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

**Table 1: Statistics of L1 part of PROPHET.**

derived from causal inference. CIL estimates how strongly each news article supports the answer to a given question.

We use Bernoulli variables to model the occurrence of events. Let $Y \in \{0,1\}$ indicate whether the event asked by the question occurs, and let $X_i \in \{0,1\}$ indicate whether the event described in the $i$-th news article occurs. Each variable $X_i$ is associated with a happening date $T_i$. We use the notation $T_i \prec T_j$ to represent that the occurrence of the $i$-th news is before that of the $j$-th. Note that the date of $Y$ is later than any date of $X_i$.

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

$$\texttt{CIL}_i = P(Y = \hat{Y}|\text{do}(X_i = 1)) - P(Y = \hat{Y}|\text{do}(X_i = 0)), \tag{3}$$

where $\text{do}(\cdot)$ is the intervention operation from causal inference, representing that $X_i$ is forced to happen or not.

The CIL estimation is composed by interventional probabilities. As similar to the core idea in causal inference, we derive these interventional probabilities to observational probabilities. We model all $X_i$ and $Y$ as a Structural Causal Model (SCM), where variables are nodes and causal relationships are directed edges. However, determining the complete SCM is extremely challenging due to incomplete knowledge and the need for intensive expert analysis. Therefore, calculating CIL via methods that rely on a complete SCM is impractical. Thus we introduce 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 common-sense assumption aligns causal relationships with the flow of time and eliminates cycles in the SCM. Note that $Y$ is the chronologically last variable in this model.

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

$$\forall i, j, \quad \text{if } (T_i - T_j) > w \text{ days, then}(X_i, X_j) \notin \text{edges of SCM}. \tag{5}$$

This assumption posits that the causal influence of distant past news events is mediated by more recent, intermediate news events.

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

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

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

- $\mathcal{N}_i = \{\, j \mid 0 \le T_i - T_j \le w \,\}$ 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 $j$th 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 $\mathcal{N}_i$.

Proof. Let $\mathcal{N}_i = \{j \mid 0 \le T_i - T_j \le w\}$ 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 $X_i$.

$$\begin{aligned} &P(Y = \hat{Y}|\text{do}(X_i = 1)) \\ &= \sum_{\mathbf{x}_{\mathcal{N}_i}} \sum_{\mathbf{x}_{\mathcal{M}_i}} P(Y = \hat{Y}|X_i = 1, \mathbf{X}_{\mathcal{N}_i} = \mathbf{x}_{\mathcal{N}_i}, \mathbf{X}_{\mathcal{M}_i} = \mathbf{x}_{\mathcal{M}_i}) \\ &\quad \times P(\mathbf{X}_{\mathcal{N}_i} = \mathbf{x}_{\mathcal{N}_i}, \mathbf{X}_{\mathcal{M}_i} = \mathbf{x}_{\mathcal{M}_i}|\text{do}(X_i = 1)). \end{aligned} \tag{7}$$

Based on the adjustment formula in causal inference [20], 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: $P(\mathbf{X}_{\mathcal{N}_i}, \mathbf{X}_{\mathcal{M}_i}|\text{do}(X_i = 1)) = P(\mathbf{X}_{\mathcal{N}_i}, \mathbf{X}_{\mathcal{M}_i})$.

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 $\mathbf{X}_{\mathcal{M}_i}$.

$$\begin{aligned} &P(Y = \hat{Y}|X_i = 1, \mathbf{X}_{\mathcal{N}_i} = \mathbf{x}_{\mathcal{N}_i}, \mathbf{X}_{\mathcal{M}_i} = \mathbf{x}_{\mathcal{M}_i}) \\ &= P(Y = \hat{Y}|X_i = 1, \mathbf{X}_{\mathcal{N}_i} = \mathbf{x}_{\mathcal{N}_i}). \end{aligned} \tag{8}$$

Substituting this back into Equation (7):

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

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

$$\begin{aligned} P(Y = \hat{Y}|\text{do}(X_i = 1)) = \sum_{\mathbf{x}_{\mathcal{N}_i}} &P(Y = \hat{Y}|X_i = 1, \mathbf{X}_{\mathcal{N}_i} = \mathbf{x}_{\mathcal{N}_i}) \\ &\times P(\mathbf{X}_{\mathcal{N}_i} = \mathbf{x}_{\mathcal{N}_i}). \end{aligned} \tag{10}$$

This completes the proof. □

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 [3], we leverage LLMs to estimate these probabilities. Note that since each $X_j \in \mathbf{X}_{\mathcal{N}_i}$ is a Bernoulli variable, the summation is

over all $2^{|\mathcal{N}_i|}$ permutations of their values. We construct prompts similar to the approach of Halawi et al. [9] to query the LLM for these conditional and joint probabilities. The specific prompts used are detailed in Appendix 6 (e-f). For our experiments, we use a dependency window of $w = 30$ days.

Importantly, trained on vast amounts of observational data, LLM excels at providing observational probabilities. However, LLMs are not strong enough to directly compute intervened probabilities [3]. Our derivation provides the necessary bridge. By computing both terms and substituting them into Equation 3, we can now calculate the CIL scores for each news article.

## 3.3 News Events Composition

In practical computation of CIL (Eq. (6)), we observe that the number of news articles within the $w$-day dependency window (we use $w = 30$ days) can be extremely large. Directly enumerating all $2^{|\mathcal{N}_i|}$ value combinations of $\mathbf{X}_{\mathcal{N}_i}$ would lead to exponential computational costs, making the naive implementation infeasible. To address this issue, we design a *News Events Composition* procedure to aggregate and compress related news items into fewer representative nodes before enumeration. This step preserves the key causal information needed for CIL computation while significantly reducing complexity.

The motivation behind this approach is that many news articles within a short period may describe similar or redundant events, differing only in details or reporting styles, and can be generalized to larger-granularity events. Thus, we aggregate temporally and semantically similar articles into concise, representative summaries.

Formally, when computing CIL for $X_i$, we proceed as follows:

- **Individual summarization.** We first employ the LLM to produce a concise, factual summary for each news article $X_j$ in the dataset. These summaries capture the essential event semantics while discarding irrelevant text details, serving as a standardized basis for grouping.
- **Grouping within the dependency window.** We identify the set $\mathcal{N}_i = \{j \mid 0 \leq T_i - T_j \leq w\}$, i.e., all news events occurring within $w$ days before $X_i$. We then divide these preceding events into groups of at most 10 articles each, ordered chronologically.
- **Group-level relevance extraction.** For each group, we concatenate the summaries of its articles into a single text block. We then query the LLM to extract and condense only the information relevant to $X_i$ from this block. This step removes unrelated details and focuses on causally pertinent content.
- **Clustering of extracted summaries.** The relevance-extracted summaries from all groups are embedded into a semantic vector space (e.g., via sentence embeddings[4]) and clustered based on semantic similarity. Clustering captures redundancy among different groups, as reports on similar events may appear in multiple sources or days.
- **Final cluster summarization.** For each cluster, we again use the LLM to summarize the shared content into a final, compact representation of that event cluster. These final summaries constitute the reduced set of nodes that we enumerate when computing the CIL for $X_i$.

[4]https://sbert.net/

Through this multi-stage compression pipeline, we replace a large, potentially redundant set $\mathcal{N}_i$ with a much smaller set of semantically distinct representative events. This significantly decreases the number of permutations over $\mathbf{X}_{\mathcal{N}_i}$ in Equation equation 6, thereby reducing the overall computational load, while still retaining the necessary causal context for accurate interventional probability estimation.

## 3.4 Construction

After calculating the CIL for all pieces of news, we construct the benchmark with them. For each question, if it contains news articles with CIL$>0.7$, 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 sufficiently 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 use QwQ-32B [22] to compute all observation probabilities and Qwen2.5-32B [22] to do other data processing. Note that:

**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.

## 3.5 Statistics and Properties of PROPHET

We perform basic statistics of PROPHET, and, assisted by CIL, we further explore the key properties of the future forecasting task. Currently, we harvest 612 data points in L1 and 43 data points in L2. The statistics of the crawled news articles are shown in Table 1. During the construction process, we only discard obviously irrelevant news, meaning that we did not significantly alter the data distribution of valid news. Thus, the retained articles can reflect the real distribution of situations surrounding queried events.

For each question, we retain 100 top relevant news articles, resulting in an average of 559.51 articles per question. The average token count per article is 1249.88, which poses a challenge for methods that attempt to simply concatenate all news into a single prompt. The average CIL score is $-0.013$, indicating that slightly more negative-supportive news exists compared to positive ones.

Beyond these summary statistics, we conduct more in-depth analyses, visualized in Figure 3 and reveal several key insights:

- **Article count distribution.** Figure 3(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 3(b), the distribution of CIL values is sharply peaked around zero, indicating that

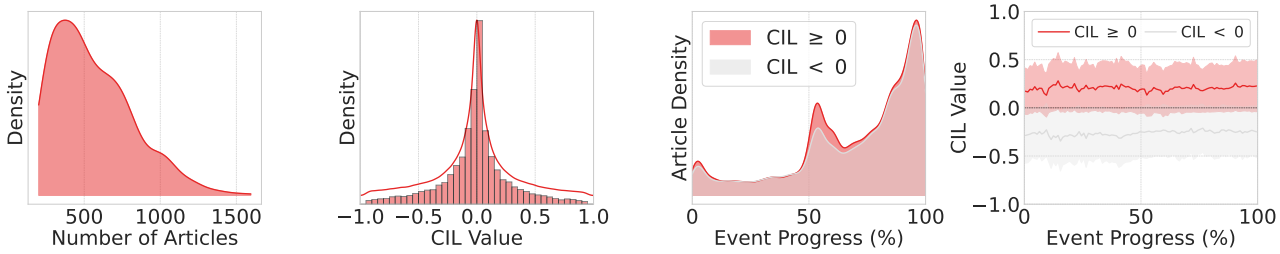


**Figure 3: In-depth statistics on PROPHET. (a) Distribution of article counts per question. (b) Overall distribution of CIL values across all articles. (c) Article density along event progress (date ranging from the earliest to the latest news) for positive (CIL ≥ 0) and negative (CIL < 0). (d) Mean CIL values and standard deviations as a function of forecasting progress.**

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 3(c) illustrates article density over the course of event progress (date ranging from the earliest to the latest news) for positive (CIL $\geq 0$) and negative (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 3(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.

## 4 Experiments

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

**Naive RAG** Naive RAG refers to the straightforward form of retrieval-augmented generation, where the system directly retrieves relevant documents from the knowledge base $\mathbb{X}$ and conditions the generation process on them without iterative reasoning or tool use. Formally, Naive RAG can be represented as:

$$\mathbb{R} = \text{Retrieve}(Q, B, \mathbb{X}), Y = \text{Generate}(Q, B, \mathbb{R}), \tag{11}$$

where $\mathbb{R}$ denotes the set of retrieved articles, Retrieve is typically a dense or sparse retriever, and Generate is an LLM that outputs the probability estimation. In this setting, there is no explicit multi-step reasoning over retrieval results; the model must encode all reasoning based solely on $\mathbb{R}$. Prompt is in the Appendix prompt (g).

**Agentic RAG** Agentic RAG extends Naive RAG by introducing an *agent* that can iteratively retrieve, reason, and act on intermediate conclusions, often following the ReAct paradigm [33]. In ReAct, the model alternates between reasoning steps $\tau$, action steps $\alpha$ (tool invocation for retrieval), and receiving observation $o$ of the action, enabling multi-step evidence gathering before producing its final probability estimation:

$$(Q, \tau_1, \alpha_1, o_1, \tau_2, \alpha_2, o_2, \ldots, \tau_L, \alpha_L, o_L) \tag{12}$$

All agents adopt the ReAct framework. Each agent is equipped with the same retrieval tool, prompts, and interaction structure, with the only variation being the underlying LLM. The retrieval tool implements document retrieval: the agent issues a natural language query, which is converted into a dense vector representation via a fixed embedding model, followed by cosine similarity search over the news corpus $\mathbb{X}$ to obtain the most relevant articles. Prompt is in the Appendix prompt (h). The tool is in the Appendix 6.

We show the evaluation results in Section 4.3 and 4.4. Then we soly evaluate the reasoning abilities of each model in Section 4.2 and analyze the reasoning performances along the forecasting timeline in Section 4.5. We finally demonstrate cases in Section 6.

### 4.1 Validity of CIL and PROPHET

To validate the effectiveness of our proposed CIL metrics, we design experiments to test their practical impact. The underlying intuition is that, if CIL is indeed effective, forecasts based on the articles with the highest CIL scores should significantly outperform forecasts obtained by directly answering the question without reference to these articles. To examine this, we conduct hypothesis tests to determine whether there is a statistically significant difference between the two approaches.

We partition the entire news corpus into two disjoint subsets according to the distribution of CIL scores. For each subset we evaluate two strategies: (1) **Top CIL**: retrieve the top-20 articles ranked by CIL score and feed them into the generator; (2)**without RAG**: let the generator condition solely on the query.
Every question is repeated 10 times to obtain stable estimates.

Let $X = (x_i)_{1\times 10n}$, where each $x_i$ is the Berrir score of a single response to one question under the *with-RAG* strategy; analogously, let $Y = (y_i)_{1\times 10n}$ denote the scores produced *without-RAG*. since every question is repeated 10 times, the vector contains $10n$ such individual scores. Define the paired differences $Z = (x_i - y_i)_{1\times 10n}$. We conduct one-sample $t$-tests on $Z$ with significance level $\alpha = 0.05$ under two one-sided null hypotheses:

$$\mathcal{H}_0^{\leq}: \mu(Z) \leq 0.077 \qquad \text{and} \qquad \mathcal{H}_0^{\geq}: \mu(Z) \geq 0.027.$$

Rejecting $\mathcal{H}_0^{\leq}$ implies that the average Berrir score of *with-RAG* significantly exceeds that of *without-RAG* by at least 0.077, i.e. the retrieved articles are beneficial for answering the question. Rejecting $\mathcal{H}_0^{\geq}$ implies that the improvement is *no more than* 0.027, i.e. the articles provide negligible gain.