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

Anonymous ACL submission

Abstract

Predicting future events from Web news is one of the ultimate aspirations of artificial intelligence. Recent large language model (LLM)-based systems have shown remarkable potential and drawn significant interest from the research community. Several benchmarks evaluate forecasting capability by formalizing event prediction as a retrieval-augmented generation (RAG)-and-reasoning task, where each question is answered with relevant news articles retrieved from the Web. However, because none of them check whether the questions can be supported by valid or sufficient rationales, some questions in these benchmarks are inherently non-inferable. To address this, we introduce PROPHET, a benchmark of inferable forecasting questions paired with relevant news for retrieval. To ensure inferability, we propose Causal Intervened Likelihood (CIL), a statistical measure that assesses inferability via causal inference. We first collect recent trend forecasting questions, then filter them with CIL, yielding an inferable benchmark for future forecasting. Through extensive experiments, we demonstrate the validity of CIL, conduct in-depth investigations into future forecasting with the aid of CIL, and evaluate several representative prediction methods on PROPHET, drawing valuable insights for this task.

1 Introduction

The quest to forecast future events from Web information has long been a central pursuit in artificial intelligence (AI), with the potential to transform finance (Li et al., 2024), climate science (Wang and Karimi, 2024), and social policy (Rotaru et al., 2022). The recent rise of large language model (LLM)-based systems has demonstrated an unprecedented capacity for forecasting and drawn significant research attention (Halawi et al., 2024; Hsieh et al., 2024; Pratt et al., 2024).

To evaluate these LLM-based forecasting systems, prior works construct benchmarks from real-world questions (Halawi et al., 2024; Guan et al., 2024; Karger et al., 2024; Zeng et al., 2025), framing future forecasting as a retrieval-augmented generation (RAG)-and-reasoning task: systems first search the Web or databases for news articles related to the question, then reason over the retrieved knowledge base. However, truly evaluating LLM-based forecasting requires the questions to be *inferable*—the supporting knowledge base must contain sufficient information to substantiate the answers. Unlike traditional knowledge-intensive benchmarks such as HotpotQA (Yang et al., 2018) and 2WikiMultiHopQA (Ho et al., 2020), where answers are guaranteed to lie in the corpus, forecasting requires rationales (i.e. facts and reasoning clues) that the knowledge base may only partially provide (Zhao et al., 2024). For example, the question "Will Bitcoin rise to $100,000 by December 2024?" is only meaningfully answerable when the retrieved news covers leading indicators such as macroeconomic conditions, ETF approvals, or major holders' actions; if only off-topic articles are available, no system can reasonably infer the

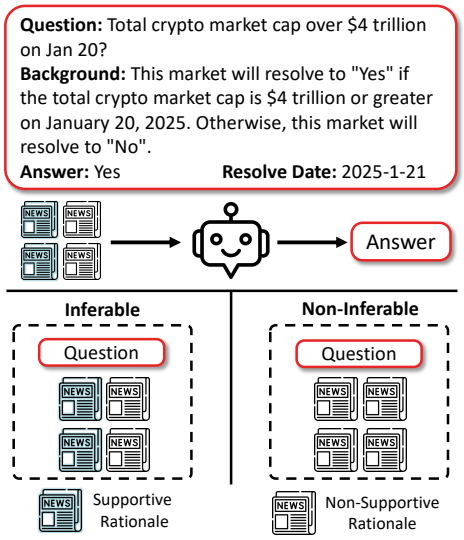

Figure 1: The upper Figure demonstrates future forecasting task. The lower half shows both inferable and non-inferable scenarios.

outcome. Collecting real-world questions without nuanced validation can thus yield non-inferable cases (Birur et al., 2024), where benchmarks measure guessing rather than genuine forecasting ability.

To overcome this challenge, we introduce PROPHET, an inferable future forecasting benchmark designed to provide a more accurate evaluation. For reproducibility, PROPHET is an RAG task where each real-world question pairs with relevant downloaded news articles for retrieval. We are next motivated to select inferable questions based on their related articles. The hardest part is estimating inferability without observing the full real-world event evolution; even with such observation, it is difficult given the lack of expert knowledge across diverse domains. Manual annotation is also impractical, as judging whether a set of articles is collectively sufficient demands deep, up-to-date understanding of finance, geopolitics, science, and more. Our key innovation is Causal Intervened Likelihood (CIL), a statistical measure that assesses inferability via causal inference. CIL regards each article as an event and measures the effect of intervening on its occurrence: the change in answer probability when forcing the event to happen versus not happen. This causal framing distinguishes articles that genuinely drive the outcome from those that merely co-occur with it, sidestepping the calibration pitfalls of using raw conditional likelihoods alone. CIL provides a robust estimate of whether a question is answerable. We then filter prediction questions with CIL to ensure benchmark inferability, enabling a fair and accurate evaluation. Assisted by CIL, PROPHET serves as a well-formulated RAG-and-reasoning task with hidden rationales (Zhao et al., 2024).

To validate CIL, we conduct extensive experiments that rigorously test how well it represents question inferability. The results show a strong correlation between CIL scores and system performance in both retrieval and prediction accuracy. Further, CIL enables in-depth investigations into future forecasting, drawing out innate properties of this complicated task. Finally, we evaluate several state-of-the-art prediction systems on PROPHET, highlighting their strengths, weaknesses, and directions for future research. We will also regularly update the dataset to maintain timeliness and minimize data leakage from model evolution. To summarize our contribution:

- We are the first to introduce CIL for inferability estimation of future forecasting. We provide a feasible method for calculating it and attest to its validity.
- Assisted by CIL, we establish an automatic pipeline to construct the future forecasting benchmark PROPHET where the real-world prediction questions are insufficiently inferable.
- We evaluate several baselines for future forecasting. The results show the pros and cons of these systems and present great potential and development directions.

2 Preliminaries

2.1 Future Forecasting

Future forecasting means predicting whether an event will occur in the future based on Web news. We formalize the task as binary question answering. Given a prediction question Q such as "Will Tim Walz win the VP debate against J.D. Vance?" or "Will Bitcoin rise to $100,000 by December 2024?", a background B describes the context of Q. A large set of news articles $\mathbb{X}$ serves as the retrieval knowledge base. The forecasting system must answer the question as formalized:

$$Y = \text{Reason}(Q, B, \mathbb{X}), \tag{1}$$

where $Y \in [0, 1]$ is the predicted probability that the event in Q occurs. A ground-truth answer $\hat{Y} \in \{0, 1\}$ paired with a resolved date D records whether the event in Q finally occurs and when the question resolves. Following prior work (Halawi et al., 2024; Karger et al., 2024), we use the Brier Score (Brier, 1950) for evaluation:

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

N is the number of questions. We report the Brier Score multiplied by 100 to 2 decimal places.

We formalize future forecasting as an RAG task. As an RAG, it differs from traditional datasets such as HotpotQA (Yang et al., 2018) and 2WikiMultiHopQA (Ho et al., 2020): the knowledge base $\mathbb{X}$ stores rationales and clues for answering Q (Zhao et al., 2024), and the task tests two entangled abilities of the system, retrieval and reasoning.

Existing benchmarks are built by harvesting real-world questions and pairing them with news before the resolved date D (Halawi et al., 2024; Guan et al., 2024; Karger et al., 2024), without validating inferability. $\mathbb{X}$ may lack sufficient supportive information for the question, so methods are needed to ensure that the questions are inferable.

2.2 Causal Inference

Causal inference is a vital statistical tool to determine causal relationships between variables (Pearl, 2010). A mere correlation between two variables may arise by chance or hidden factors; causal inference aims to establish direct causality. For instance, ice-cream sales and drowning incidents are not causally linked, although both are affected by hot weather. It uses concepts such as structural causal models, interventions, and counterfactual inference, with applications in medicine, economics, and social sciences.

Structural causal model (SCM) SCM represents and analyzes causal relationships via causal graphs and structural equations. It relies on a directed graph where nodes are variables X and edges denote direct causal influences, capturing dependencies and pathways of causation. Each variable is determined by its direct causes (parent nodes). SCM enables the identification of causal effects and the exploration of interventions (e.g., "what would happen if we intervened on X?"), and is widely applied in epidemiology and economics to disentangle complex causal mechanisms (Stolfo et al., 2023).

Interventional distribution An atomic intervention $\text{do}(X_i = x)$ fixes X_i at a value x (e.g., setting the dose of a new drug for a group in a medical trial). Interventions change the original structure and variable distributions: after $\text{do}(X_i = x)$, the distribution is $P(\cdot|\text{do}(X_i = x)) \doteq P_m(\cdot|X_i = x)$, showing how other variables respond.

A critical distinction lies between interventional and observational probabilities. The observational probability $P(Y|X = x)$ is the likelihood of Y given that we passively observe $X = x$, capturing mere statistical association. The interventional probability $P(Y|\text{do}(X = x))$ is the likelihood of Y when we actively set X to x, describing the causal effect of X on Y by simulating a controlled experiment and removing spurious correlations. Interventional probabilities are essential for predicting the effects of actions but cannot be computed directly from passively collected data; a primary challenge of causal inference is determining whether and how they can be obtained from observational probabilities (Pearl, 2010).

3 PROPHET Benchmark

In this section, we introduce PROPHET, a future forecasting benchmark with inferability estimation and selection. We first describe data collection (Section 3.1), introduce Causal Intervened Likelihood (CIL) (Section 3.2), describe benchmark construction (Sections 3.4 and 3.3), and report statistics (Section 3.5).

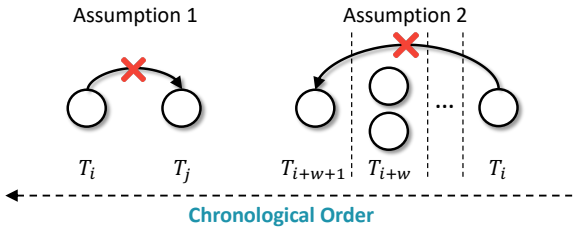

Figure 2: Illustration of assumptions. Nodes are variables that are in chronological order with time T.

3.1 Data Collection

Our objective is to gather a dataset of recent, prominent prediction questions. We source questions from Polymarket[1], whose questions span a wide range of domains—from scientific breakthroughs to social and economic trends—ensuring a representative benchmark. The questions are trending and attention-attracting, making them both contextually relevant and of broad interest to the forecasting community.

To avoid model leakage, we carefully select questions. We include all Polymarket questions resolved between 2025-01-01 and 2025-01-31, and filter out meaningless ones (e.g., personal inquiries or those with little community interest). We then collect as many relevant news articles as possible: using an LLM, we generate three types of search queries per question—entities, resolving steps, and similar historical events (Appendix prompts (a–c)). We query the open-source MediaCloud platform[2] to gather comprehensive news URLs, download contents via Newspaper3k[3], and discard articles published after the resolved date.

However, many downloaded articles are irrelevant; some carry incorrect publish dates that could cause leakage. To address this, we use an LLM to rate each article's relevance to the question, where the highest score means it directly answers the question and the lowest means total irrelevance. We remove articles with the highest and lowest scores (relevance prompt in Appendix (d)).

3.2 Causal Intervened Likelihood

To measure the sufficiency of supportive rationales for each question and build an inferable benchmark, we introduce **Causal Intervened Likeli-**

[1]https://polymarket.com/
[2]https://www.mediacloud.org
[3]https://newspaper.readthedocs.io/

hood (CIL), derived from causal inference. CIL estimates how strongly each news article supports the answer to a question. The key idea is to treat each candidate rationale as an event whose occurrence may or may not causally drive the answer; an article that genuinely supports the answer should change the predicted outcome when we counterfactually flip its occurrence. We model event occurrence with Bernoulli variables: $Y \in \{0, 1\}$ indicates whether the event in the question occurs, and $X_i \in \{0, 1\}$ indicates whether the event in the i-th article occurs. Each X_i has a happening date T_i. We write $T_i \prec T_j$ to mean the i-th news occurs before the j-th. The date of Y is later than any T_i.

Intuitively, if the occurrence of the i-th article ($X_i = 1$) is a necessary condition for $Y = \hat{Y}$, then $\mathrm{do}(X_i = 0)$ would significantly increase the probability of $Y \neq \hat{Y}$. Conversely, an article that is merely correlated with the answer but not causally responsible should leave the answer probability largely unchanged under intervention. We thus define the CIL of the i-th article as the gap between these two interventional probabilities:

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

where $\mathrm{do}(\cdot)$ is the causal-inference intervention, forcing X_i to happen or not.

`CIL` is composed of interventional probabilities. As in standard causal inference, we reduce these interventional probabilities to observational ones. We model all X_i and Y as a Structural Causal Model (SCM), with variables as nodes and causal relations as directed edges. Determining the full SCM is impractical due to incomplete knowledge and intensive expert analysis, so methods relying on a complete SCM are infeasible. We instead make two key assumptions.

Assumption 1. Temporality *For any two news events, the one that occurs later cannot causally affect the one that occurred earlier:*

$$\forall i, j, \quad \text{if } T_i \prec T_j, \text{then} (X_j, X_i) \notin \text{edges of SCM}. \tag{4}$$

This commonsense assumption aligns causality with the flow of time and rules out cycles in the SCM. Note that Y is chronologically the last variable in this model.

Assumption 2. w-day Dependency Window *Direct causal influence between news events is time-limited: X_j can have a direct causal effect on X_i only if X_j occurs within w-day window before X_i:*

$$\forall i, j : (T_i – T_j) > w \Rightarrow (X_i, X_j) \notin \text{edges of SCM}. \tag{5}$$

This assumes that distant past news affects X_i only through more recent intermediate events.

With these assumptions, we derive the CIL estimation. We show the calculation for $P(Y = \hat{Y}|\mathrm{do}(X_i = 1))$; the term $P(Y = \hat{Y}|\mathrm{do}(X_i = 0))$ is computed analogously.

Proposition. *The intervened probability $P(Y = \hat{Y} \mid \mathrm{do}(X_i = 1))$ can be expressed purely in terms of observational probabilities:*

$$P(Y{=}\hat{Y} \mid \mathrm{do}(X_i{=}1)) = \sum_{\mathbf{x}_{\mathcal{N}_i}} P(Y{=}\hat{Y} \mid X_i{=}1, \mathbf{X}_{\mathcal{N}_i}{=}\mathbf{x}_{\mathcal{N}_i})\, P(\mathbf{X}_{\mathcal{N}_i}{=}\mathbf{x}_{\mathcal{N}_i}). \tag{6}$$

- $\mathcal{N}_i = \{\, j \mid 0 \leq T_i - T_j \leq w \,\}$ is the set of indices to news articles published within the w-day window prior to X_i.
- $\mathbf{X}_{\mathcal{N}_i}$ denotes the vector of binary random variables $(X_j)_{j \in \mathcal{N}_i}$, where $X_j \in \{0, 1\}$ indicates the occurrence of jth news event.
- $\mathbf{x}_{\mathcal{N}_i}$ denotes one specific assignment of values to $\mathbf{X}_{\mathcal{N}_i}$, i.e., a binary vector $(x_j)_{j \in \mathcal{N}_i}$ with each $x_j \in \{0, 1\}$.
- The summation $\sum_{\mathbf{x}_{\mathcal{N}_i}}$ runs over all $2^{|\mathcal{N}_i|}$ possible binary value combinations of $\mathbf{x}_{\mathcal{N}_i}$, thereby enumerating every possible pattern of possibility of events in $\mathcal{N}_i$.

We leave the proof to the Appendix. The remaining task is to compute the observational probabilities $P(Y = \hat{Y}|X_i = 1, \mathbf{X}_{\mathcal{N}_i} = \mathbf{x}_{\mathcal{N}_i})$ and $P(\mathbf{X}_{\mathcal{N}_i} = \mathbf{x}_{\mathcal{N}_i})$. Inspired by Bynum and Cho, we leverage LLMs to estimate them. Since each X_j is Bernoulli, the summation enumerates all $2^{|\mathcal{N}_i|}$ value combinations. We use prompts similar to Halawi et al. to query the LLM for these conditional and joint probabilities (Appendix prompts (e–f)). We use a dependency window of $w = 30$ days.

Crucially, our derivation (Eq. 6) decomposes the causal query into observational conditional probabilities. While LLMs may struggle with direct causal reasoning, extensive pre-training makes them state-of-the-art density estimators for observational text distributions. We thus use LLMs strictly to estimate these statistical dependencies ($P(Y|X)$), while the causal validity is guaranteed by the structural constraints of our SCM (Assumptions 1 & 2), not by the model's reasoning.

3.3 News Events Composition

When computing CIL (Eq. equation 6) in practice, the number of articles in the w-day dependency window ($w = 30$) can be very large. Directly enumerating all $2^{|\mathcal{N}_i|}$ value combinations of $\mathbf{X}_{\mathcal{N}_i}$ is computationally infeasible, and in our pilot study a single 30-day window routinely contained hundreds of articles, making naive enumeration intractable for the full benchmark. We therefore design a *News Events Composition* procedure that aggregates and compresses related news into fewer representative nodes before enumeration, preserving the causal information needed for CIL while greatly reducing complexity.

The motivation is that many articles within a short period describe similar or redundant events, differing only in detail or style, and can be generalized to larger-granularity events. For instance, multiple outlets typically issue near-duplicate coverage of the same press conference, and at the level of causal effect on the answer they collectively act as a single event. We thus aggregate temporally and semantically similar articles into concise summaries. Formally, when computing CIL for X_i:

- **Individual summarization.** We employ the LLM to produce a concise, factual summary for each article X_j, capturing the essential event semantics while discarding irrelevant details, serving as a standardized basis for grouping.
- **Grouping within the dependency window.** We identify $\mathcal{N}_i = \{j \mid 0 \leq T_i - T_j \leq w\}$, i.e., all events within w days before X_i, and split them into groups of at most 10 articles, ordered chronologically.
- **Group-level relevance extraction.** For each group, we concatenate the summaries into a single block and query the LLM to extract and condense only the content relevant to X_i, removing unrelated details.
- **Clustering of extracted summaries.** Relevance-extracted summaries are embedded into a semantic vector space (e.g., via sentence embeddings[4]) and clustered by semantic similarity, capturing redundancy across groups.
- **Final cluster summarization.** For each cluster, we use the LLM to summarize the shared content into a final compact representation. These summaries form the reduced node set we enumerate when computing CIL for X_i.

[4]https://sbert.net/

This pipeline replaces a large, potentially redundant set $\mathcal{N}_i$ with a much smaller set of semantically distinct representative events, sharply reducing the number of permutations over $\mathbf{X}_{\mathcal{N}_i}$ in Eq. equation 6 while retaining the causal context needed for accurate interventional estimation.

3.4 Construction

After computing CIL for all news, we construct the benchmark. For each question, if it has any article with CIL> 0.7, we add it to the chosen set *L1*; otherwise it goes to *L2*. *L1* forms the main part of our benchmark since its questions are sufficiently supported, serving as an RAG benchmark, while *L2* lacks sufficient support. We use QwQ-32B (Qwen et al., 2025) to compute observational probabilities and Qwen2.5-32B (Qwen et al., 2025) for other data processing. Note that:

Causality Assumptions. Our assumptions reflect general commonsense and capture the dominant patterns in news–event relationships. We do not model global causality; we model only the causality required for the task with appropriate parameters. In practice, this means our SCM is intentionally a coarse approximation that trades fine-grained fidelity for tractability, an explicit modeling choice that lets us scale to thousands of articles per question while keeping the resulting CIL scores informative for filtering.

Probability Computing. In pilot experiments, different LLMs gave slightly different probability scores in CIL. We therefore query a single LLM multiple times for reliable estimation, averaging across runs to dampen sampling noise. Later experiments show that CIL is model-agnostic: different models reach the same conclusions, validating that the metric tracks an underlying causal signal rather than artifacts of any particular estimator.

3.5 Statistics and Properties of PROPHET

We compute basic statistics of PROPHET and, assisted by CIL, explore key properties of the future forecasting task. We harvest 612 data points in L1 and 43 in L2. The crawled news statistics are in the Appendix table. During construction we discard only obviously irrelevant news, so the data distribution of valid news is not significantly altered. Thus the retained articles reflect the real distribution of situations surrounding queried events.

Each question yields on average 559.51 articles, with 1249.88 average tokens per article—posing a challenge for methods that simply concatenate

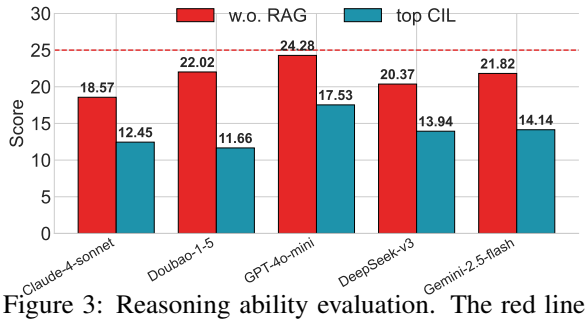

Figure 3: Reasoning ability evaluation. The red line stands for random results which is 25.0.

all news into one prompt. The mean CIL score is -0.013, indicating slightly more negative-supportive than positive-supportive news. Deep analysis is in the Appendix.

4 Experiments

We first conduct experiments to show the validity of CIL estimation and our benchmark in Section 4.1, then evaluate various forecasting methods on PROPHET, covering both Naive RAG and Agentic RAG baselines.

Naive RAG Naive RAG is the straightforward form of retrieval-augmented generation: the system directly retrieves relevant documents from $\mathbb{X}$ and conditions generation on them without iterative reasoning or tool use, formalized as:

$$\begin{aligned} \mathbb{R} &= \text{Retrieve}(Q, B, \mathbb{X}), \\ Y &= \text{Generate}(Q, B, \mathbb{R}), \end{aligned} \tag{7}$$

where $\mathbb{R}$ is the retrieved articles, Retrieve a dense or sparse retriever, and Generate an LLM that outputs the probability estimate. There is no explicit multi-step reasoning over retrieval results; the model encodes all reasoning based solely on $\mathbb{R}$. Prompt is in Appendix prompt (g).

Agentic RAG Agentic RAG extends Naive RAG by introducing an *agent* that iteratively retrieves, reasons, and acts on intermediate conclusions, following the ReAct paradigm (Yao et al., 2023). The model alternates between reasoning steps τ, action steps α (tool invocation for retrieval), and observations o, enabling multi-step evidence gathering before the final answer:

$$(Q, \tau_1, \alpha_1, o_1, \tau_2, \alpha_2, o_2, \ldots, \tau_L, \alpha_L, o_L) \tag{8}$$

All agents follow the ReAct framework. Each agent shares the same retrieval tool, prompts, and interaction structure; only the underlying LLM varies. The retrieval tool implements document retrieval: the agent issues a natural-language query that is encoded into a dense vector via a fixed embedding model and matched against $\mathbb{X}$ by cosine similarity. Prompt and tool details are in the Appendix (prompt (h)).

We report results in Sections 4.3 and 4.4, then isolate the reasoning ability of each model in Section 4.2, analyzing performances along the forecasting timeline with case studies in the Appendix.

4.1 Validity of CIL and PROPHET

To validate our CIL metric, we design experiments testing its practical impact. The intuition is that if CIL is effective, forecasts based on the highest-CIL articles should significantly outperform forecasts that ignore them. Equivalently, articles selected by CIL should contribute concrete answer-relevant signal beyond what the LLM already knows from its parametric memory, so removing them should hurt and including them should help. We use hypothesis tests to determine whether the difference between the two approaches is statistically significant, rather than relying on point estimates that can be confounded by per-question variance and LLM stochasticity.

We partition the news corpus into two disjoint subsets by CIL score distribution and evaluate two strategies on each: (1) **Top CIL**: feed the top-20 articles ranked by CIL to the generator; (2) **w.o. RAG**: answer using the question alone. Every question is repeated 10 times for stable estimates.

Let $X = (x_i)_{1\times 10n}$, where each x_i is the Brier score of one response to a question under the *with-RAG* strategy; analogously $Y = (y_i)_{1\times 10n}$ are scores *without RAG* ($10n$ scores per vector since each question is repeated 10 times). Define paired differences $Z = (x_i - y_i)_{1\times 10n}$. We run one-sample t-tests on Z at $\alpha = 0.05$ under two one-sided null hypotheses:

$$\begin{aligned} \mathcal{H}_0^{\leq} &: \mu(Z) \leq 0.077, \\ \mathcal{H}_0^{\geq} &: \mu(Z) \geq 0.027. \end{aligned}$$

Rejecting $\mathcal{H}_0^{\leq}$ means the average Brier score with RAG exceeds that without RAG by at least 0.077, i.e., the retrieved articles help answer the question. Rejecting $\mathcal{H}_0^{\geq}$ means the improvement is at most 0.027, i.e., the articles give negligible gain.

Under L1, $\mathcal{H}_0^{\leq}$ is rejected ($p = 0.0264$), showing that forecasts from the top-20 CIL-scored articles achieve a significantly larger improvement (over 7.7) than direct answers. Under L2, $\mathcal{H}_0^{\leq}$ cannot be rejected ($p = 1.0000$), indicating the improvement is not statistically significant. This pattern shows CIL successfully identifies highly supportive articles. LLM-predicted probabilities may have

Model	w.o. RAG	AM	MDC	MM	MDD	MBD	IL	IB
Claude-4-sonnet	18.57	19.00/19.49	19.15/18.72	18.12/18.00	19.15/19.64	18.41/18.87	18.24/19.45	19.69/19.14
Doubao-1-5	22.02	20.73/21.01	21.45/21.00	22.72/21.09	20.53/20.25	21.89/21.23	20.33/21.18	21.85/22.05
GPT-4o-mini	24.28	27.17/27.20	28.32/28.92	28.34/27.89	29.41/28.95	29.08/30.34	27.52/28.06	29.10/28.46
DeepSeek-v3	20.37	21.56/21.48	21.94/21.74	22.22/21.04	21.33/22.60	22.10/22.27	21.60/22.96	23.08/22.55
Gemini-2.5-flash	21.82	21.63/21.48	23.69/22.19	22.58/23.63	21.11/22.68	22.06/20.15	21.54/21.16	21.99/22.06

Table 1: Naive RAG on the PROPHET dataset. Lower is better. Each cell reports the score with $n = 10/n = 20$ articles.

Subset	Hypothesis	Decision	p-value	$\bar{d}$
L1	$\mathcal{H}_0^{\leq} : \mu(Z) \leq 0.077$	Rejected	0.0264	0.0835
	$\mathcal{H}_0^{\geq} : \mu(Z) \geq 0.027$	Not rejected	1.0000	
L2	$\mathcal{H}_0^{\leq} : \mu(Z) \leq 0.077$	Not rejected	1.0000	0.0176
	$\mathcal{H}_0^{\geq} : \mu(Z) \geq 0.027$	Rejected	0.0474	

Table 2: Hypothesis-testing results for subsets L1 and L2. $\bar{d} = \bar{x} - \bar{y}$ denotes the mean improvement of *Top CIL* over *without RAG*.

calibration errors, but for retrieval the relative ranking of evidence dominates absolute values. The significant gap between Top-CIL and the random baseline ($p = 0.0264$) confirms that CIL robustly distinguishes supportive rationales from noise despite potential calibration biases.

4.2 Reasoning Performances

We select the top-10 CIL articles per question for prediction and compare with the no-RAG baseline. Results are in Figure 3. The Brier Scores of top-10 CIL selection are significantly better than those without RAG for all tested models, further showing that CIL identifies high-quality articles that boost forecasting performance.

The strong top-10 CIL performance also suggests minimal answer leakage in the dataset. If substantial leakage were present, RAG-assisted prediction would not improve so consistently over the baseline, as retrieved content would simply repeat ground-truth answers rather than genuinely aiding reasoning. These results indicate that the dataset preserves integrity while still allowing meaningful gains through retrieval.

4.3 Evaluation On Naive RAG Baselines

We evaluate naive Retrieval-Augmented Generation (RAG) baselines on PROPHET to establish a reference and examine the task's practical challenges. The generators are Claude-sonnet-4, Doubao-1.5, GPT-4o-mini, DeepSeek-v3, and Gemini-2.5-flash. The retrievers are seven popular open-source embedding models: all-MiniLM-L6-v2 (AM), msmarco-distilbert-cos-v5 (MDC), msmarco-MiniLM-L6-cos-v5 (MM), msmarco-distilbert-dot-v5 (MDD), msmarco-bert-base-dot-v5 (MBD), instructor-large (IL), and instructor-base (IB). Each LLM–embedding pair is evaluated with retrieval sizes $n = 10$ and $n = 20$, reporting Brier Scores (Table 1). Key takeaways:

Limited capability of Naive RAG. Across all configurations, naive RAG does not consistently outperform the "w.o. RAG" baseline and often degrades performance. Simply appending retrieved documents to the LLM input—without filtering, temporal alignment, or causal reasoning—is insufficient for PROPHET forecasting. Retrieved context often contains redundant or irrelevant information that may conflict with the model's internal knowledge, confusing the generator and inflating the Brier Score. Identifying truly predictive evidence from historical data is also hard, suggesting naive RAG lacks mechanisms to reason about event timelines, causal links, or uncertainty.

Small differences across current embedding models. Changing among the seven embedding models yields only modest variation for a given LLM, suggesting current general-purpose retrievers are not well suited to temporally or causally grounded forecasting and that task-specific embeddings may be needed.

Increasing retrieval size has limited benefit. Going from $n = 10$ to $n = 20$ seldom improves and sometimes degrades performance: more documents add noise and increase reasoning difficulty. Selecting high-quality evidence matters more than retrieval quantity.

Overall, the results highlight the need for causally aware retrieval strategies and task-tailored reasoning methods for forecasting.

4.4 Agentic RAG Performances

We also evaluate Agentic RAG. AM and MDC are different embedding models used in the tool. From Table 3, Agentic RAG achieves consistent gains over the w.o. RAG baseline and outperforms naive RAG (Section 4.3) for several LLMs; e.g., GPT-4.1 and Gemini-2.5-pro show Brier reductions exceeding -2.0. Unlike naive RAG's simple concatenation, Agentic RAG lets the LLM iteratively retrieve, inspect, and integrate evidence, improving temporal and causal relevance while reducing noise. Gains often correlate with deeper reasoning, as in-

	w.o. RAG	AM			MDC		
	BS	BS (δ)	Step	Thought	BS (δ)	Step	Thought
Claude-4-sonnet	18.57	18.65 (+0.08)	3.67	195.65	17.89 (-0.68)	3.58	211.99
Claude-3.5-sonnet	22.22	21.53 (-0.69)	1.05	169.75	21.54 (-0.68)	1.02	202.83
Doubao-1.5	20.20	20.40 (+0.20)	1.02	129.97	24.53 (+4.33)	1.08	166.08
Gemini-25-pr0	21.41	20.51 (-0.90)	0.95	139.64	19.26 (-2.15)	0.95	151.01
GPT-4.1	21.64	21.14 (-0.50)	0.61	91.24	18.44 (-3.20)	0.18	101.94
GPT-4o	22.31	23.33 (+1.02)	1.02	102.64	22.47 (+0.16)	1.04	125.38

Table 3: Agentic RAG results. Lower is better. δ shows differences between Agentic and w.o. RAG. Thought is average thought token.

dicated by higher *Step* counts and longer *Thought* lengths: stronger models tend to issue more targeted follow-up queries when initial evidence is ambiguous, whereas weaker models stop early and rely on parametric priors.

Overall, Agentic RAG is a promising direction for forecasting, combining adaptive retrieval with reasoning to better exploit knowledge, and our results suggest that future systems should invest in retrieval policies that are explicitly aware of causal and temporal evidence rather than treating retrieval as a one-shot black box.

5 Related Work

5.1 Future Forecasting and Benchmarks

Early forecasting benchmarks such as MCNC (Granroth-Wilding, 2016), SCT (Mostafazadeh et al., 2017), and CoScript (Yuan et al., 2023) target script learning and commonsense reasoning in synthetic scenarios, lacking real-world grounding. Time-series datasets GDELT (Leetaru and Schrodt, 2013) and ICEWS (Schrodt et al., 2012) track real events but do not formalize prediction as retrieval-augmented reasoning or ensure answerability. ECARE (Du et al., 2022) and EV2 (Tao et al., 2024) advance event reasoning yet remain confined to abstract settings, limiting practical forecasting and causal inference. FutureX (Zeng et al., 2025) construct online event forecasting evaluation by collecting questions from prediction market and open leaderboards.

With LLMs, benchmarks such as Halawi et al., OpenEP (Guan et al., 2024), and ForecastBench (Karger et al., 2024) shift to real-world questions and news search, but (1) lack explicit inferability validation, and (2) prioritize dynamic sources over reproducibility under evolving archives. PROPHET addresses both via Causal Intervened Likelihood filtering.

5.2 RAG and Benchmarks

Foundational: QA datasets 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), grounded on Wikipedia-style bases, are widely adapted for RAG evaluation.
Domain-Agnostic: RAGBench (Friel et al., 2024) spans biomedical, legal, customer support, and finance; CRAG (Wang et al., 2024a) provides factual QA across five domains, simulating web and knowledge-graph search.
Domain-Specific: LegalBench-RAG (Wang et al., 2024b), WeQA (Meyur et al., 2024), PubHealth (Zhang et al., 2023), and MTRAG (Tang and Yang, 2024) target niche applications for precise in-domain evaluation.
Capability-Oriented: RGB (Liu et al., 2024) probes noise robustness, negative rejection, information integration, and counterfactual robustness; TRIAD (Zong et al., 2024) assesses retrieval quality, fidelity, and utility in 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.

6 Conclusion

We address the challenge of building an inferable RAG benchmark for evaluating future forecasting systems by introducing PROPHET, rigorously validated for inferability by our Causal Intervened Likelihood (CIL) estimation. Using causal inference to quantify the inferability of questions from their news articles, PROPHET ensures questions are answerable through retrieved rationales, yielding a more accurate assessment of model capabilities. Experiments confirm that CIL correlates with system performance, and evaluations of state-of-the-art systems on PROPHET reveal key strengths and limitations in retrieval and reasoning. PROPHET thus enables inferable evaluation that drives progress toward 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*.

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.

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.

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*.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

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*.

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*.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*.

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.

Zhiyuan Zeng, Jiashuo Liu, Siyuan Chen, Tianci He, Yali Liao, Yixiao Tian, Jinpeng Wang, Zaiyuan Wang, Yang Yang, Lingyue Yin, and 1 others. 2025. Futurex: An advanced live benchmark for llm agents in future prediction. *arXiv preprint arXiv:2508.11987*.

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, 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.

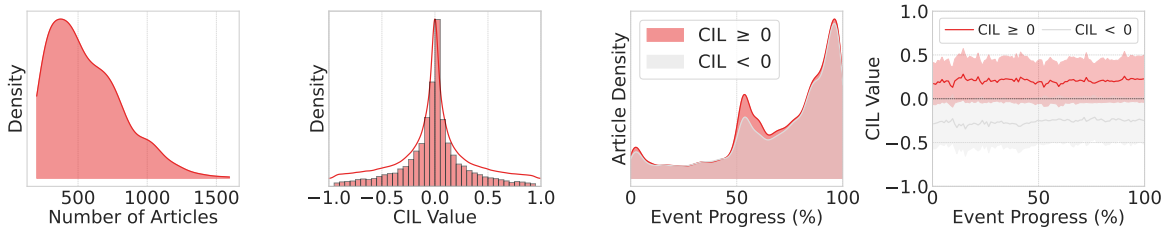

Figure 4: In-depth statistics on PROPHET. (a) Distribution of article counts per question. (b) Overall distribution of CIL values across all articles. (c) Article density along event progress (date ranging from the earliest to the latest news) for positive (CIL ≥ 0) and negative (CIL < 0). (d) Mean CIL values and standard deviations as a function of forecasting progress.

Appendix

Proof of Proposition

Proof. Let $\mathcal{N}_i = \{j \mid 0 \leq T_i - T_j \leq w\}$ be the set of indices for variables within the w-day window before X_i, and let $\mathcal{M}_i = \{j \mid T_i - T_j > w\}$ be the set for those outside the window. Using the law of total probability, we can expand the post-intervention probability by marginalizing over all variables that are not descendants of X_i. Due to Assumption 1 (Temporality), we only need to consider variables that precede X_i.

$$\begin{aligned} &P(Y = \hat{Y} \mid \mathrm{do}(X_i = 1)) \\ &= \sum_{\mathbf{x}_{\mathcal{N}_i}} \sum_{\mathbf{x}_{\mathcal{M}_i}} P\big(Y = \hat{Y} \mid X_i = 1, \\ &\qquad \mathbf{X}_{\mathcal{N}_i} = \mathbf{x}_{\mathcal{N}_i}, \mathbf{X}_{\mathcal{M}_i} = \mathbf{x}_{\mathcal{M}_i}\big) \\ &\quad \times P\big(\mathbf{X}_{\mathcal{N}_i} = \mathbf{x}_{\mathcal{N}_i}, \\ &\qquad \mathbf{X}_{\mathcal{M}_i} = \mathbf{x}_{\mathcal{M}_i} \,\big|\, \mathrm{do}(X_i = 1)\big). \end{aligned} \tag{9}$$

Based on the adjustment formula in causal inference, an intervention on X_i only affects its descendants. Due to Assumption 1, the variables $\mathbf{X}_{\mathcal{N}_i}$ and $\mathbf{X}_{\mathcal{M}_i}$ all occur before X_i and thus cannot be its descendants. Therefore, the intervention on X_i does not affect their joint probability: $P(\mathbf{X}_{\mathcal{N}_i}, \mathbf{X}_{\mathcal{M}_i} | \mathrm{do}(X_i = 1)) = P(\mathbf{X}_{\mathcal{N}_i}, \mathbf{X}_{\mathcal{M}_i})$.

Furthermore, Assumption 2 states there are no direct causal paths from variables in $\mathbf{X}_{\mathcal{M}_i}$ to Y that are not mediated by variables closer to Y (including those in $\mathbf{X}_{\mathcal{N}_i}$ and X_i). This implies that given X_i and its more immediate predecessors $\mathbf{X}_{\mathcal{N}_i}$, Y becomes conditionally independent of the more distant predecessors $\mathbf{X}_{\mathcal{M}_i}$.

$$\begin{aligned} &P\big(Y = \hat{Y} \mid X_i = 1, \\ &\quad \mathbf{X}_{\mathcal{N}_i} = \mathbf{x}_{\mathcal{N}_i}, \mathbf{X}_{\mathcal{M}_i} = \mathbf{x}_{\mathcal{M}_i}\big) \\ &= P(Y = \hat{Y} \mid X_i = 1, \mathbf{X}_{\mathcal{N}_i} = \mathbf{x}_{\mathcal{N}_i}). \end{aligned} \tag{10}$$

Substituting this back into Equation (9):

$$\begin{aligned} &P(Y = \hat{Y} \mid \mathrm{do}(X_i = 1)) \\ &= \sum_{\mathbf{x}_{\mathcal{N}_i}} \sum_{\mathbf{x}_{\mathcal{M}_i}} P(Y = \hat{Y} \mid X_i = 1, \mathbf{X}_{\mathcal{N}_i} = \mathbf{x}_{\mathcal{N}_i}) \\ &\quad \times P(\mathbf{X}_{\mathcal{N}_i} = \mathbf{x}_{\mathcal{N}_i}, \mathbf{X}_{\mathcal{M}_i} = \mathbf{x}_{\mathcal{M}_i}) \\ &= \sum_{\mathbf{x}_{\mathcal{N}_i}} P(Y = \hat{Y} \mid X_i = 1, \mathbf{X}_{\mathcal{N}_i} = \mathbf{x}_{\mathcal{N}_i}) \\ &\quad \times \sum_{\mathbf{x}_{\mathcal{M}_i}} P(\mathbf{X}_{\mathcal{N}_i} = \mathbf{x}_{\mathcal{N}_i}, \mathbf{X}_{\mathcal{M}_i} = \mathbf{x}_{\mathcal{M}_i}). \end{aligned} \tag{11}$$

The inner summation over all configurations of $\mathbf{X}_{\mathcal{M}_i}$ is simply the marginal probability $P(\mathbf{X}_{\mathcal{N}_i} = \mathbf{x}_{\mathcal{N}_i})$.

$$\begin{aligned} &P(Y = \hat{Y} \mid \mathrm{do}(X_i = 1)) \\ &= \sum_{\mathbf{x}_{\mathcal{N}_i}} P(Y = \hat{Y} \mid X_i = 1, \mathbf{X}_{\mathcal{N}_i} = \mathbf{x}_{\mathcal{N}_i}) \\ &\quad \times P(\mathbf{X}_{\mathcal{N}_i} = \mathbf{x}_{\mathcal{N}_i}). \end{aligned} \tag{12}$$

This completes the proof. □

Deep Analysis of PROPHET

Beyond these summary statistics, we conduct more in-depth analyses, visualized in Figure 4 and reveal several key insights:

- **Article count distribution.** Figure 4(a) shows that most questions have between 400 and 800 associated articles, with a long tail of dense coverage for certain high-profile events. This indicates an uneven information supply across questions.
- **CIL value distribution.** As shown in Figure 4(b), the distribution of CIL values is sharply peaked around zero, indicating that most of the events are non-supportive. It's hard for methods to mine useful information for precise forecasting.
- **Temporal density of positive/negative articles.** Figure 4(c) illustrates article density over the course of event progress (date ranging from the earliest to the latest news) for positive ($\texttt{CIL} \geq 0$) and negative ($\texttt{CIL} < 0$) articles. We observe peaks at early stages, implying heightened reporting activity both in the initial phase and just before resolution.
- **Temporal evolution of sentiment.** As shown in Figure 4(d), the mean CIL values remain relatively stable over time for both positive and negative groups, with only small oscillations.

Overall, these statistical properties indicate that PROPHET offers both rich contextual diversity and robust inferability, making it a challenging benchmark for forecasting systems that must reason over temporally distributed Web information.

Temporal Analysis

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. 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 run different models based on the top-10 CIL articles in various prediction progress. The results are in Figure 5. As the prediction process progresses, the difficulty of prediction decreases, but early predictions still face great difficulties.

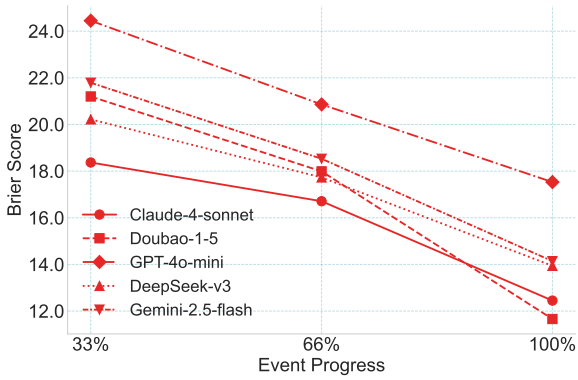

Figure 5: Temporal analysis. Results are on top-10 CIL articles on various forecasting progress.

Case Studies

To evaluate the proposed supportiveness metric in RAG, we consider two representative cases shown in Figure 6. The first question asks whether the total crypto market cap will exceed $4 trillion on January 20, 2025. Among the retrieved articles, the blue-highlighted news discusses substantial rebounds in Bitcoin and Ether alongside ETF approvals—information directly indicative of a "Yes" outcome. Conversely, the white-highlighted news describes global market declines and cryptocurrency crashes, contradicting the supportive stance. Our metric assigns higher scores to the supportive article, reflecting its stronger evidential link to the answer.

The second question concerns whether Joe Rogan will attend the presidential inauguration. The supportive article details Rogan's past political activities and his collaboration with Donald Trump during the campaign, which strengthens the likelihood of his attendance. In contrast, the non-supportive article mentions Rogan only in the context of broader shifts in media influence, without providing any concrete statement or event suggesting his presence at the inauguration. Our metric correctly ranks the supportive article higher, aligning with the ground truth answer "Yes".

Across both cases, the metric effectively distinguishes articles offering direct, answer-relevant evidence from those that are tangential or context-only, thereby enhancing RAG answer accuracy.

# Question	# News	Avg Article Words	Avg CIL	Avg Pos CIL	Avg Neg CIL	Earliest News Date	Latest News Date
612	559.51	1249.88	-0.013	0.2041	-0.2503	2023,1,1	2025,1,1

Table 4: Statistics of L1 part of PROPHET.

Figure 6: Case studies. The supportive and non-supportive articles are selected by CIL scores.

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).

Prompts

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

Agentic RAG Tool

We show the article retrieval tool for all agentic RAG methods:

```
class AgenticRAGTool:
    name = "query_rag"
    description = (
        "To search in a previously built
             RAG index to find the most
            relevant chunks of text."
    )
    parameters = [
        {
            'name': 'query',
            'type': 'string',
            'description': 'the query
                text to search for
                relevant text chunks.',
            'required': True
        },
        {
            'name': 'top_k',
            'type': 'integer',
            'description': 'the number
                of top relevant chunks
                to retrieve (default is
                3).',
            'required': False
        }
    ]
```

Listing 1: Python Class Definition for Agentic RAG Tool

(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: Q
Question Background: B
Question Date: D
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: Q
Question Background: B
Question Date: D
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: Q
Question Background: B
Question Date: D
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: Q
Question Background: B
Article: {article}
Please rate the relevance of the article to the question, at the scale of 1-6
1 – completely irrelevant
2 – slightly relevant
3 – somewhat relevant
4 – relevant
5 – highly relevant
6 – completely relevant directly reveals the answer
Guidelines:
- scale 1-5 don't reveal the answer, only scale 6 reveals the answer.
- If the article has events of similar types which may happened on different subjects, it also consider relevant to the question to some degree.
- 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, don't generate other words in the response:
Rating:

(e) Probability $P(Y = \hat{Y}|X_i = 1, \mathbf{X}_{\mathcal{N}_i} = \mathbf{x}_{\mathcal{N}_i})$

Given some event: {events}
Most important event: {current event}
Instructions: Based on the information, especially the most important event, compute the probability that the answer for the following question is yes. Question: Q
The mentioned named entities are real in life, no need to check that. You compute the probability based on your world knowledge about the mentioned named entities and the internal logic in the situation. the Output your answer (a probability number between 0 and 1).

(f) Probability $P(\mathbf{X}_{\mathcal{N}_i} = \mathbf{x}_{\mathcal{N}_i})$

Given a situation describing some events: {events}
Instructions: Reason the probability of the situation. The mentioned named entities are real in life, no need to check that. You compute the probability based on your world knowledge about the mentioned named entities and the internal logic in the situation. Output your answer (a probability number between 0 and 1).

(g) Naive RAG

Your task is to answer the prediction question. The question asks about whether a future event would happen or not. Output a probability between 0 and 1 representing the likelihood of the event happening. Output the probability in the following format:
<Probability>...</Probability>
Use the provided articles as context to inform your prediction. The articles are provided below.
Question: Q
Background: B
Create date of the question: D
Context Articles: {retrieved articles}

(h) Agentic RAG

You are a prophet agent who can answer a query based on the data and the background information. Use the tools below to help you answer the question.
Your task is to answer the prediction question. The question asks about whether a future event would happen or not. Output a probability between 0 and 1 representing the likelihood of the event happening. Output the probability in the following format:
<Probability>...</Probability>
Question: Q
Background: B
Create date of the question: D