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

Zhengwei Tao
tttzw@pku.edu.cn
Peking University
China

Pu Wu
puwu1997@126.com
Peking University
China

Zhi Jin†
zhijin@whu.edu.cn
Wuhan University
China

Xiaoying Bai†
baixy@aibd.ac.cn
AIBD
China

Haiyan Zhao
zhhy@sei.pku.edu.cn
Peking University
China

Chengfeng Dou
chengfengdou@pku.edu.cn
Peking University
China

Xiancai Chen
xiancaich@stu.pku.edu.cn
Peking University
China

Jia Li
jia.li@whu.edu.cn
Wuhan University
China

Linyu Li
linyuli@stu.pku.edu.cn
Peking University
China

Chongyang Tao
chongyang@buaa.edu.cn
Beihang University
China

Wentao Zhang
wentao.zhang@pku.edu.cn
Peking University
China

Abstract

Predicting future events based on news on the Web 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 downloaded from the Web. However, because there is no consideration of 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 future forecasting. Through extensive experiments, we first demonstrate the validity of CIL and in-depth investigations into future forecasting with the aid of CIL. Subsequently, we evaluate several representative prediction methods on PROPHET. The overall results draws valuable insights for task of future directions. Data is public on https://github.com/TZWwww/PROPHET.

1 Introduction

The quest to forecast future events based on information on the Web has long been a central pursuit in the field of artificial intelligence (AI). The ability to foresee outcomes and trends holds the promise of revolutionizing numerous sectors covering finance [15], climate science [31], and social policy [23]. 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 [9, 12, 21].

To evaluate the abilities of these LLM-based future forecasting systems, pilot works construct several benchmarks based on real-world forecasting questions [8, 9, 13, 35]. 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 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 knowledge-intensive benchmarks such as HotpotQA [32] and 2WikiMultiHopQA [11]. 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 [37]. 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 [1].

†Corresponding authors.

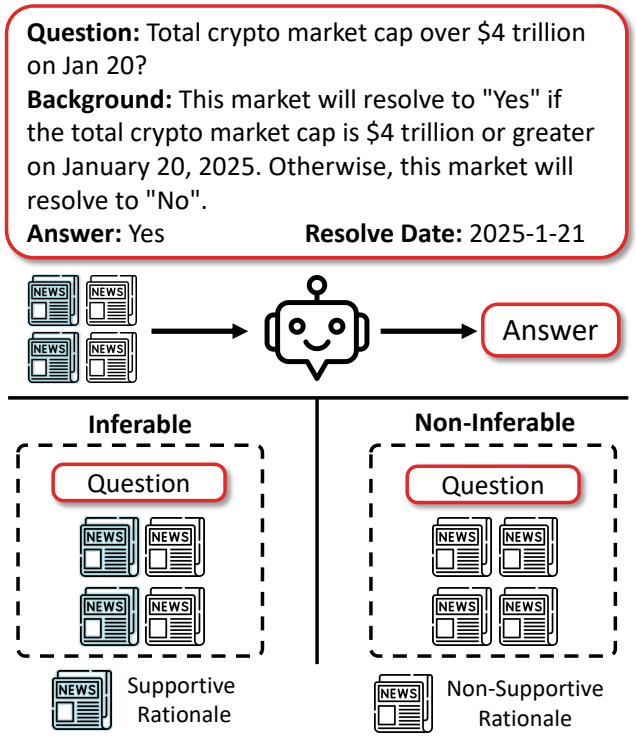

Figure 1: The upper Figure demonstrates the task of future forecasting. The lower half shows both inferable and non-inferable scenarios.

To overcome this challenge, 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 real-world prediction question pairs with relevant downloaded news articles for retrieval from the Web. 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 CIL to ensure the inferability of the benchmark, providing a fair and accurate evaluation of the systems' forecasting ability. Assisted by CIL, PROPHET performs as a well-formulated RAG-and-reasoning task with hidden rationale [37].

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. We provide a feasible method for calculating this metric, which we also testify 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 based on their related news 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 Preliminaries

2.1 Future Forecasting

Future forecasting stands for predicting whether a certain event will happen in the future based on news information from the Web. We now formalize the task as a binary question-answering task. Given a prediction question 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 B that describes the context of Q. A large set of news articles $\mathbb{X}$ serves as a knowledge base to retrieve. 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 of how likely the event in Q would occur. A ground truth answer $\hat{Y} \in \{0, 1\}$ paired with a resolved date D represents whether the event in Q finally occurs and the date the question resolves. As the same in previous works [9, 13], we use Brier Score [2] as the metric 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 in the dataset. In this work, we report the Brier Score expanded by 100 times and keep 2 decimal places.

We formalize future forecasting as an RAG task. As an RAG, it features distinctly compared with traditional dataset such as HotpotQA [32] and 2WikiMultiHopQA [11]. The knowledge base $\mathbb{X}$ stores the rationales and clues for answering Q [37]. 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 D [8, 9, 13] 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.

2.2 Causal Inference

Causal inference is a vital statistical method to determine causal relationships between variables [20]. 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 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 the 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 [26].

Interventional distribution An SCM allows the study of interventions. An atomic intervention $\text{do}(X_i = x)$ fixes 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 the structural causal model, interventions can be understand as changing of the original structure and variable distributions. After $\text{do}(X_i = x)$, the resulting distribution is $P(\cdot|\text{do}(X_i = x)) \doteq P_m(\cdot|X_i = x)$, which shows how other variables respond.

A critical distinction in causal inference lies between interventional and observational probabilities. The observational probability, denoted as the conditional probability $P(Y|X = x)$, represents the likelihood of event Y given that we have passively observed variable X to be x. This probability captures mere statistical association or correlation. In contrast, the interventional probability, denoted as $P(Y|\text{do}(X = x))$, represents the likelihood of Y if we were to actively intervene and set the value of X to x. This distribution describes the causal effect of X on Y by simulating a controlled experiment and removing spurious correlations. While interventional probabilities are essential for predicting the effects of actions, they cannot be computed directly from passively collected data. Therefore, a primary challenge in causal inference is to determine whether and how these interventional probabilities can be calculated from available observational probabilities[20].

3 PROPHET Benchmark

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

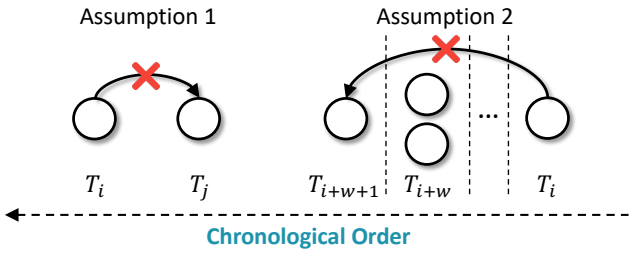

Figure 2: Illustration of assumptions. Nodes are variables that are in chronological order with time T.

3.1 Data Collection

Our objective is to gather a dataset that encompasses recent and prominent prediction questions. To achieve this, we have sourced questions from well-known platforms: Polymarket[1]. 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 the platform. 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. We aim to use the latest forecast questions. We include all questions on Polymarket resolved between 2025,1,1 and 2025,1,31. We filter 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 as much as possible. Using LLM, 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 6 (a-c). Then we searched on the MediaCloud open-source platform[2] with these queries. MediaCloud's vast news repository helped us gather comprehensive news URLs. After gathering the news URLs, we use Newspaper3k[3] API to download the news articles. We discard news that is later than the resolved date of the question.

However, many downloaded articles were irrelevant. There are also news articles tagged with wrong publish date, which could incur data leakage. To address these, we use LLM to rate the relevant score of each article to the question. We set the highest score, indicating this article directly answering of the question, while the lowest score showing total irrelevance. We remove both articles with the highest and lowest scores. The relevance rating prompt is in the Appendix 6 (d).

3.2 Causal Intervened Likelihood

To measure the sufficiency of supportive rationales for each question and to construct an inferable benchmark, we introduce a statistical estimation named **Causal Intervened Likelihood (CIL)**,

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