned} \tag{9}$$

Then take Equation (9) into Equation (8), and interchange the order of summation,

$$\begin{aligned}
&P(\mathcal{Y}=\hat{\mathcal{Y}}|do(\mathcal{X}_i=1)) \doteq P_m(\mathcal{Y}|\mathcal{X}_i=1)\\
=&\sum_{n_1,}\cdots P_m(\mathcal{Y}=\hat{\mathcal{Y}}|\mathcal{X}_i=1,\mathcal{X}_{n_1},\cdots)\\
&\times P_m(\mathcal{X}_{n_1},\cdots|\mathcal{X}_i=1)\sum_{m_1}\cdots P(\mathcal{X}_{m_1},\cdots)\\
=&\sum_{n_1,}\cdots P_m(\mathcal{Y}=\hat{\mathcal{Y}}|\mathcal{X}_i=1,\mathcal{X}_{n_1},\cdots)\\
&\times P_m(\mathcal{X}_{n_1},\cdots|\mathcal{X}_i=1)\\
&0<\forall n_j, \mathrm{G}(\mathcal{X}_i)-\mathrm{G}(\mathcal{X}_{n_j})\leqq w,\\
&\forall m_j, \mathrm{G}(\mathcal{X}_i)-\mathrm{G}(\mathcal{X}_{m_j})>w.
\end{aligned} \tag{10}$$

Under the do operation, $\mathcal{X}_i$ is independent to $\mathcal{X}_{n_j}, \forall n_j$. Owing to Assumptions 1 and 3, the concurrent and later variables don't influence $\mathcal{X}_i$. Therefore, the intervened distribution equals to origin distribution.

$$\begin{aligned}
& P(\mathcal{Y}=\hat{\mathcal{Y}}|do(\mathcal{X}_i=1)) \doteq P_m(\mathcal{Y}|\mathcal{X}_i=1) \\
= & \sum_{n_1,} \cdots P_m(\mathcal{Y}=\hat{\mathcal{Y}}|\mathcal{X}_i=1,\mathcal{X}_{n_1},\cdots)P_m(\mathcal{X}_{n_1},\cdots) \\
= & \sum_{n_1,} \cdots P(\mathcal{Y}=\hat{\mathcal{Y}}|\mathcal{X}_i=1,\mathcal{X}_{n_1},\cdots)P(\mathcal{X}_{n_1},\cdots) \\
& \forall n_j, 0 < \mathrm{G}(\mathcal{X}_i) - \mathrm{G}(\mathcal{X}_{n_j}) \leqq w.
\end{aligned} \tag{11}$$

□

A.2 Construction Details

During constructing, we use `gpt-4o-mini-2024-07-18` for all LLM callings. We set window size w to 3 which is enough large in our pilot study. For computing each probability in CIL, we call twice `gpt-4o-mini-2024-07-18` and get the average score. The constructing prompts we use are shown in prompts (a-f).

A.3 Future Forecasting Analysis Assisted by CIL

We calculate the distribution of the CIL metric and the number of news articles over time. We regard the time span between the oldest news and the resolved date as the whole progress of a question. Then we compute the progress of each news by the percentage of its date in this progress. The results are in Figure 5. We explore some key properties of future forecasting based on these studies.

① As the approaching the resolved date, both news of high and low CIL increase. News of high CIL increase is consistent with human intuition. As time progresses, the prediction of future events will become more certain. However, we also find low CIL news increases indicating that as time progresses, there will also be an increase in the generation of misleading information. It challenges the model to resist this bias for precise predicting.

② We mainly discovery two volume distributions of news articles. The first type of distribution is characterized by a very low number of news articles early on, with a sudden surge close to the end time. The second type of distribution is characterized by a uniform distribution of news over time. This reflects two ways in which people pay attention to events. However, the first type brings difficulties for early prediction since it lack valid information at an early date.

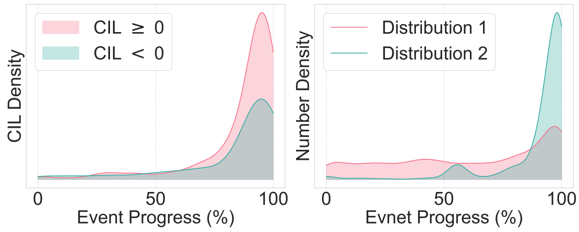

Figure 5: In-depth analysis. The horizontal axis represents the entire prediction process.

A.4 Evaluated Methods

We introduce the methods that we evaluate in this work. For the retrieval methods:

`Naive RAG`: Since the news articles are long, we first summarize the news articles in advance. This RAG method then retrieves relevant news articles via embedding similarity between the question and news summary. We use all-MiniLM-L6-v2 models in SentenceTransformer[6]. After retrieving the news, we use the scratchpad prompt for reasoning.

`APP`: This is the method introduced by Halawi et al. (2024). It also first summarizes the news articles. Then it uses LLM to compute the relevance score. After that, it also uses scratchpad prompt for reasoning.

`Rankllama`: This is a retrieval method where it can understand the complicated retrieval instructions (Ma et al., 2024). It uses the model to encode the question and the news articles. We use summaries of the news. After retrieval, it answers in scratchpad prompt as well.

`HyDE`: Given a query, this method uses an instruction-following language model (e.g., InstructGPT) to generate a "hypothetical document" that captures relevance patterns (Gao et al., 2023). In event prediction scenario, we generate potential future events that could effect the answer. Then retrieve relevant news articles.

The reasoning methods are:

`ScrathPAD`: This is the zero-shot ScrathPAD prompting method based on LLMs. We use the scratchpad prompt introduced by Halawi et al. (2024).

`CoT`: Chain of Thought is a technique that enables AI models to mimic human-like step-by-step reasoning by breaking down complex problems into intermediate logical steps, significantly improving interpretability and accuracy in tasks such as mathematical reasoning and NLP (Wei et al., 2022).

[6]https://sbert.net

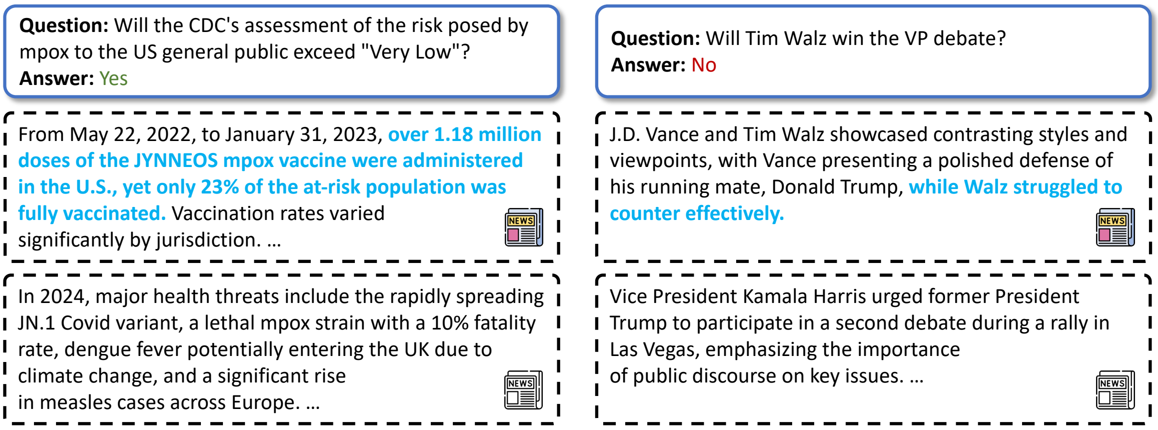

Figure 6: Case studies.

`Long-CoT`: Long-CoT is on LLMs trained with reinforcement learning to perform advanced reasoning through internal CoT such as OpenAI-O1 (OpenAI, 2024), achieving state-of-the-art performance in competitive programming, mathematics, and scientific benchmarks, even surpassing human experts in some domains.

Type	Benchmark	W	G	R	I
Script Learning	MCNC (Granroth-Wilding, 2016)	✗	✗	✓	-
	SCT (Mostafazadeh et al., 2017)	✗	✗	✓	-
	CoScript (Yuan et al., 2023)	✗	✗	✓	-
Time Series	GDELT (Leetaru and Schrodt, 2013)	✓	✗	✓	-
	ICEWS (Schrodt et al., 2012)	✓	✗	✓	-
Event Reasoning	ECARE (Du et al., 2022)	✓	✗	✓	-
	EV2 (Tao et al., 2024)	✗	✗	✓	-
Open Event Prediction	Halawi et al. (2024)	✓	✓	✗	✗
	OpenEP (Guan et al., 2024)	✓	✓	✗	✗
	ForecastBench (Karger et al., 2024)	✓	✓	✗	✗
	`PROPHET` (Ours)	✓	✓	✓	✓

Table 4: Comparison with other forecasting benchmarks. `W`: real-world questions. `G`: News Grounded. `R`: reproductive. `I`: inferable validation.

A.5 Evaluation Details

All experiments in this work are under twice runs. We report the mean and std values. We list the versions of LLMs we use in Table 5. The reasoning prompts are in prompts (g-h).

Model	Version
GPT-4o	gpt-4o-2024-08-06
GPT-4o-mini	gpt-4o-mini-2024-07-18
O1-mini	o1-mini-2024-09-12
Claude	claude-3-5-sonnet-20240620
Gemini	gemini-1.5-pro-latest
Qwen2.5-32B	Qwen2.5-32B-Instruct-GPTQ-Int4
Qwen2.5-7B	Qwen2.5-7B-Instruct-GPTQ-Int4

Table 5: Evaluated model versions.

A.6 Case of `CIL`

In this section, we showcase articles of high and low `CIL` scores. In Figure 6 we illustrate two questions. Each question is paired with `CIL-High` and `CIL-Low` articles. We find our `CIL` estimation precisely captures supportiveness for answering the question. For example, the first question asks the CDC's reaction to mpox. The `CIL-High` states the situation of vaccination of U.S. while the `CIL-Low` only mentions the global situation of mpox. Owing to the low vaccination rates of the U.S., it is likely that the CDC would pose the assessment of mpox exceeding "Very Low". In the second example, the `CIL-High` tells that Walz struggled to counter J.D. Vance effectively while `CIL-Low` merely mentions Kamala Harris wants to raise a debate. `CIL-High` contributes more to the correct answer.

A.7 Prompts

We list all prompts in the following Figures (a-h).

(a) Entity Query Generation

I will provide you with a forecasting question and the background information for the question. Extract the named entities, events of the question. Each entity and event are up to 5 words. The named entities can only be people, organizations, countries, locations while can not be date or time. Put all result items in a list that I can parse by JSON as ["entity 1", "entity 2", "event 1", "event 2", ...].
Question: $\mathcal{Q}$
Question Background: $\mathcal{B}$
Question Date: *date*
Output:

(b) Resolving Steps Query Generation

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. The articles should be mainly about event arguments such as subjects, objects, locations, organizations of the events in question and background information. You must generate this exact amount of queries: num keywords. Put all result items in a list that I can parse by JSON as ["step 1", "step 2", "step 3", ...].
Question: $\mathcal{Q}$
Question Background: $\mathcal{B}$
Question Date: *date*
Output:

(c) Similar Events Query Generation

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 of similar events on Google News to help answer the question. The similar events are events happened on other similar entities in the history. Or events happended on question entities but on other date. You must generate this exact amount of queries: num keywords. Put all result items in a list that I can parse by JSON as ["event 1", "event 2", "event 3", ...].
Question: $\mathcal{Q}$
Question Background: $\mathcal{B}$
Question Date: *date*
Output:

(d) News Article Relevance Rating

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: $\mathcal{Q}$
Question Background: $\mathcal{B}$
Resolution Criteria: $\mathcal{R}$
Article: *articles*
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:
- If the article has events of similar types which may happened on different subjects, it also consider relevant to the question.
- You don't need to access any external sources. Just consider the information provided.
- 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: Thoughts: { insert your thinking } Rating: { insert your rating }

(e) Conditional Probability

Given a background that in the meantime:
— These events happened: *news of* $\mathcal{X}_{n_1}$
— These events didn't happen: *news of* $\mathcal{X}_{n_2}$
Most importantly: — These events happened: *news of* $\mathcal{X}_i$
Answer the question: $\mathcal{Q}$
Instructions:
1. Provide at least 3 reasons why the answer might be no.
{ Insert your thoughts }
2. Provide at least 3 reasons why the answer might be yes.
{ Insert your thoughts }
3. Rate the strength of each of the reasons given in the last two responses. Think like a superforecaster (e.g. Nate Silver).
{ Insert your rating of the strength of each reason }
4. Aggregate your considerations.
{ Insert your aggregated considerations }
5. Output your answer (a number between 0 and 1) with an asterisk at the beginning and end of the decimal.
{ Insert your answer }"

(f) Probability

Given a situation that in the meantime:
— These events happened: news of $\mathcal{X}_{n_1}$
— These events didn't happen: news of $\mathcal{X}_{n_2}$
Instructions:
Use your world knowledge and commonsense to reason the probability if the situation can happen. Generate the thoughts first:
{ Insert your thoughts }
Then output your answer (a probability number between 0 and 1) with an asterisk at the beginning and end of the decimal.
{ Insert your answer }

(g) ScratchPAD

Question: $\mathcal{Q}$
Question Background: $\mathcal{B}$
Resolution Criteria: $\mathcal{R}$
We have retrieved the following information for this question: *retrieved articles*
Instructions:
1. Provide at least 3 reasons why the answer might be no.
{ Insert your thoughts }
2. Provide at least 3 reasons why the answer might be yes.
{ Insert your thoughts }
3. Rate the strength of each of the reasons given in the last two responses. Think like a superforecaster (e.g. Nate Silver).
{ Insert your rating of the strength of each reason }
4. Aggregate your considerations.
{ Insert your aggregated considerations }
5. Output your answer (a number between 0 and 1) with an asterisk at the beginning and end of the decimal.
{ Insert your answer }

(h) CoT and Long-CoT

Question: $\mathcal{Q}$
Question Background: $\mathcal{B}$
Resolution Criteria: $\mathcal{R}$
We have retrieved the following information for this question: retrieved articles
Think step by step. Reason and finally output your answer (a number between 0 and 1) with an asterisk at the beginning and end of the decimal.,