

LLM-as-a-Prophet: Understanding Predictive Intelligence with Prophet Arena *

Qingchuan Yang ^{1†}, Simon Mahns ^{2†}, Sida Li ^{3†}, Anri Gu ^{3†}, Jibang Wu ^{4‡}, Haifeng Xu ^{3‡}

¹ University of Southern California ² Meta ³ The University of Chicago ⁴ New York University

Abstract

Forecasting is not only a fundamental intellectual pursuit but also is of significant importance to societal systems such as finance and economics. With the rapid advances of large language models (LLMs) trained on Internet-scale data, it raises the promise of employing LLMs to forecast real-world future events, an emerging paradigm we call “LLM-as-a-Prophet”. This paper systematically investigates such predictive intelligence of LLMs. To this end, we build `Prophet Arena`, a general evaluation benchmark that continuously collects live forecasting tasks and decomposes each task into distinct pipeline stages, in order to support our controlled and large-scale experimentation. Our comprehensive evaluation reveals that many LLMs already exhibit impressive forecasting capabilities, reflected in, e.g., their small calibration errors, consistent prediction confidence and promising market returns. However, we also uncover key bottlenecks towards achieving superior predictive intelligence via LLM-as-a-Prophet, such as LLMs’ inaccurate event recalls, misunderstanding of data sources and slower information aggregation compared to markets when resolution nears.

1 Introduction

Forecasting is a fundamental intellectual pursuit that has shaped human progress from the earliest scientific inquiries to modern economics and finance. In machine learning, forecasting has also been a central theme, with rich traditions ranging from time-series analysis (Box & Jenkins, 1976; Lim & Zohren, 2021), online learning (Foster & Vohra, 1999), to conformal prediction (Barber et al., 2023). Yet, somewhat surprisingly, the challenge of open-domain forecasting, producing accurate predictions across a wide range of topics without domain-specific tuning or specialized datasets, remains largely unexplored. Achieving reliable foresight in this setting would represent a qualitative leap in AI capability, with far-reaching societal implications, from enhancing market efficiency to guiding high-stakes policy decisions (Arrow et al., 2008).

At its core, forecasting is the process of connecting present knowledge to anticipate future outcomes. Large language models (LLMs) seem natural candidates for this role. Trained on massive corpora of human knowledge through the seemingly narrow objective of next-word prediction, LLMs have developed emergent capabilities that extend far beyond their training objective (Bubeck et al., 2023). This motivates the hypothesis that the ability to predict the next word may also give rise to the ability to predict the next event. If substantiated, this would position LLMs not only as repositories of human knowledge but also as instruments of reliable foresight, leading to the prospect of **LLM-as-a-Prophet**: *Can AI systems reliably predict the future by connecting the dots across existing real-world information?*

In this paper, we seek to systematically examine the prospects and challenges of building general-purpose systems for open-domain forecasting. On one hand, forecasting is a natural next pursuit given the rapid progress

*Correspondence to contact@prophetarena.co.

Part of the work by Qingchuan Yang and Jibang Wu are done while they are at the University of Chicago. We thank Alex Gu, Chaplin Huang and Lucien Liu for help at the early stage of this work.

[†]Equal contribution.

[‡]Equal advising.

of AI, as it draws on a combination of advanced capabilities that current models are only beginning to demonstrate: information retrieval, complex reasoning and data analysis. Moreover, as many established benchmarks are approaching saturation and are increasingly prone to training-data contamination (Deng et al., 2024), open-domain forecasting provides a forward-looking and contamination-free setting with objectively measurable outcomes, making it a rigorous testbed for evaluating advanced model intelligence (Zou et al., 2022; Karger et al., 2025). On the other hand, we observe that current LLMs often struggle with key requirements for reliable foresight, including calibrated uncertainty estimation (Geng et al., 2023) and robust reasoning (Zhou et al., 2024) in the presence of noisy or incomplete evidence. As a result, their forecasting results may at times resemble guesswork rather than deliberated prediction, raising the possibility that fundamental barriers must be addressed before such evaluations can serve as a meaningful benchmark at the present time (Paleka et al., 2025a). Toward this end, we introduce *Prophet Arena*, a general framework for evaluating LLMs on live, real-world forecasting questions in a controlled and extensible way. Our goal is not only to assess the current forecasting performance of LLMs, but also to use forecasting as a lens for studying core components of intelligence, including reasoning, calibration, evidence aggregation. By doing so, we aim to identify which capabilities are emerging, which remain limited, and how forecasting evaluation can guide the development of more reliable predictive intelligence.

Organization of this paper. In §2, we present our design of *Prophet Arena*, a live and extensible benchmark for studying predictive intelligence. We define the key concepts and notation used throughout, detail the modular forecasting pipeline – from event and market extraction to probabilistic prediction and evaluation – and discuss the design principles that differentiate *Prophet Arena* from prior forecasting benchmarks. In §3, we introduce our three primary evaluation dimensions with their formal metrics and practical motivations, providing the foundation for rigorous and multi-faceted model assessment. In §4, we conduct an in-depth examination of LLM-as-a-Prophet, including mechanistic studies on knowledge internalization, context construction, and reasoning synthesis. We conclude in §5 with a synthesis of key findings, limitations, and future research directions toward more reliable predictive intelligence.

Summary of Our Contributions. We summarize the main contributions of this work as follows:

- We introduce the notion of *LLM-as-a-Prophet*, framing open-domain forecasting of real-world events as the next grand challenge of language models’ intelligence.
- We develop *Prophet Arena*, a live and extensible benchmark that continuously collects live real-world forecasting tasks across diverse domains. The framework decomposes forecasting into distinct stages in order to support controlled evaluations and incorporates multiple scoring metrics, capturing statistical accuracy, calibration, and economic value, in order to provide a comprehensive view of forecasts’ quality.
- We conduct thorough evaluations on state-of-the-art LLMs across 1300+ resolved real-world events, and perform comprehensive experiments to highlight several caveats in LLM forecasting that may have been overlooked in previous studies. Our results reveal both emerging strengths – such as calibrated uncertainty and reasoning alignment – as well as persistent bottlenecks in information aggregation and foresight near event resolution.

1.1 Connection and Comparison to Previous Works.

Understanding LLMs’ Forecasting Capabilities. The central goal of our work is to understand the novel paradigm of *LLM-as-a-Prophet* and analyze how different capabilities (e.g., knowledge internalization, source usage, etc.) affect LLMs’ predictive intelligence. This research goal shares similarity to a few recent works on understanding and diagnosing *special aspects* of LLMs’ forecasting capabilities. For instance, (Dai et al., 2025; Zhu et al., 2025) study LLMs’ *temporal generalization* capability by challenging LLMs to forecast future events curated from news articles. Both works show that LLMs’ forecasting accuracy degrades over time, even when they are armed with retrieval augmented generation (RAG). Paleka et al. (2025b) study whether LLMs can make consistent forecasts. For example, a logical AI should not predict that both the Democratic and Republican parties have 60% probability of winning the 2024 US presidential election. Towards that end, they build a proper-scoring-rule forecasting benchmark to measure the consistency of LLMs’ predictions. Similar to both works’ research methodology, we also build a benchmark platform *Prophet Arena* to study our research question. However, to our knowledge, our work is the first to investigate the general predictive intelligence of the LLM-as-a-Prophet paradigm.

Forecasting Benchmarks. Forecasting has recently become a popular challenge for benchmarking LLMs’ capabilities. Besides being a real challenge to LLM, it also avoids the thorny issue of benchmark contamination due to the evaluating LLMs on “future” events (Dai et al., 2025; Karger et al., 2025). To our knowledge, (Jin

et al., 2021) is perhaps one of the earliest to test such forecasting capabilities of language models. They introduced ForecastQA, an evaluation dataset consisting of 10,392 crowdsourced multiple-choice questions, and the performance of language models at the time (mainly BERT models) still significantly lags behind human performance. Since (Jin et al., 2021), there has been a progressive line of work developing more and more challenging forecasting benchmarks for more advanced models, by integrating prediction market events (Zou et al., 2022), extracting events from news articles (Zhang et al., 2024; Wang et al., 2025), curating future-oriented questions from websites (Wildman et al., 2025), using open-ended queries (Guan et al., 2024), and lately developing dynamic benchmarks with live event and leaderboard updates (Zeng et al., 2025; Karger et al., 2025; Bianchi et al., 2025).

We also build a benchmark Prophet Arena, but as mentioned above, the main goal of our work is to understand LLM-as-a-prophet and analyze how its different capabilities affect the forecasting. We built Prophet Arena mainly for this analysis purpose, similar to that of (Dai et al., 2025; Paleka et al., 2025b). Towards this end, our work departs from these benchmark researches in multiple ways. First, we focus on comprehensive evaluation about LLMs’ predictive intelligence from various dimensions such as the quality, trustworthiness and economic values of forecasts. In contrast, the works above mainly focus on curating the right dataset but do not dive deep into evaluation metrics. In fact, almost all the aforementioned works only report forecasting accuracy, which cannot fully capture the quality of forecasting (as we illustrate in Section 3). The only exception is (Karger et al., 2025), which uses Brier scores (one of the three metrics we examine). Second, because our goal is not to rank LLMs or agents as in the above benchmark papers but rather to understand frontier LLMs’ forecasting capabilities, we need a natural baseline forecaster to know how good frontier LLMs are. Towards that end, we build Prophet Arena based on events from prediction markets, and hence can use the market’s forecasts as a cost-effective and justified baseline forecaster (Arrow et al., 2008). Finally, besides reporting different evaluation metrics, we further dive into the LLM-as-a-Prophet paradigm and offer insights about how the forecasts are affected by different capabilities such as knowledge internalization and source usage.

2 Prophet Arena: A Live Benchmark for Predictive Intelligence

2.1 Key Definitions and Notations

(*Note: To aid exposition, we borrow the terminology of prediction markets for the definitions below. The general concept of forecasting, however, extends beyond market-based contexts to any form of future prediction.*)

Event. Let $\{E_i\}_{i=1}^K$ denote the set of forecasting *events*. An *event* is the overarching question or subject concerning a future real-world occurrence. It serves as a high-level container for one or more tradable *markets*. We consider an event **not** as a tradable asset; rather, it sets the context, scope, and resolution criteria for the markets that fall under it.

- **Example E_1 :** “Who will win the 2025-26 NBA Championship?”
- **Example E_2 :** “Which individuals will President Trump officially meet in 2025?”

Market. Each event E_i contains one or more (binary) *markets* $\{M_{ij}\}_{j=1}^{N_i}$. A *market* is a specific, tradable proposition under an event that will ultimately resolve to either **Yes** (True) or **No** (False). It represents a verifiable and unambiguous instantiation of the events overarching question.

- **Example market under E_1 :** “The Boston Celtics will win the 2025 NBA Championship.”
- **Example market under E_2 :** “President Trump will officially meet with Emmanuel Macron.”

Markets within the same event may be either *mutually exclusive* (e.g., only one NBA team can win the championship) or *non-exclusive* (e.g., a president may meet multiple individuals, or none). The only requirement is that each market be binary, thereby defining a well-posed resolution outcome.

Event Resolution. An event E_i is said to *resolve* at time τ_i , when all of its markets M_{ij} have their outcomes $o_{ij} \in \{0, 1\}$ determined, where $o_{ij} = 1$ indicates that market M_{ij} resolved to **Yes**.⁴

Contract. For a given market M_{ij} , a *Yes contract* q_{ij}^Y is a binary random variable that pays out 1 if $o_{ij} = 1$ and 0 otherwise. A corresponding *No contract* q_{ij}^N is defined analogously, paying out when the **Yes** contract does

⁴In practice, different markets within an event may resolve at different times; we take τ_i to be the latest resolution time among them.

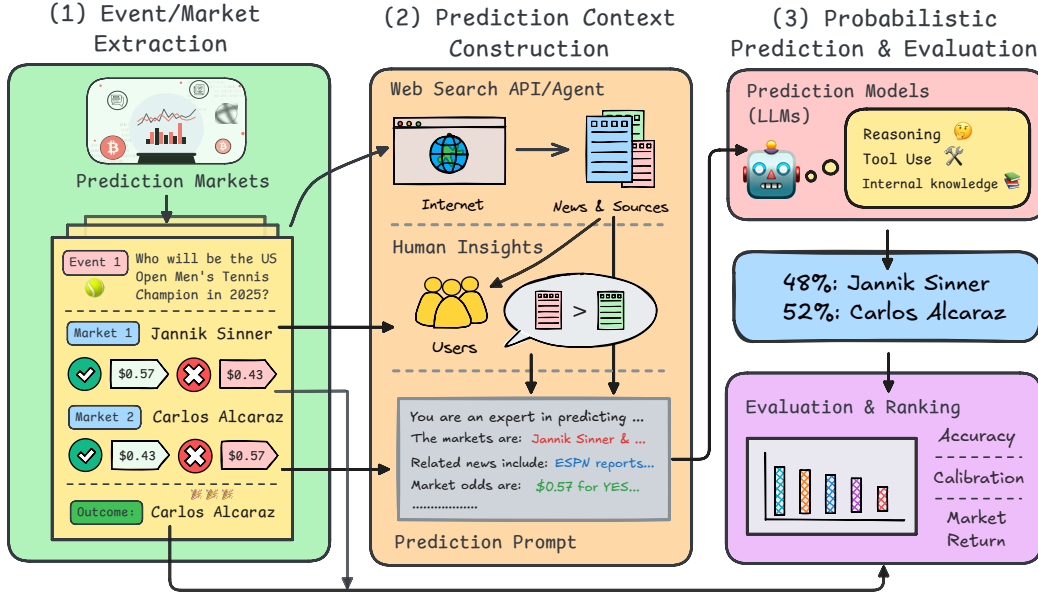


Figure 1: **Overview of the Prophet Arena pipeline.** The outcomes of events are unknown at extraction time. LLM predictions are generated at pre-scheduled forecast times, and evaluation occurs only after the events resolve, comparing model forecasts against the realized outcomes.

not. Each contract comes with a price/value $q_{ij} \in [0, 1]$, representing its market price or implied probability; this will be formally defined in §3.

(Hereafter, when the event index i is clear from context, we drop the subscript i and write M_j, o_j .)

2.2 The Full Prophet Arena Pipeline

Prophet Arena is implemented as a live, continuously updated pipeline for evaluating forecasts on real-world events. As illustrated in Fig. 1, it consists of three main stages – (1) event and market extraction, (2) prediction context construction, and (3) probabilistic forecasting with evaluation – which together form an end-to-end workflow for assessing predictive intelligence at scale.

(1) Event and Market Extraction. We continuously collect forecasting events from active prediction markets across diverse domains such as politics, economics, sports, entertainment, and science. Events are selected based on their popularity, diversity, and recurrence patterns, ensuring both temporal and topical coverage. Each day, 20 new events are added to our event pool. For the evaluations in this paper, we apply a cutoff date of October 11, 2025, and include only events that resolved prior to this date. An additional filter is applied to rule out predictions that are too close to the event resolution time (we explain the filtering detail in §3.2). The resulting dataset comprises 1,367 resolved events encompassing 72,136 markets. The use of live markets guarantees that all questions pertain to genuine future outcomes free from training-data contamination and that their resolutions can be objectively verified once finalized.

(2) Prediction Context Construction. Once an event E_i is extracted, Prophet Arena schedules a series of *prediction times* $T_i = \{t_i^{(1)}, t_i^{(2)}, \dots\}$ before the event resolves, enabling models to update their forecasts as market conditions and public information evolve. Accordingly, both the model-predicted probabilities and contract prices are time-dependent (i.e. our notation is actually time-dependent, e.g. $p_{ij} \equiv p_{ij}^{(t)}$). For clarity of exposition, however, we assume in the main text that each model issues a single forecast per event (i.e. $|T_i| = 1$); details on the full scheduling algorithm and multi-horizon aggregation are provided in §A.2.

For each event, Prophet Arena constructs a unified *prediction context* C_i that all models receive identically. This context contains:

- **Relevant information sources** retrieved by an LLM-based search agent using web queries that collect titles, snippets, timestamps, and URLs of recent news or reports; and
- **Market snapshots** including the latest Yes/No contract prices and trading volumes, from which implied probabilities (i.e. market contract prices $q_{ij} \in [0, 1]$) are derived.

Providing identical contexts isolates differences in models reasoning and calibration rather than their retrieval capabilities. The search component in `Prophet Arena` is fully *searcher-agnostic*: new search agents can be added or replaced without altering the forecasting protocol or evaluation procedure. In all experiments presented in this paper, we use a single LLM-based searcher instantiated with GPT-4o equipped with web access.

(3) Probabilistic Forecasting and Evaluation. Given an event E_i and its constructed context C_i , each model produces a probabilistic forecast for every market M_{ij} within the event. The model outputs a predicted probability $p_{ij} \in [0, 1]$, interpreted as its belief that the market will resolve to `Yes`, together with a short natural-language rationale. All forecasts are logged for later analysis but only the probabilities are used for quantitative evaluation. Once the event resolves and the true outcomes o_{ij} become available, `Prophet Arena` evaluates the forecasts along multiple complementary dimensions – which are detailed in §3. Together with the event extraction and context construction stages, this final step completes the `Prophet Arena` pipeline – turning live, real-world questions into a continuous, contamination-resistant benchmark for predictive intelligence.

(*) Bonus feature: Market Baseline. To establish a fair and interpretable anchor for comparing different LLM forecasters, we include a *Market Baseline* – a synthetic forecaster whose prediction (p_{ij}^{mb}) is defined as the market-consensus probability that market (M_{ij}) will resolve to `Yes`. In practice, this probability is inferred from the normalized contract prices. For instance, if the `Yes` and `No` contract prices are ($q_{ij}^Y \equiv q_{ij} = 0.8$) and ($q_{ij}^N \equiv 1 - q_{ij} = 0.2$), respectively, then the market baseline assigns an 80% probability to the `Yes` outcome (i.e. $p_{ij}^{\text{mb}} \equiv q_{ij}^Y$).⁵ As we will demonstrate in §3, this market baseline serves as an informative benchmark for assessing *forecast difficulty* and contextualizing model performance. When LLMs outperform the market baseline, they demonstrate genuine predictive advantage over the aggregated human consensus in the real-world market.

2.3 Design Choices of `Prophet Arena` and Differences from Other Forecasting Benchmarks

Because our goal is to thoroughly evaluate what capabilities affect LLMs’ predictive intelligence, `Prophet Arena` is built with different design choices compared to several recent forecasting benchmarks, which are designed primarily to rank LLMs or agents based solely on forecasting accuracy. We further illustrate these differences in Table 1, whereas highlight the four key design choices of `Prophet Arena` below.

1. **Probabilistic forecasts:** Future events are intrinsically random, so we elicit each market’s probabilistic forecast from each model, rather than a single choice of the most likely outcome. Notably, probabilistic forecasting is also the standard for real-world forecasting platforms in general, including leading ones such as [Metaculus \(2015\)](#) and [Good-Judgement-Open \(2015\)](#).
2. **Multi-horizon protocol:** For each event, the models are asked to make prediction across various timestamps before the event resolves. This enables temporal analyses of information assimilation and model behaviors.
3. **Modularized forecasting pipeline:** The process of prediction is broken down from source collection to probability elicitation to market actionability. This enables a comprehensive and controlled study of LLMs’ forecasting performance at different modules.
4. **Market return metrics:** `Prophet Arena` allows the evaluation of market profitability of each forecast, measuring models’ *relative* advantage over market consensus.

Benchmark	Live events	Probabilistic	Multi-horizon	Modularized	Return metrics
MIRAI	–	–	✓	–	–
FORECASTBENCH	✓	✓	✓	–	–
FUTUREBENCH	✓	–	–	–	–
FUTUREX	✓	–	–	–	–
Prophet Arena	✓	✓	✓	✓	✓

Table 1: **Comparison between `Prophet Arena` and related forecasting benchmarks.** We compare against MIRAI (Ye et al., 2024), ForecastBench (Karger et al., 2025), FutureBench (Bianchi et al., 2025), and FutureX (Zeng et al., 2025).

⁵In practice, contract prices may not sum exactly to one because of exchange transaction fees, requiring slight normalization.

3 Evaluating Forecasts: Caveats, Metrics, and Results

The golden metric in machine learning for evaluating prediction of binary (or multiary) outcomes is the *accuracy* measure (i.e., 0-1 loss). Indeed, several recent studies on benchmarking LLMs’ forecasting capabilities (e.g., Zeng et al. (2025); Karger et al. (2025); Wildman et al. (2025)) have used models’ accuracy as an indicator of their forecasting capabilities. These studies ask LLMs to predict a deterministic outcome and then evaluate this prediction’s accuracy (i.e., 0-1 loss). As a first caveat, we point out the limitation of measuring probabilistic forecasts via the accuracy metric. This limitation is due to the random nature of future events, which make them intrinsically different from typical classification tasks (e.g., recognizing whether a cat is in an image or not) with a deterministic answer. This is also why real-world forecasting platforms, including leading ones such as Metaculus (2015) and Good-Judgement-Open (2015), almost always elicit probabilistic forecasting.⁶

Caveat 1: Accuracy cannot fully measure the quality of probabilistic forecasts.

To see this, consider a binary random event E with groundtruth probability $p^* = 0.6$. Suppose Alice has a perfect forecast of the probability as $p^A = 0.6$ whereas Bob’s forecast is $p^B = 1$. To evaluate their predictions’ accuracy, each forecaster must give a predicted *outcome* $o \in \{0, 1\}$. In this case, it is natural for both to predict $o = 1$ (as they both believe $o = 1$ is more probable), which yields the same accuracy $\mathbb{E}_{o \sim p^*} \Pr(o = 1) = 0.6$. Therefore, the accuracy metric fails to distinguish Alice’s perfect forecast of the event’s probability from Bob’s forecast, which is extremal and not precise. The intrinsic reason of this failure is that using accuracy alone ignores prediction confidence, which is reflected in forecasted probabilities. In this example, Bob is overly confident about the outcome $o = 1$ compared to Alice, despite having the same accuracy as her. Nevertheless, correctly measuring the confidence of a forecast is crucial for downstream decision making tasks, such as risk control in financial markets (Rigotti & Shannon, 2005).

3.1 Evaluating Forecasting from Three Dimensions.

Given the **Caveat 1** above, we identify three different dimensions of forecasting evaluation in order to obtain a comprehensive understanding of LLMs’ forecasting capabilities. These three dimensions are (1) *forecasting loss* (Gneiting & Raftery, 2007; Brier, 1950) which measures the *absolute* quality of a probabilistic prediction; (2) *calibration error* (DeGroot & Fienberg, 1983; Guo et al., 2017) which measures trustworthiness (formally, statistical consistency) of a probabilistic prediction; and (3) *market return* (Mallikarjuna & Rao, 2019) which measures the *relative* advantage over current market’s consensus.⁷ As we illustrate below, these dimensions capture fundamentally different aspects of a forecast, enabling us to build a comprehensive understanding of LLMs predictive intelligence. For each dimension, we choose the most standard metrics.

3.1.1 Scoring Rules to Measure Forecasting Loss

The standard approach for evaluating probabilistic forecasts is through *proper scoring rules* (Gneiting & Raftery, 2007), which quantifies the discrepancy between predicted probabilities and realized outcomes. A scoring rule is *proper* if it ensures that the perfect probabilistic forecast achieves the optimal expected score. Prophet Arena adopts one of the most popular proper scoring rules, the **Brier score** (Brier, 1950), which is defined for each event E_i as⁸

$$BS_i := \frac{1}{m_i} \sum_{j=1}^{m_i} (p_{ij} - o_{ij})^2. \quad (1)$$

Therefore, the overall Brier score of an LLM is the average across all n events, $BS := \frac{1}{n} \sum_{i=1}^n BS_i$. The Brier score is the mean squared error for probabilistic forecasts, hence smaller scores indicate better forecasts. It is easy to verify that Brier score separates the quality of Alice’s and Bob’s forecasts in the above example: Bob’s expected Brier score is $\mathbb{E}_{o \sim p^*} (p^B - o)^2 = 0.4$, whereas Alice’s is $\mathbb{E}_{o \sim p^*} (p^A - o)^2 = 0.24$.

⁶Some classification algorithms, such as logistic regression, also compute probabilities. However, unlike forecasting, such probabilities are not a requirement on the output but rather for the sake of algorithm design. Indeed, these algorithms often convert the probability to a deterministic output as its final prediction.

⁷Market consensus about an event’s forecast is generally difficult to obtain. However, a good proxy of such data is available for Prophet Arena as it fetches forecasting events from prediction markets, which are widely believed to offer good approximation of the market’s consensus on the event’s probability (Arrow et al., 2008; Berg et al., 2008).

⁸Careful readers might notice that definition given here has some subtle difference from a “textbook version.” We delay a detailed explanation to §B.1.

3.1.2 Calibration Errors to Measure Trustworthiness

In addition to achieving low Brier scores, a good forecast should also be *trustworthy* – that is, when it assigns probability \tilde{p} to `Yes`, the event should indeed occur about \tilde{p} fraction of the time. This notion of trustworthiness is conventionally studied in the machine learning literature as *calibration* (DeGroot & Fienberg, 1983; Guo et al., 2017; Kalai & Vempala, 2024).

Formally, in our binary market setting, let $M = \{1, \dots, m\}$ denote the set of markets. Suppose a forecaster provides predicted probabilities \tilde{p}_k for all markets in a set $M_k \subset M$ ($m_k = |M_k|$ denotes its cardinality). Then the **expected calibration error (ECE)** of this forecaster is defined as

$$ECE = \frac{1}{m} \sum_k \left| \sum_{j \in M_k} \mathbb{P}(o_j = 1 | p_j = \tilde{p}_k) - m_k \tilde{p}_k \right|. \quad (2)$$

Intuitively, ECE captures how much a probabilistic forecast differs from the real averaged probability, given this forecast level. A lower ECE means the forecast is more trustable, though it does not necessarily imply small forecasting loss. This difference is reflected in our LLM evaluations below, whereas an illustrative mathematical example can be found in §B.2.

In practice, however, exactly repeated predictions are rare, so computing $\mathbb{P}(o_j = 1 | p_j = \tilde{p}_i)$ in Eq. (2) is infeasible. A standard solution – used throughout the applied literature – is to approximate ECE by partitioning predictions into probability bins, and comparing each bins empirical accuracy to its average predicted probability. This binned version is what most prior work simply calls “ECE.” Since the implementation is routine and well-known, we defer the details to §B.3.

3.1.3 Market Return to Measure Economic Values

Prophet Arena uses events from real-world prediction markets, hence allows us to evaluate the economic value of a forecast based on current market prices. We thus introduce *Average Return* as a natural metric to capture *how profitable it would be to trade in the market using LLM forecasts?* Formally, consider a (binary) market with market price q^Y per share for the `Yes` contract and q^N for the `No` contract. Given forecasted probability p for an `Yes` outcome, we use the following natural betting strategy: allocate a unit budget (\$1) to buy $1/q^Y$ shares of `Yes` contracts if $\frac{p}{q^Y} \geq \frac{1-p}{q^N}$, or to buy $1/q^N$ shares of `No` contracts otherwise. After the event resolution, if the bought contracts match the event outcome, the return is simply the number of these contracts; otherwise, the return is 0. The **Average Return** of an LLM is then defined as the average of these returns across all markets, under a unit budget allocated to each market.

Remark: Three notes on market return.

First, the betting strategy above can be shown to maximize expected return under risk neutrality. We defer its proof to §B.5, where we describe the more general optimization framework for betting strategies under different risk-preferences.

Second, while both market return and Brier score assess forecast quality, they differ fundamentally. The Brier score is an *absolute* metric, measuring a forecasts closeness to the ground truth, independent of external factors like market prices. In contrast, market return is a *relative* metric, capturing how much a forecast outperforms the markets current belief (interpreted as the contract price (Wolfers & Zitzewitz, 2006)). This distinction is already evident in our evaluations. In §B.4, we provide a concrete example to illustrate that forecasts with worse Brier scores can achieve higher market returns.

Third, calibration is not related to the total market return, but is indicative about how balanced the market return is from betting on the `Yes` contracts and `No` contracts. To formalize this intuition, we prove in §B.6 that a well-calibrated and symmetric forecaster – intuitively, one that is not systematically more aggressive or conservative than the market – will have balanced returns from both contract types.

3.2 Evaluation Results and Analysis Across Different Dimensions

3.2.1 Caveats from Temporal Analysis of Forecasting Results

Before presenting our main evaluation results, we first examine how predictive accuracy varies with the time remaining before event resolution. As shown in Fig. 2, the average Brier score is plotted across lead-time

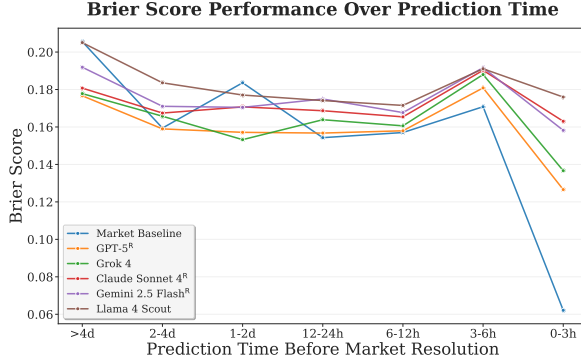


Figure 2: **Brier score as a function of prediction time before resolution.** Predictions made closer to the market resolution time tend to be better in Brier score, as additional information becomes available. While some LLMs outperform the market baseline at longer horizons, the market adjusts rapidly to new signals and achieves the highest short-term accuracy near event resolution.

intervals, where, for instance, the “0-3 h” bin represents predictions made within three hours before resolution. For both the market baseline and representative LLMs, accuracy generally improves as the resolution time approaches – reflecting that additional information becomes available and predictive signals strengthen.⁹ Interestingly, the market baseline lags behind several frontier LLMs when predictions are made far in advance, suggesting that LLMs can effectively synthesize broader prior knowledge and reason under noisy settings at long horizons. However, as resolution nears, markets incorporate breaking information and news updates more rapidly than LLMs, quickly surpassing LLMs in short-term accuracy. *This observation motivates two methodological takeaways that inform our evaluation design:*

Caveat 2: Raw Brier scores can be misleading, and having a baseline is instrumental in understanding predictive intelligence.

Predictive difficulty varies drastically with lead time: the same event may be highly uncertain a week in advance but almost deterministic hours before resolution. Hence, comparing raw Brier scores across horizons conflates forecasting skill with intrinsic event difficulty. Including the *market baseline* as a dynamic reference forecaster provides a meaningful normalization, offering a heuristic measurement of “how predictable an event is” at each point in time.

Caveat 3: Predictions too close to resolution should be excluded.

Near-resolution forecasts are dominated by real-time information access rather than reasoning ability. Because Prophet Arena holds the retrieval component fixed across models, such predictions no longer reflect intrinsic model capabilities. Consequently, we exclude all forecasts made within three hours before event resolution from subsequent evaluations.

3.2.2 Evaluation Results and Discussions

Throughout the main body of the paper, we highlight (the same) five representative LLMs¹⁰ out of the 23 evaluated in total (the full results are available in Table 5 of §C.1). As shown in Table 2, frontier proprietary models demonstrate similar Brier score performances as the the *Market Baseline*, and consistently outperform the later in terms of calibration and market return.

Notably, the relative rankings of models differ depending on which evaluation metric is used, illustrating the complementary perspectives offered by accuracy, calibration, and profitability. Concretely, Brier scores fall in a narrow band [0.18, 0.22] (for reference, pure random guess has expected Brier score 0.25). By contrast, calibration differences are more pronounced: strong models typically achieve $ECE \leq 0.05$, whereas weaker ones fall in the [0.06, 0.2] range. Nevertheless, all the selected LLMs demonstrate better calibration than the market baseline. For market performance, even GPT-5^R, the top-ranked model, fails to reach break-even (Average Return < 1), and most models fall below 0.9. Since event-level payoffs depend heavily on market-implied probabilities, the resulting returns exhibit substantial variance, as evidenced by the wide confidence

⁹In certain domains, such as live sports, markets may remain open during the event itself, allowing traders to incorporate real-time developments as outcomes unfold.

¹⁰Models are chosen to span proprietary and open-source families, reasoning and non-reasoning variants, and a range of performance levels in the full ranking.

LLM	Forecasting Loss		Calibration Error		Market Return	
	↓ Brier (95% CI)	Rank	↓ ECE	Rank	↑ Average (95% CI)	Rank
GPT-5 ^R \triangle	0.184 (\pm 0.006)	①	0.042	②	0.943 (\pm 0.042)	①
Grok-4 ^R \triangle	0.189 (\pm 0.005)	②	0.043	③	0.864 (\pm 0.052)	④
Claude Sonnet 4 ^R \triangle	0.194 (\pm 0.006)	③	0.041	①	0.909 (\pm 0.101)	②
Gemini 2.5 Flash ^R \triangle	0.197 (\pm 0.007)	④	0.067	⑤	0.883 (\pm 0.053)	③
Llama-4-Scout \triangle	0.219 (\pm 0.008)	⑤	0.060	④	0.805 (\pm 0.040)	⑤
Market Baseline	0.187 (\pm 0.006)	N/A	0.069	N/A	0.899 (\pm 0.043)	N/A

Table 2: **Evaluation of five representative LLMs.** For Brier and Average Return, bootstrapped 95% confidence intervals are reported. Superscript ^R is used to denote a reasoning model, with its reasoning configuration in § A.5. The full results for all 23 LLMs are provided in Table 5. Although our benchmark updates in real time, for the purpose of writing this paper we need to fix a dataset. Our evaluation is conducted on 1,367 events that were resolved before October 11, 2025.

intervals. In § C.2, we further discuss the *Sharpe ratio* (Sharpe, 1998) of our betting strategy, which normalizes Average Return by volatility, providing a more stable comparison of models’ economic performance.

Overall, our results suggest that absolute forecasting skill and relative profitability against prediction markets are still challenging for today’s LLMs. A more fine-grained investigation could help deepen our understanding about what could make an LLM a good prophet. Fig. 3 draws for the best and worst models (Left & right) their reliability diagrams regarding calibration errors (Guo et al., 2025), where predicted probability (x -axis) is compared against realized frequency (y -axis) at different probability bins. While their calibration is similar in intermediate ranges, the stronger model – OpenAI o3 – performs much better in the extreme bins (0-0.1 and 0.9-1.0), where it almost always predicts correctly. Because such extreme forecasts occur frequently, this advantage helps explain the gap in both Brier score and market return.

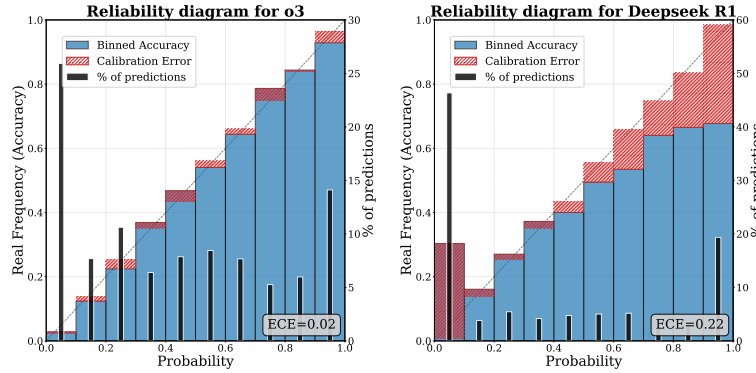


Figure 3: **Reliability diagrams for the best and worst LLMs ranked by calibration score (ECE).** The black histogram indicates fraction of predicted probabilities in each bin. The calibration error within each bin appears as the height of the red rectangle, i.e., the gap between accuracy and confidence. The reported ECE score (bottom right) corresponds to a weighted sum of these errors, weighted by the distribution. Superscript ^R denotes a reasoning model.

4 An In-Depth Analysis of LLM-as-a-Prophet

In this section, we deploy our tailored Prophet Arena framework to perform a systematic investigation into the emerging paradigm of LLM-as-a-Prophet. Due to the resource constraint, our experiment results in this section uses a subset of 100 events sampled uniformly from our full benchmark.¹¹ For clarity, this section prioritizes the presentation of interesting findings and novel insights, during which we shall also point the reader to corresponding appendices for more thorough evaluations and discussions.

We begin with a series of sanity checks to ensure that existing LLMs demonstrate reasonable level of understanding and reasoning capabilities to make forecasts. Our analysis proceeds from two complementary perspectives. The first adopts a mechanistic approach, designing controlled experiments to uncover the principles and failure modes that govern model behavior. The second follows a granular evaluation procedure, assessing how well current LLMs can reason about future events when embedded in realistic forecasting environments.

¹¹We make this dataset publicly available at <https://huggingface.co/datasets/prophetarena/Prophet-Arena-Subset-100>.

Together, these experiments establish an empirical foundation for understanding what it means for a model to “forecast”, and delineate the boundary of its current predictive intelligence.

4.1 Robustness and Consistency Checking

The basis to accomplish the general forecasting tasks is to have a good understanding of the question as well as the nature of probability. Hence, we examined two fundamental capabilities of existing LLM models: (i) *robustness of probability elicitation*, where calibration remains stable under prompt variations and alternative probability estimation methods, and (ii) *logical consistency*, where most LLMs correctly understand structures such as mutually exclusive or nested markets. For the majority of models, both capabilities appear already reliable and largely mature. Detailed results are provided in §C.3 and C.4.

4.2 Mechanistic Analysis of LLM-as-a-Prophet

We begin by taking a mechanistic view of LLM predictive intelligence, aiming to uncover how it arises from the interaction of distinct model capabilities, in aim to identify the bottlenecks and causal links that determine their effectiveness. To this end, we design experiments that examine the key factors behind strong forecasting performance, starting from a model’s internal knowledge (section 4.2.1), to the quality and accessibility of external sources (section 4.2.2), to its ability to integrate those sources effectively (section 4.2.3).

4.2.1 Can Internalized Knowledge Become Foresight?

Forecasting begins with what a model already knows — the *internalized* knowledge. A key question is whether models possess accurate representations of past outcomes and can effectively leverage this knowledge to inform present forecasts. To investigate this, we examine how well models recall and interpret historical events, where we retrieve 100 past events from Kalshi that occurred prior to each models knowledge cutoff date and evaluate their responses under the recall prompt described in §C.7.2.

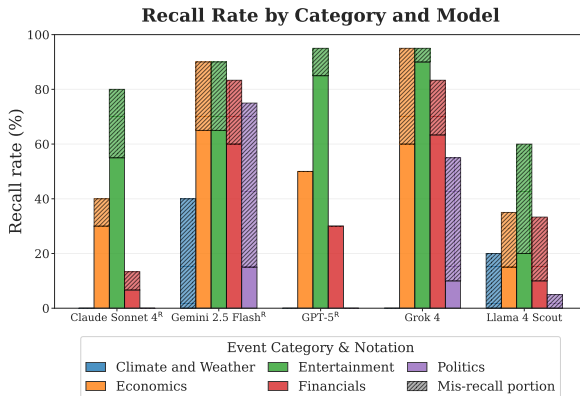


Figure 4: Recall rate by event category and model using the knowledge internalization recall prompt (§E.3.1). Shaded region represents the percentage of events reported to be recognized by the LLM but having a wrong recall.

Event recall varies by topics and models. In Fig. 4, we observe that models most reliably recall events in *Entertainment*. By contrast, *Climate and Weather* and *Politics* display low recall and frequent mis-recall. Two factors likely contribute. First, *Weather* prompts often require fine-grained, date-stamped facts (e.g., *Highest temperature in Miami on Aug 29, 2023?*), which are less likely to be memorized. Second, due to 2023 regulatory constraints on Kalshi’s election markets, *Politics* in our dataset skews toward politics-adjacent indicators (e.g., *“Biden 538 approval rating on Aug 30, 2023?”*, which varies daily and may be sparsely represented in training corpora (Inc. (2025))).

Despite these broad patterns, models differ in recall accuracy. For *Economics* and *Politics*, GPT-5 (High) correctly answered all events it claimed to recall. In contrast, models like Llama 4 Scout and Gemini 2.5 Flash (Reasoning) reported recognizing events in all categories. However, for categories like *Climate and Weather* as well as *Politics*, almost all of their recalled events were false.

Event recall is approximate, not precise. Consider the event “*Billboard Hot 100 #1, Jul 13, 2023?*”. Gemini 2.5 Flash recalled (§D.2) that Olivia Rodrigos *Vampire* displaced Morgan Wallens *Last Night* at number one, which is correct. However, it aligned the answer with the chart dated July 15, 2023, treating it as equivalent to July 13, 2023. Thus, while the model demonstrated knowledge of the outcome, it failed to recall the precise

alignment between dates and chart releases. This illustrates that event recall in LLMs is approximate: models often retain coarse associations (the correct song and transition) but lack fidelity on exact temporal details. An intriguing open question is to quantify the effect of inaccurate past knowledge internalization on the model’s forecasting capability in future events.

4.2.2 How Do Contexts Shape Forecasts?

Besides the internalized knowledge, making good forecasts also requires good context information — Prophet Arena retrieves relevant news sources from the Internet and provide the live market data from Kalshi. We first analyze how these different information sources affect model performance. Fig. 5 shows the average Brier scores across all tested LLMs under four conditions: access to both market data and news sources, market data only, news sources only, and none. The results demonstrate a clear performance hierarchy; as expected, models with access to both market data and news sources output the best predictions, while those without access to either source of information exhibit the poorest performance.

Beyond performance differences, the variance patterns reveal deeper insights. Interestingly, models using only market data perform only slightly worse, on average, than those with both market data and news sources; however, the key difference lies in the variability of predictions. Combining multiple high-quality sources substantially reduces variance in prediction quality, suggesting that sources still offer valuable signals and perspectives that enable more consistent forecasting. Thus, while market data is powerful precisely because it aggregates information from a plethora of sources and trends, adding a few carefully chosen high-quality sources can still help stabilize and refine the signals the information provides. §C.8 includes a deeper analysis of LLMs abilities to find and utilize high-quality sources.

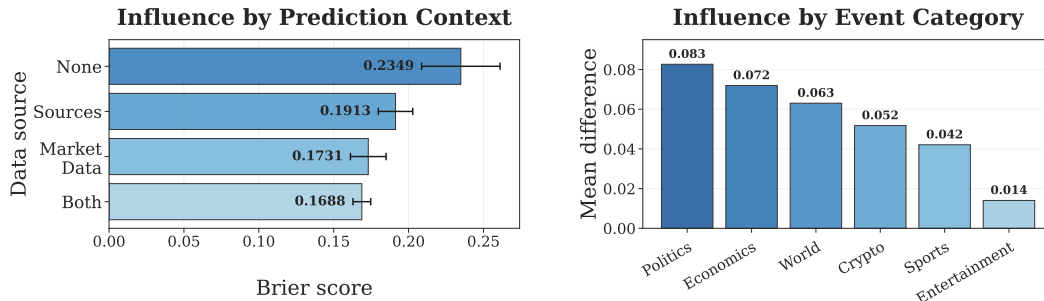


Figure 5: **Prediction quality across different contexts and event categories.** Left panel shows average Brier scores across evaluated LLMs for prompts with varying information availability (none, sources only, market data only, or both). Error bars represent the interquartile range (IQR) of model performance. Right panel displays mean Brier scores of events in each category.

Sources can clarify or confound predictions. As shown in Fig. 5, on average, adding sources improves mean Brier score. However, the effect is heterogeneous: not only does the magnitude of the effect vary based on event category, it also does not necessarily strictly improve prediction quality, as shown in the case study in §D.3. Regarding category differences, the benefit of adding sources is not uniform. In areas like politics, where events can be interpreted through multiple perspectives, incorporating information from varied outlets and institutions appears to add useful context. In contrast, in domains such as entertainment or sports, the marginal value of additional sources seems smaller. This highlights that more information is not necessarily better; the effectiveness of sources depends on their relevance to the prediction task.

4.2.3 How Models Engage with Sources

Having established that access to contextual information improves forecasting performance, we next examine how models actually engage with those sources. In particular, we study how models calibrate their beliefs in light of new evidence: do they tend to keep their prior belief or open to adjust in response to external signals? Do they amplify confidence when multiple sources are consistent, or moderate it when signals diverge?

Sources drive LLMs’ forecasts to be more conservative than markets. Fig. 6 compares model prediction probabilities to the market baseline’s probabilities on those markets resolved to Yes (i.e., the “winner”). Since market data is included in the LLM prompts — most models generally align closely with market predictions. However, across a large majority of events, LLMs consistently output more conservative probabilities.

A representative example is Llama 4 Scout, shown in Fig. 6d: even when markets assign near-certain probabilities (close to one), the model remains hesitant, rarely producing equally extreme predictions. This reflects a systematic conservatism where the model, when being fed with additional external sources, tend to underweight outcomes the market views as almost certain. As illustrated in Fig. 6a & Fig. 6b, although both GPT-5 and Grok 4 adopt a cautious approach in their probability assignments, they generally track the market more closely across the full probability range and avoid the same degree of reluctance. Claude Sonnet 4 exhibits a slightly more assertive pattern: in the mid-probability range (around 0.5-0.6), it occasionally assigns probabilities slightly above the market-implied probabilities. Nevertheless, across most events, it displays significant conservatism, and particularly at the higher end of the probability spectrum, Claude Sonnet 4 also is reluctant to place extreme probability predictions. Overall, while conservatism is a common trait across models, the extent of hesitation varies, with some models exhibiting considerably stronger reluctance than others. Similar trends have appeared in other models as well.

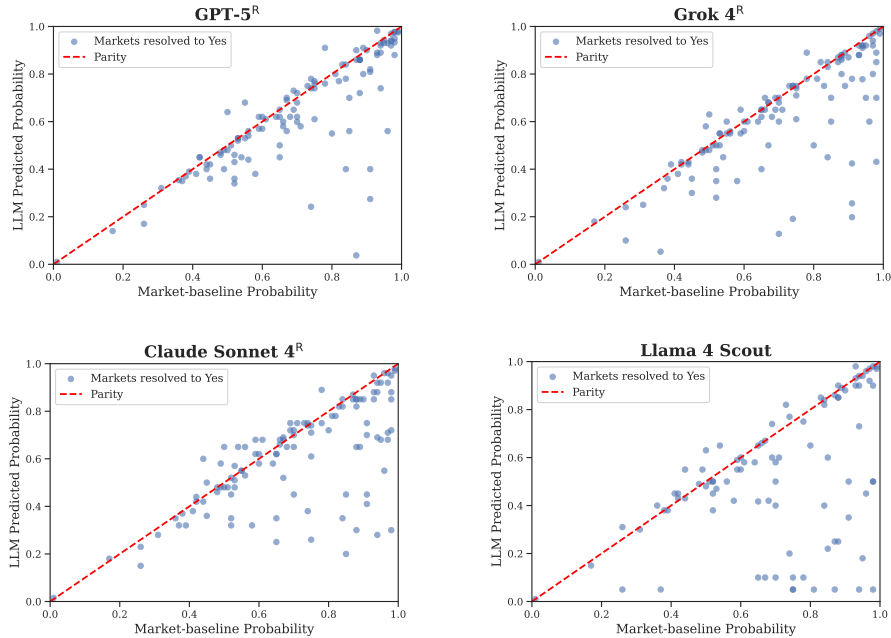


Figure 6: **LLM-predicted probabilities versus market baselines for markets resolved to Yes.** Each scatter plot compares model-predicted probabilities with market-implied probabilities. Points along the diagonal indicate cases where the models forecast exactly matches the markets implied probability of a Yes resolution.

4.3 Granular Analysis of LLM-as-a-Prophet

The reported probabilities of LLMs compress a rich decision-making process into a single number: two models may yield similar predicted probabilities and scores while relying on dramatically different reasoning processes. To address this, we open the black box and evaluate the process underlying each forecast. Specifically, we adopt an LLM-as-a-judge framework (Zheng et al., 2023) to assess the soundness of reasoning along five critical dimensions: source selection, evidence extraction, reasoning synthesis, reasoning-to-prediction alignment, and recognition of prediction uncertainty. Full prompts and additional details of the evaluation framework are provided in §C.6.

As shown in Table 3, the models demonstrate broadly comparable performance in source utilization, evidence extraction, and uncertainty analysis. However, substantial disparities emerge in the reasoning synthesis and reasoning-prediction alignment categories. For example, when comparing GPT-5 with Gemini 2.5 Flash, the differences in reasoning synthesis (0.95) and reasoning-to-prediction alignment (0.30) are markedly larger than the relatively minor gaps in source use (0.12), evidence extraction (0.00), and uncertainty analysis (0.20). Due to the significant deficiencies in those two key categories, the other models demonstrate a significant gap in prediction quality relative to GPT-5, supporting the findings in Table 2. These findings indicate a potential ceiling effect: once models attain proficiency in retrieval and evidence extraction, further performance gains depend primarily on advances in higher-order reasoning rather than incremental improvements in information access.

LLM	Sources	Evidence	Reas. Synth.	Align.	Uncert.	Average Score
GPT-5 ^R	3.69	3.66	4.14	3.97	3.94	3.88
Gemini 2.5 Flash ^R	3.57	3.66	3.19	3.67	3.74	3.57
Grok 4	3.40	3.51	3.33	3.48	3.66	3.48
Claude Sonnet 4 ^R	3.53	3.47	2.93	3.39	3.75	3.41
Llama 4 Scout	2.97	2.88	2.29	2.37	2.87	2.68

Table 3: **LLM performance on reasoning evaluation criteria across dataset events.** Each dimension is scored on a standardized 5-point scale, where 1 and 5 indicate poor and excellent performance, respectively. Average scores are presented for each model, with **bold** values indicating the best-performing model for each criterion. Models are ordered by descending overall average score.

5 Conclusion

This paper systematically evaluates the prospects and challenges of using LLMs to forecast future events, a paradigm coined LLM-as-a-Prophet. Towards that end, we build `Prophet Arena`, a benchmark that allows modularized analysis about various aspects of existing LLMs’ predictive intelligence. Our thorough experiments demonstrates the promise of LLM-as-a-Prophet reflected in the small forecasting loss, calibration error, strong reasoning synthesis and alignment of frontier models. However, we also identify key bottlenecks and highlight avenues for further progresses, such as better curation of context sources, more accurate internalization of knowledge and improving forecasts near events’ resolution.

References

- Anthropic. Introducing claude 4. <https://www.anthropic.com/news/claude-4>, May 2025.
- Kenneth J Arrow, Robert Forsythe, Michael Gorham, Robert Hahn, Robin Hanson, John O Ledyard, Saul Levmore, Robert Litan, Paul Milgrom, Forrest D Nelson, et al. The promise of prediction markets. *Science*, 320(5878):877–878, 2008.
- Rina Foygel Barber, Emmanuel J Candes, Aaditya Ramdas, and Ryan J Tibshirani. Conformal prediction beyond exchangeability. *The Annals of Statistics*, 51(2):816–845, 2023.
- Joyce Berg, Robert Forsythe, Forrest Nelson, and Thomas Rietz. Results from a dozen years of election futures markets research. *Handbook of experimental economics results*, 1:742–751, 2008.
- Federico Bianchi, Junlin Wang, Zain Hasan, Shang Zhu, Roy Yuan, Clementine Fourrier, and James Zou. Back to the future: Evaluating ai agents on predicting future events. <https://www.together.ai/blog/futurebench>, 2025. Online; accessed 10 September 2025.
- George EP Box and Gwilym M Jenkins. Time series analysis. forecasting and control. *Holden-Day Series in Time Series Analysis*, 1976.
- Glenn W Brier. Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1):1–3, 1950.
- Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023.
- Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.
- Hui Dai, Ryan Teehan, and Mengye Ren. Are llms prescient? a continuous evaluation using daily news as the oracle. In *Forty-second International Conference on Machine Learning (ICML)*, 2025.
- Morris H DeGroot and Stephen E Fienberg. The comparison and evaluation of forecasters. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 32(1-2):12–22, 1983.

- Chunyuan Deng, Yilun Zhao, Xiangru Tang, Mark Gerstein, and Arman Cohan. Investigating data contamination in modern benchmarks for large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 8698–8711, 2024.
- Thomas J DiCiccio and Bradley Efron. Bootstrap confidence intervals. *Statistical science*, 11(3):189–228, 1996.
- Dean P Foster and Rakesh Vohra. Regret in the on-line decision problem. *Games and Economic Behavior*, 29(1-2):7–35, 1999.
- Gemini Team. Introducing gemini 2.0: our new ai model for the agentic era. <https://blog.google/technology/google-deepmind/google-gemini-ai-update-december-2024>, Dec 2024. Accessed: 2025-09-22.
- Jiahui Geng, Fengyu Cai, Yuxia Wang, Heinz Koepl, Preslav Nakov, and Iryna Gurevych. A survey of confidence estimation and calibration in large language models. *arXiv preprint arXiv:2311.08298*, 2023.
- Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102(477):359–378, 2007.
- Good-Judgement-Open. Good judgement open. <https://www.gjopen.com/>, 2015. Online; accessed 10 September 2025.
- Yong Guan, Hao Peng, Xiaozhi Wang, Lei Hou, and Juanzi Li. Openep: Open-ended future event prediction. *arXiv preprint arXiv:2408.06578*, 2024.
- Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pp. 1321–1330. PMLR, 2017.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.
- Kalshi Inc. Kalshi - prediction market for trading the future. <https://www.kalshi.com/>, 2025.
- Woojeong Jin, Rahul Khanna, Suji Kim, Dong-Ho Lee, Fred Morstatter, Aram Galstyan, and Xiang Ren. Forecastqa: A question answering challenge for event forecasting with temporal text data. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 4636–4650, 2021.
- Adam Tauman Kalai and Santosh S Vempala. Calibrated language models must hallucinate. In *Proceedings of the 56th Annual ACM Symposium on Theory of Computing*, pp. 160–171, 2024.
- Ezra Karger, Houtan Bastani, Chen Yueh-Han, Zachary Jacobs, Danny Halawi, Fred Zhang, and Philip Tetlock. Forecastbench: A dynamic benchmark of ai forecasting capabilities. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025.
- Kimi Team. Kimi k2: Open agentic intelligence. *arXiv preprint arXiv:2507.20534*, 2025.
- Bryan Lim and Stefan Zohren. Time-series forecasting with deep learning: a survey. *Philosophical Transactions of the Royal Society A*, 379(2194):20200209, 2021.
- Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.
- Mejari Mallikarjuna and R Prabhakara Rao. Evaluation of forecasting methods from selected stock market returns. *Financial Innovation*, 5(1):40, 2019.
- Meta Team. The Llama 4 herd: The beginning of a new era of natively multimodal AI innovation. *AI at Meta*, April 2025. URL <https://ai.meta.com/blog/llama-4-multimodal-intelligence/>. Blog post.
- Metaculus. Metaculus: Forecasting for a complex world. <https://www.metaculus.com/>, 2015. Online; accessed 10 September 2025.

- OpenAI. Hello gpt4o. <https://openai.com/index/hello-gpt-4o/>, May 2024.
- OpenAI. Introducing gpt4.1 in the api. <https://openai.com/index/gpt-4-1/>, Apr 2025a.
- OpenAI. Introducing gpt-5. <https://openai.com/index/introducing-gpt-5-for-developers/>, Aug 2025b.
- OpenAI. Introducing openai o3 and o4-mini. <https://openai.com/index/introducing-o3-and-o4-mini/>, Apr 2025c.
- Daniel Paleka, Shashwat Goel, Jonas Geiping, and Florian Tramèr. Evaluating forecasting is more difficult than other llm evaluations. In *ICML 2025 Workshop on Assessing World Models*, 2025a.
- Daniel Paleka, Abhimanyu Pallavi Sudhir, Alejandro Alvarez, Vineeth Bhat, Adam Shen, Evan Wang, and Florian Tramèr. Consistency checks for language model forecasters. In *The Thirteenth International Conference on Learning Representations*, 2025b.
- Luca Rigotti and Chris Shannon. Uncertainty and risk in financial markets. *Econometrica*, 73(1):203–243, 2005.
- William F Sharpe. The sharpe ratio. *Streetwise—the Best of the Journal of Portfolio Management*, 3(3):169–85, 1998.
- Amir Taubenfeld, Tom Sheffer, Eran Ofek, Amir Feder, Ariel Goldstein, Zorik Gekhman, and Gal Yona. Confidence improves self-consistency in LLMs. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 20090–20111, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.1030. URL <https://aclanthology.org/2025.findings-acl.1030/>.
- Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D Manning. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. *arXiv preprint arXiv:2305.14975*, 2023.
- Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.
- Zhen Wang, Xi Zhou, Yating Yang, Bo Ma, Lei Wang, Rui Dong, and Azmat Anwar. Openforecast: A large-scale open-ended event forecasting dataset. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 5273–5294, 2025.
- Gwenyth Portillo Wightman, Alexandra Delucia, and Mark Dredze. Strength in numbers: Estimating confidence of large language models by prompt agreement. In *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)*, pp. 326–362, 2023.
- Jack Wildman, Nikos I Bosse, Daniel Hnyk, Peter Mühlbacher, Finn Hambly, Jon Evans, Dan Schwarz, Lawrence Phillips, et al. Bench to the future: A pastcasting benchmark for forecasting agents. *arXiv preprint arXiv:2506.21558*, 2025.
- Adam Wilmot. Legal breakdown of kalshi and sports prediction markets, March 2025. URL <https://www.legalsportsreport.com/229042/legal-breakdown-of-kalshi-and-sports-prediction-markets/>.
- Justin Wolfers and Eric Zitzewitz. Interpreting prediction market prices as probabilities, 2006.
- xAI. Grok 3. <https://x.ai/news/grok-3>, Feb 2025a. Accessed: 2025-09-22.
- xAI. Grok 4. <https://x.ai/news/grok-4>, Jul 2025b.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

- Chenchen Ye, Ziniu Hu, Yihe Deng, Zijie Huang, Mingyu Derek Ma, Yanqiao Zhu, and Wei Wang. Mirai: Evaluating llm agents for event forecasting. *CoRR*, abs/2407.01231, 2024. URL <https://doi.org/10.48550/arXiv.2407.01231>.
- Zhiyuan Zeng, Jiashuo Liu, Siyuan Chen, Tianci He, Yali Liao, Jinpeng Wang, Zaiyuan Wang, Yang Yang, Lingyue Yin, Mingren Yin, et al. Futurex: An advanced live benchmark for llm agents in future prediction. *arXiv preprint arXiv:2508.11987*, 2025.
- Zhihan Zhang, Yixin Cao, Chenchen Ye, Yunshan Ma, Lizi Liao, and Tat Seng Chua. Analyzing temporal complex events with large language models? a benchmark towards temporal, long context understanding. In *Long Papers*, pp. 1588–1606. Association for Computational Linguistics (ACL), 2024.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623, 2023.
- Zhanke Zhou, Rong Tao, Jianing Zhu, Yiwen Luo, Zengmao Wang, and Bo Han. Can language models perform robust reasoning in chain-of-thought prompting with noisy rationales? *Advances in Neural Information Processing Systems*, 37:123846–123910, 2024.
- Chenghao Zhu, Nuo Chen, Yufei Gao, Yunyi Zhang, Prayag Tiwari, and Benyou Wang. Is your llm outdated? a deep look at temporal generalization. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 7433–7457, 2025.
- Andy Zou, Tristan Xiao, Ryan Jia, Joe Kwon, Mantas Mazeika, Richard Li, Dawn Song, Jacob Steinhardt, Owain Evans, and Dan Hendrycks. Forecasting future world events with neural networks. *Advances in Neural Information Processing Systems*, 35:27293–27305, 2022.

♣ Appendix: Table of Contents

A. Prophet Arena Pipeline Details	Page 18
B. Technical Details	Page 21
C. Additional Experiment Results	Page 25
D. Case Studies	Page 33
E. Prompts	Page 40

A Definitions and Prophet Arena Pipeline Details

A.1 Event Extraction

Prophet Arena sources unresolved events from Kalshi, a live prediction platform with events spanning across finance, sports, politics, sports, and entertainment. To ensure informativeness and comparability, we filter by *Popularity* (volume/liquidity/volatility), *Diversity* (domain balance), and *Recurrence* (repeated formats). Prophet Arena periodically retrieves 20 unresolved events each day at 12 AM (UTC).

Event. Defined formally, let $\{E_i\}_{i \in [K]}$ denote the set of evaluated *events*. An *event* is the overarching question or subject concerning a future real-world occurrence. It serves as a high-level container for one or more tradable *markets*. In many prediction markets, an event itself is **not** a tradable asset; rather, it sets the context, scope, and resolution criteria for the markets that fall under it.

- **Example 1:** “Who will win the 2025-26 NBA Championship?”
- **Example 2:** “Which individuals will President Trump officially meet in 2025?”

Market. Each event E_i contains *markets* $\{M_{ij}\}_{j \in [N_i]}$. A *market* is a specific, tradable proposition under an event that resolves to a definitive Yes or No outcome. Each market represents a potential, verifiable answer to the event’s overarching question. For a given market, a *Yes contract* is a 0-1 random variable that achieves value 1 if the Yes outcome is realized, and 0 otherwise. A “NO” contract is defined similarly, and always pays out in the opposite direction as the “YES” contract.

- **Under Event 1, a market could be:**
“The Boston Celtics will win the 2025 NBA Championship.”
- **Under Event 2, a market could be:**
“President Trump will officially meet with Emmanuel Macron in 2025.”

Event Resolution. An event E_i resolves at time τ_i with realized outcome indicator $o_{ij} \in \{0, 1\}$, where $o_{ij} = 1$ means the corresponding market M_{ij} of the event is realized (Yes). When the event index i is clear from context, we often drop the subscript i and write M_j, o_j .

A.2 Prediction Context

Prediction scheduling. For each event E_i , the benchmark specifies a finite set of pre-resolution forecast times (horizons) $\mathcal{T}_i \subset (-\infty, \tau_i)$. We construct \mathcal{T}_i by placing each subsequent forecast halfway between the current forecast time and the event close time τ_i . Let the first forecast time be $t_i^{(0)} < \tau_i$ and define the initial gap $\Delta_i^{(0)} := \tau_i - t_i^{(0)} > 0$. For $k \geq 0$,

$$t_i^{(k+1)} = \frac{t_i^{(k)} + \tau_i}{2} \iff t_i^{(k)} = \tau_i - 2^{-k} \Delta_i^{(0)}, \quad \Delta_i^{(k+1)} = \frac{\Delta_i^{(k)}}{2}.$$

To avoid excessive clustering near τ_i , we enforce a minimum time gap $\delta_{\min} > 0$ between the last forecast and the close time.

Context construction. For each event E_i and forecasting time $t \in \mathcal{T}_i$ we construct a curated context $C_{i,t}$ shared across models; this isolates forecasting ability in $p_{ij,t}$ from retrieval variability. The context $C_{i,t}$ consists of two components: relevant news sources and market data. The relevant news sources are retrieved by LLM searchers. The prompt for search is shown below. The market snapshot is fetched from Kalshi API, and contains three fields for each market: `last_price` (price of the last transaction), `yes_ask` (asking price for buying Yes), `no_ask` (asking price for buying No). From that snapshot we extract the *implied probability* $q_{ij,t} \in [0, 1]$ for a Yes contract at time t (with No priced at $1 - q_{ij,t}$).¹²

At each forecasting time $t_i \in \mathcal{T}_i$, LLM searchers will be dispatched and live market snapshot will be retrieved. Together, relevant news sources from LLM searchers and market snapshots from Kalshi will serve as the prediction context.

¹²Implied probabilities are reverse-engineered from contract prices; transaction fees may cause YES/NO prices not to sum to 1.

Importantly, Prophet Arena is *searcher-agnostic*: the search component is an independent, pluggable module and adding new searchers does not change the forecasting protocol or the scoring of $p_{ij,t}$. Prophet Arena actively updates to include new LLM searchers. In this paper, we instantiate a single LLM searcher using GPT-4o with web search enabled and all experiments use that configuration (§E.1.1).

A.3 Probabilistic Forecasting to Account for Uncertainty

LLM-predicted Probabilities. Given $(E_i, M_{ij}, C_{i,t}, t)$, a model must output an *LLM probability*,

$$p_{ij,t} \in [0, 1],$$

interpreted as its belief that M_{ij} will realize `Yes` at time t , accompanied by a natural-language rationale (logged for analysis but not used in scoring). The prediction prompts are documented in §E.1.2.

Implied Probabilities. At each $t \in \mathcal{T}_i$, a `Yes` (resp. `No`) contract is valued at $q_{ij,t}$ (resp. $1 - q_{ij,t}$). The *implied probability* q_{ij} represents the (human) market-consensus belief that the `Yes` outcome will come true.

A.4 Resolution and Evaluation

Edge and Utility. We denote a (yes) *edge* $e_{ij} := \frac{p_{ij}}{q_{ij}}$ as the likelihood-ratio between the LLM-predicted and implied probabilities. A larger edge signals the LLM to be more confident (than the market) that a market will be realized. Similarly, we define the (no) edge to be $\tilde{e}_{ij} := \frac{1-p_{ij}}{1-q_{ij}}$.

We assume that the *price* of a single `Yes` contract simply equals the implied probability q_{ij} , and the price of a single `No` contract is thus $1 - q_{ij}$.¹³ In the sequel, our strategy is limited to buying contracts (taking a long position). But all contracts can be purchased in fractional amounts.

Our simulated trading policies (used only for *relative* metrics) considers the utility function with risk-aversion hyperparameter $\gamma \in [0, 1]$. It maps any wealth (i.e. payoff) to a utility.

Prediction Evaluation. After the close time τ_i , we will retrieve the outcomes of event E_i from Kalshi. For each market M_{ij} and each horizon $t \in \mathcal{T}_i$, we score $(p_{ij,t}, o_{ij})$ with *absolute* proper scoring rules (e.g., Brier-based) and, separately, evaluate *relative* Average Return by simulating trades against $q_{ij,t}$ under U_γ . Scores are then aggregated across j, i , with the specific calculations detailed in the subsequent evaluation section.

The below table summarizes the notations that will appear in the later math expressions:

E_i ,	$i \in [K]$, the i -th event in our evaluation set.
M_{ij} (M_j)	$j \in [N_i]$, the j -th market of the i -th event, we drop the i subscript when the market is obvious (same for below).
p_{ij} (p_j)	the LLM-predicted probability that M_{ij} will realize.
q_{ij} (q_j)	the market implied probability that M_{ij} will realize.
e_{ij}/\tilde{e}_{ij} (e_j/\tilde{e}_j)	the (yes/no) edge of M_{ij} .
o_{ij} (o_j)	the indicator of whether M_{ij} is realized.
$U_\gamma(w)$	the utility function with risk-aversion hyperparameter $\gamma \in [0, 1]$. It maps any wealth (i.e. payoff) to a utility.

A.5 Comprehensive List of Evaluated LLMs

Remark. The column **Default Reasoning** specifies the reasoning configuration used when a model is referenced without qualifiers (e.g. GPT-5, or when we write GPT-5^R in Table 2). In the full evaluation table (Table 5), some models appear multiple times under different reasoning settings; each such variant is treated as a distinct model (e.g., GPT-5 (Minimal)).

Reasoning configuration is inherently model-dependent. For “hybrid reasoning” models that allow toggling between thinking and non-thinking modes, the configuration is specified as either *enabled* or *disabled*. For

¹³In practice, (1) we actually “reverse-engineer” the implied probabilities from market contract prices, and (2) the prices of `Yes/No` contracts might not sum to 1 due to transaction fees taken by the exchange.

LLM	Citation	Open Weight?	Default Reasoning
GPT-5	OpenAI (2025b)	No	High
o3	OpenAI (2025c)	No	High
o3-Mini	OpenAI (2025c)	No	Medium
o4-Mini	OpenAI (2025c)	No	High
Gemini 2.5 Pro	Comanici et al. (2025)	No	Enabled
Gemini 2.5 Flash	Comanici et al. (2025)	No	Enabled
Grok-4	xAI (2025b)	No	Enabled
Grok-3-Mini	xAI (2025a)	No	Enabled
GPT-4.1	OpenAI (2025a)	No	N/A
Claude Sonnet 4	Anthropic (2025)	No	Enabled
Kimi-K2	Kimi Team (2025)	Yes	N/A
GPT-4o	OpenAI (2024)	No	N/A
Llama 4 Maverick	Meta Team (2025)	Yes	N/A
Llama 4 Scout	Meta Team (2025)	Yes	N/A
DeepSeek-V3	Liu et al. (2024)	Yes	N/A
DeepSeek-R1	Guo et al. (2025)	Yes	Enabled
Qwen3-235B	Yang et al. (2025)	Yes	N/A
Gemini 2.0 Flash	Gemini Team (2024)	No	N/A
Gemini 2.0 Flash (Lite)	Gemini Team (2024)	No	N/A

Table 4: **Comprehensive list of LLMs evaluated in Prophet Arena** (as of submission time).

models that expose explicit control over reasoning effort, the configuration is expressed in levels (e.g., *minimal*, *medium*, *high*).

B Technical Details

B.1 Further Intuitions behind Brier Score

In Prophet Arena, the set of markets $\{M_{ij}\}_{j=1}^{m_i}$ under an event E_i need not be mutually exclusive. This has two implications: (1) the realized outcome vector $\mathbf{o}_i = (o_{i1}, \dots, o_{im_i})'$ may contain multiple ones rather than being strictly one-hot, and (2) the predicted probabilities $\{p_{ij}\}_j$ may sum to more than one. Our Brier score formulation remains robust to these cases because the score is always evaluated at the *market level*, where each individual market is binary – resolving either to Yes or No, but not both. If we pool all markets across all events and relabel them M_1, \dots, M_k , the standard binary Brier score is simply

$$BS = \frac{1}{k} \sum_{j=1}^k (p_j - o_j)^2. \quad (3)$$

Our event-level definition in § 3.1.1 can be viewed as a *weighted* version of Eq. (3). Since events vary greatly in the number of associated markets, directly pooling them would let large events dominate the metric. To mitigate this, we assign each market a weight $w_{ij} = 1/m_i$, inversely proportional to the number of markets in its event. This yields the final form:

$$BS = \frac{1}{\sum_i \sum_j w_{ij}} \sum_{i=1}^N \sum_{j=1}^{m_i} w_{ij} (p_{ij} - o_{ij})^2 = \frac{1}{N} \sum_{i=1}^N \left(\frac{1}{m_i} \sum_{j=1}^{m_i} (p_{ij} - o_{ij})^2 \right).$$

This weighting ensures that each event contributes equally, regardless of its size, while still respecting the binary resolution of individual markets.

B.2 A Better-calibrated forecast can have a worse Brier score.

As mentioned in § 3.1.2, calibration errors and Brier scores measure different aspects of forecasts. This difference can be illustrated with a simple example. Suppose there are two markets M_1, M_2 with ground-truth probability $p_1^* = 0.9, p_2^* = 0.1$. Alice’s prediction is $p_1^A = 1, p_2^A = 0$ whereas Bob’s prediction is $p^B = 0.5$ for both markets. It is not difficult to see that Alice has a better/smaller Brier score, but she turns out to have a worse/larger ECE. Specifically, since Alice has two different probability predictions, her ECE is $\frac{1}{2}(|p_1^* - p_1^A| + |p_2^* - p_2^A|) = 0.1$ whereas Bob has a single probability prediction with the ECE equaling $\frac{1}{2}|p_1^* + p_2^* - 2p^B| = 0$. That is, Bob’s prediction of 0.5 probability is perfectly calibrated and trustworthy as, conditioned on this probability, events’ average probability is indeed 0.5.

B.3 Empirical Estimation of Calibration Error

In § 3.1.2, we introduced the formal definition of expected calibration error (ECE) in terms of true conditional probabilities. In practice, this definition cannot be computed exactly, as the conditional terms $\mathbb{P}(o_j = 1 \mid p_j = \tilde{p}_i)$ are not directly observable. Instead, the standard approach in the applied literature is to approximate ECE via a binned empirical estimate.

Formally, let $\{(p_j, o_j)\}_{j=1}^m$ denote a set of predicted probabilities and realized outcomes. We first partition the unit interval $[0, 1]$ into B disjoint bins $\{I_b\}_{b=1}^B$, and assign each prediction p_j to its corresponding bin. Let $M_b = \{j : p_j \in I_b\}$ be the index set of predictions falling into bin b , and $m_b = |M_b|$ its size. Define

$$\hat{p}_b = \frac{1}{m_b} \sum_{j \in M_b} p_j, \quad \hat{o}_b = \frac{1}{m_b} \sum_{j \in M_b} o_j,$$

as the average predicted probability and empirical frequency of Yes outcomes in bin b , respectively. The **empirical expected calibration error** is then given by

$$\widehat{ECE} = \frac{1}{m} \sum_{b=1}^B m_b \cdot |\hat{o}_b - \hat{p}_b|. \quad (4)$$

Intuitively, \widehat{ECE} measures the weighted average discrepancy between empirical accuracy and average predicted probability across bins, with weights proportional to bin counts. Throughout our experiments, all reported calibration results correspond to this empirical version (with $B = 10$).

B.4 Differences between Brier Scores and Market Returns

We provide a simple example to prove the following fact.

Fact 1. *There exist a binary prediction market with two forecaster A, B such that A has strictly higher market return than B but has strictly worse/higher Brier score.*

Proof. Suppose we bet on a single event with binary outcomes, with ground-truth probability of Y_{es} being 0.6 (for the event to be realized), and prediction market price 0.5. Consider two different probabilistic predictions of this events Y_{es} realization: model A predicts 0.45 and model B predicts 0.9.

The expected Brier Score of A is:

$$1 - [0.6 \cdot (0.45 - 1)^2 + 0.4 \cdot (0.45 - 0)^2] = 1 - 0.2625 = 0.7375$$

The expected Brier Score of B is:

$$1 - [0.6 \cdot (0.9 - 1)^2 + 0.4 \cdot (0.9 - 0)^2] = 1 - 0.33 = 0.67$$

So A has a higher Brier Score. However, because A predicts 0.45, lower than the 0.5 prediction market price, A will short Y_{es} (or equivalently, buy N_{o}) at 0.5. Meanwhile, B predicts 0.9, much higher than the prediction market price, so B will buy Y_{es} . Respectively, A and B s expected returns will be:

$$\begin{aligned} 0.6 \cdot (-1 + 0.5) + 0.4 \cdot (0.5) &= -0.1 \\ 0.6 \cdot (1 - 0.5) + 0.4 \cdot (-0.5) &= 0.1 \end{aligned}$$

Therefore, A has a higher Brier score, but lower returns.

□

This example uncovers a key difference between the two metrics. The Brier Score measures how close a prediction is to the ground truth and, importantly, has nothing to do with market prices. Since A s prediction above is closer to the ground truth, it receives a higher Brier Score. However, returns on the market are not only driven by the true probability but also by the market price. Therefore, even though B s prediction is exaggerated, it lies on the correct side of the market mispricing (buying “Yes” when the outcome is more likely than price suggests), thereby achieving higher returns.

B.5 A Unified Framework for Utility-Maximizing Betting Strategy

In §3.1.3, we introduced **Average Return** under a simple strategy: bet the full budget on Y_{es} whenever $p > q$, and on N_{o} otherwise, where p is the models predicted probability and q is the market price of a Y_{es} contract. Here we generalize this idea by developing a unified framework for betting strategies in binary markets. This framework serves two purposes:

1. It formalizes betting as a utility-maximization problem, allowing us to flexibly encode different risk preferences.
2. It shows that the simple strategy used in the main text is a special case of this unified framework, corresponding to the risk-neutral setting.

Contracts and market prices. Each binary market resolves to either Y_{es} or N_{o} . A share of a Y_{es} contract pays \$1 if the outcome is Y_{es} and \$0 otherwise; the price of this contract is denoted q . Symmetrically, a share of a N_{o} contract costs $1 - q$ and pays \$1 if the outcome is N_{o} . These prices are often interpreted as market-implied probabilities (Wolfers & Zitzewitz, 2006), though our framework does not rely on this interpretation.

Utility function. Let $U : \mathbb{R}_{\geq 0} \rightarrow \mathbb{R}$ denote a utility function mapping a payoff w to its perceived value (satisfaction) $U(w)$. We focus on the class of constant relative risk aversion (CRRA) utilities:

$$U_{\gamma}(w) = \begin{cases} \frac{w^{1-\gamma}}{1-\gamma}, & 0 \leq \gamma < 1, \\ \log w, & \gamma = 1 \end{cases} \quad (5)$$

where $\gamma \in [0, 1]$ indexes risk aversion. At $\gamma = 0$, U_{γ} is linear (risk-neutral); at $\gamma = 1$, it reduces to $\log w$ (log utility, risk-averse); and for $\gamma \in (0, 1)$, it interpolates smoothly between the two. Fig. 7 visualizes representative cases.

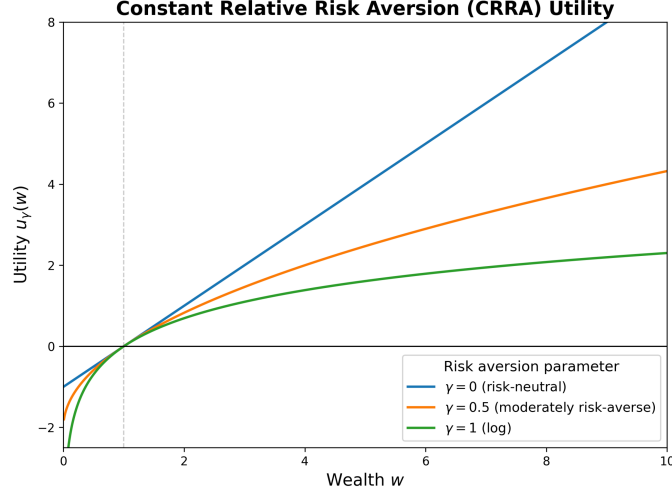


Figure 7: Visualization of CRRA utility functions.

Budget allocation as optimization. Fix a unit budget of \$1 for each market. Let a_Y and a_N denote the amounts allocated to Yes and No contracts, with $a_Y + a_N = 1$. Purchasing a_Y dollars of Yes contracts yields a_Y/q shares (i.e., the pay off is a_Y/q if the outcome is Yes); similarly, $a_N/(1-q)$ shares of No pay off if the outcome is No. Given belief (predicted probability) p and risk preference γ , the optimal allocation is the solution to

$$\max_{a_Y, a_N \geq 0 : a_Y + a_N = 1} \left[p \cdot U_\gamma \left(\frac{a_Y}{q} \right) + (1-p) \cdot U_\gamma \left(\frac{a_N}{1-q} \right) \right]. \quad (6)$$

This objective is the expected utility of betting under p .

Closed-form solutions. For $\gamma > 0$, the optimization admits a closed-form solution:

$$a_Y^* \equiv a_Y^*(\gamma) = \frac{q^{1-\frac{1}{\gamma}} p^{\frac{1}{\gamma}}}{q^{1-\frac{1}{\gamma}} p^{\frac{1}{\gamma}} + (1-q)^{1-\frac{1}{\gamma}} (1-p)^{\frac{1}{\gamma}}}, \quad a_N^* \equiv a_N^*(\gamma) = 1 - a_Y^*, \quad (7)$$

At $\gamma = 0$, we take the limit $\gamma \rightarrow 0^+$, yielding

$$a_Y^*(0) \equiv \lim_{\gamma \rightarrow 0} a_Y^*(\gamma) = \begin{cases} 1, & \text{if } p > q, \\ 0, & \text{if } p \leq q. \end{cases}$$

which recovers the strategy used in the main text (§ 3.1.3). At the other extreme, when $\gamma = 1$, the optimal strategy is to allocate $a_Y^* = p$ and $a_N^* = 1 - p$, i.e., to bet proportionally to one's own probabilities, independent of the market price. Overall, this unified framework highlights how different betting strategies arise from different risk preferences. The “all-in” rule in the main text is simply the risk-neutral optimum.

B.6 Market Returns for a Calibrated Predictor

In § 3, we introduced the evaluation pipeline in Prophet Arena, including how market return is calculated based on market decisions (buying either the YES or NO contract), which are derived from predicted and market-implied probabilities (p_j, q_j) as well as the realized outcome o_j under the following decision rule: *buy YES contract whenever $p_j > q_j$, and buy NO otherwise*.

Here we further make the connection between a (perfectly) calibrated predictor and **market returns conditioning on the type of decision made**. Before stating the main result, the following mild assumptions and definitions are needed.

Assumption B.1 (Data Generation). We assume that the triples (o_j, p_j, q_j) are drawn i.i.d. from some joint probability distribution \mathcal{D} . The expectation $\mathbb{E}[\cdot]$ below is taken over \mathcal{D} .

Assumption B.2 (Sufficiency of Calibrated Predictors). Perfect calibration implies that $\forall c \in [0, 1]$,

$$\mathbb{E}[o_j \mid p_j = c] = 1 \cdot \mathbb{P}[o_j = 1 \mid p_j = c] + 0 \cdot \mathbb{P}[o_j = 0 \mid p_j = c] = c.$$

We further assume that a calibrated predictor is **sufficient** for the market outcome: the outcome o_j is conditionally independent of the market price q_j given the forecast p_j , i.e. $\mathbb{E}[o_j \mid p_j, q_j] \equiv \mathbb{E}[o_j \mid p_j]$

Definition B.3 (Symmetric Disagreement). We say a predictor is **symmetric** (against the market) if the probability distribution of $d_j := p_j - q_j$ is **symmetric about zero**. We let $f(x)$ denote the p.d.f. of d_j and the definition implies (1) $f(x) = f(-x)$ for all $x \in \mathbb{R}$, and (2) $\mathbb{P}[p_j > q_j] = \mathbb{P}[p_j \leq q_j]$. In other words, d_j represents the “disagreement” that the predictor perceives over the market, and a **symmetric** predictor is not consistently more conservative or more aggressive.

Theorem B.4 (Symmetric Expected Returns for Calibrated Predictors). *Let $R_Y = o_j - q_j$ be the return on a YES contract and $R_N = (1 - o_j) - (1 - q_j) = q_j - o_j$ be the return on a NO contract. For a predictor that is both **perfectly calibrated** and **symmetric** against the market, the expected returns conditioned on the betting decision are equal, i.e.*

$$\mathbb{E}[R_Y \mid p_j > q_j] = \mathbb{E}[R_N \mid p_j < q_j]. \quad (8)$$

Proof. Let $E_Y = \mathbb{E}[o_j - q_j \mid p_j > q_j]$ and $E_N = \mathbb{E}[q_j - o_j \mid p_j < q_j]$ denote the average returns on YES and NO contracts, respectively. We begin by analyzing the expected outcome o_j under each condition. By the Law of Iterated Expectations and the sufficiency of the calibrated predictor (Assumption B.2), we can replace the outcome o_j with the prediction p_j inside the expectation:

$$\begin{aligned} \mathbb{E}[o_j \mid p_j > q_j] &= \mathbb{E}[\mathbb{E}[o_j \mid p_j, q_j] \mid p_j > q_j] = \mathbb{E}[p_j \mid p_j > q_j], \\ \mathbb{E}[o_j \mid p_j < q_j] &= \mathbb{E}[\mathbb{E}[o_j \mid p_j, q_j] \mid p_j < q_j] = \mathbb{E}[p_j \mid p_j < q_j]. \end{aligned}$$

Substituting these back into the expressions for E_Y and E_N :

$$\begin{aligned} E_Y &= \mathbb{E}[p_j \mid p_j > q_j] - \mathbb{E}[q_j \mid p_j > q_j] = \mathbb{E}[p_j - q_j \mid p_j > q_j], \\ E_N &= \mathbb{E}[q_j \mid p_j < q_j] - \mathbb{E}[p_j \mid p_j < q_j] = \mathbb{E}[q_j - p_j \mid p_j < q_j]. \end{aligned}$$

The theorem is proven if we can show that $\mathbb{E}[p_j - q_j \mid p_j > q_j] = \mathbb{E}[q_j - p_j \mid p_j < q_j]$. Let $d_j = p_j - q_j$ be the disagreement variable. The equality becomes:

$$\mathbb{E}[d_j \mid d_j > 0] = \mathbb{E}[-d_j \mid d_j < 0]$$

We show this using the integral definition of conditional expectation. The symmetric disagreement assumption implies $f(x) = f(-x)$ and $\mathbb{P}(d_j > 0) = \mathbb{P}(d_j < 0)$. Consider the right-hand side:

$$\mathbb{E}[-d_j \mid d_j < 0] = \frac{\int_{-\infty}^0 -x f(x) dx}{\mathbb{P}(d_j < 0)}.$$

Applying the change of variable $u = -x$, we have $x = -u$ and $dx = -du$. Therefore

$$\frac{\int_{-\infty}^0 -(-u)f(-u)(-du)}{\mathbb{P}(d_j < 0)} = \frac{-\int_{\infty}^0 u f(-u) du}{\mathbb{P}(d_j < 0)} = \frac{\int_0^{\infty} u f(-u) du}{\mathbb{P}(d_j < 0)}.$$

Using the symmetry properties $f(-u) = f(u)$ and $\mathbb{P}(d_j < 0) = \mathbb{P}(d_j > 0)$, this is equal to:

$$\frac{\int_0^{\infty} u f(u) du}{\mathbb{P}(d_j > 0)} = \mathbb{E}[d_j \mid d_j > 0].$$

Thus, we have shown $\mathbb{E}[-d_j \mid d_j < 0] = \mathbb{E}[d_j \mid d_j > 0]$, which concludes the proof. \square

B.7 Bootstrap Confidence Intervals

Most metrics in this paper can be summarized by the following two steps:

1. We first calculate a collection of scores s_1, \dots, s_N at either the event level or the market level, where N denotes the total number of events or markets.
2. We then aggregate these scores into a single statistic via a mapping $T : \mathbb{R}^N \rightarrow \mathbb{R}$ (typically the arithmetic mean).

To assess the uncertainty of the resulting point estimate, we construct a nonparametric bootstrap confidence interval (CI) at level $(1 - \alpha)\%$ (DiCiccio & Efron, 1996). Specifically, we form a symmetric interval around the point estimate using bootstrap resampling. A pseudo-code for our implementation is given below in Algorithm 1.

Algorithm 1 Symmetric Nonparametric Bootstrap CI Centered at Point Estimate $\hat{\theta}$

Require: Data s_1, \dots, s_N ; statistic $T(\cdot)$; number of bootstrap replicates B ; target level $1 - \alpha$;**Ensure:** Symmetric CI $[\hat{\theta} - h, \hat{\theta} + h]$ and achieved bootstrap mass k/B .

- 1: Compute point estimate $\hat{\theta} \leftarrow T(s_1, \dots, s_N)$.
 - 2: **for** $b = 1$ **to** B **do**
 - 3: Draw a bootstrap sample $\{s_1^{*(b)}, \dots, s_N^{*(b)}\}$ by sampling with replacement from $\{s_i\}_{i=1}^N$.
 - 4: Compute bootstrap re-estimate $\theta^{*(b)} \leftarrow T(s_1^{*(b)}, \dots, s_N^{*(b)})$.
 - 5: **end for**
 - 6: Compute deviations $d_b \leftarrow |\theta^{*(b)} - \hat{\theta}|$ for $b = 1, \dots, B$.
 - 7: Sort $\{d_b\}$ into order statistics $d_{(1)} \leq \dots \leq d_{(B)}$.
 - 8: $h \leftarrow d_{(k)}$.
 - 9: **return** Symmetric CI $[\hat{\theta} - h, \hat{\theta} + h]$.
-

C Additional Experiment Results

C.1 Full Evaluation Results

LLMs	Forecasting Loss		Calibration Error		Market Return	
	\downarrow Brier (95% CI)	Rank	\downarrow ECE	Rank	\uparrow Average (95% CI)	Rank
GPT-5 ^R (High) \triangle	0.184 (\pm 0.006)	①	0.042	④	0.943 (\pm 0.042)	⑥
GPT-5 ^R	0.187 (\pm 0.005)	②	0.044	⑥	0.890 (\pm 0.040)	13
Market Baseline	0.187 (\pm 0.006)	③	0.069	19	0.899 (\pm 0.043)	10
GPT-5 ^R (Minimal)	0.188 (\pm 0.006)	④	0.036	②	0.869 (\pm 0.044)	19
o3 ^R	0.188 (\pm 0.005)	⑤	0.030	①	0.959 (\pm 0.109)	⑤
Grok-4 ^R \triangle	0.189 (\pm 0.005)	⑥	0.043	⑤	0.864 (\pm 0.052)	21
Grok-3-Mini ^R	0.189 (\pm 0.006)	7	0.046	7	0.869 (\pm 0.043)	20
GPT-4.1	0.192 (\pm 0.007)	8	0.053	10	0.907 (\pm 0.035)	8
Gemini 2.5 Pro ^R	0.193 (\pm 0.007)	9	0.061	14	0.876 (\pm 0.050)	17
Claude Opus 4.1 ^R	0.193 (\pm 0.018)	10	0.054	11	0.982 (\pm 0.093)	①
Claude Sonnet 4 ^R \triangle	0.194 (\pm 0.006)	11	0.041	③	0.909 (\pm 0.101)	7
o3-Mini ^R	0.195 (\pm 0.005)	12	0.046	8	0.897 (\pm 0.046)	12
o4-Mini ^R (High)	0.196 (\pm 0.006)	13	0.062	16	0.874 (\pm 0.040)	18
Kimi-K2	0.197 (\pm 0.008)	14	0.048	9	0.966 (\pm 0.124)	③
Gemini 2.5 Flash ^R \triangle	0.197 (\pm 0.007)	15	0.067	17	0.883 (\pm 0.053)	15
GPT-4o	0.198 (\pm 0.007)	16	0.058	12	0.970 (\pm 0.104)	②
DeepSeek-V3	0.201 (\pm 0.008)	17	0.061	15	0.963 (\pm 0.103)	④
Gemini 2.5 Flash	0.203 (\pm 0.006)	18	0.073	20	0.859 (\pm 0.042)	22
Llama-4-Maverick	0.208 (\pm 0.008)	19	0.067	18	0.904 (\pm 0.050)	9
Llama-4-Scout \triangle	0.219 (\pm 0.008)	20	0.060	13	0.805 (\pm 0.040)	24
Gemini 2.0 Flash (Lite)	0.224 (\pm 0.013)	21	0.091	22	0.855 (\pm 0.074)	23
Gemini 2.0 Flash	0.224 (\pm 0.013)	22	0.079	21	0.876 (\pm 0.078)	16
Qwen3-235B	0.234 (\pm 0.007)	23	0.118	23	0.898 (\pm 0.111)	11
DeepSeek-R1 ^R	0.303 (\pm 0.018)	24	0.165	24	0.884 (\pm 0.058)	14

Table 5: The full evaluation results for all 23 LLMs (including variants) and the market baseline. \triangle denotes models selected for presentation in the main text.

C.2 Sharpe Ratio for Average Return

As discussed in §3.2 and Table 5, Average Return often exhibits wide confidence intervals, reflecting the high variance of event-level payoffs. In most events, models earn nothing or only marginal gains, while in rare cases they achieve outsized returns. To account for this variability, we complement Average Return with the *Sharpe ratio* (Sharpe, 1998), a standard metric that normalizes expected returns by their volatility. Formally, for asset returns R_a , the Sharpe ratio is defined as

$$S_a = \frac{\mathbb{E}[R_a - R_b]}{\sqrt{\text{Var}[R_a - R_b]}}, \quad (9)$$

where R_b denotes a reference risk-free return. In the Prophet Arena setting, R_a corresponds to the payoff from each event under the trading strategy of §3.1.3, and we set $R_b = 1$, the return from abstaining (i.e., keeping the budget unbet). The expectation and variance in Eq. (9) are estimated by the sample mean and variance over n events.

Sharpe ratios for all evaluated models are reported in Table 6, providing a volatility-adjusted comparison of economic performance.

LLMs	↑ Sharpe Ratio	Rank	LLMs	↑ Sharpe Ratio	Rank
o3 ^R	-0.0131	①	GPT-4.1	-0.0707	13
GPT-5 ^R	-0.0212	②	Claude Opus 4.1 ^R	-0.0763	14
Gemini 2.5 Pro ^R	-0.0230	③	Grok-4 ^R △	-0.0771	15
Gemini 2.5 Flash ^R △	-0.0270	④	Grok-3-Mini ^R	-0.0821	16
o3-Mini ^R	-0.0285	⑤	o4-Mini ^R (High)	-0.0853	17
Kimi-K2	-0.0373	⑥	Market Baseline	-0.0897	18
DeepSeek-V3	-0.0381	7	DeepSeek-R1 ^R	-0.1199	19
GPT-4o	-0.0383	8	GPT-5 ^R (Minimal)	-0.1400	20
Qwen3-235B	-0.0389	9	Gemini 2.5 Flash	-0.1483	21
Claude Sonnet 4 ^R △	-0.0404	10	Gemini 2.0 Flash	-0.1604	22
GPT-5 ^R (High) △	-0.0435	11	Llama-4-Scout △	-0.1799	23
Llama-4-Maverick	-0.0637	12	Gemini 2.0 Flash (Lite)	-0.1842	24

Table 6: **Sharpe ratio performance for all 23 LLMs and market baseline.** △ denotes models selected for presentation in the main text.

C.3 Probability Elicitation Methods

§2 introduced our default approach: directly prompting an LLM to verbalize the probability that a market resolves to Yes. In this experiment, we conduct ablation studies over alternative confidence estimation methods. The goal is twofold: (i) test the robustness of our default prompt against reasonable variations, and (ii) illustrate how Prophet Arena can serve as a testbed for comparing black-box confidence estimation methods under forecasting settings.

Setup. We evaluate the five representative LLMs from main text on the PROPHET-ARENA-SUBSET-100 dataset. Metrics are Brier and ECE scores. We consider two families of confidence estimation methods (seven variants total):

1. Verbalized Probability (Tian et al., 2023)

- *Prompt variation*: we modify our default prompt to make it: (A) more concise, (B) more verbose, and (C) rewritten by another LLM (Grok 4). Key logistics and formatting instructions are preserved in all variations. These variation prompts are available in §E.4.
- *Prompt ensemble* (Wightman et al., 2023): for each market, we average the probabilities elicited from the default and all variation prompts (i.e. (A), (B), (C) above).
- *Bi-direction** (ours): in addition to eliciting p_{ij} (probability of Yes) using default prompt, we also elicit p_{ij}^o (probability of NO), and calibrate via $\frac{1}{2}(p_{ij} + (1 - p_{ij}^o))$.

2. Self-consistency (Wang et al., 2022)

Method Type / Name	Grok 4	Gemini 2.5 Flash	Claude Sonnet 4	GPT-5	Llama 4 Scout
<i>Verbalized Prob</i>					
Default	0.186/0.117	0.166/0.036	0.173/0.046	0.165/0.020	0.196/0.153
Variation A (Concise)	0.180/0.102	0.172/0.036	0.169/0.035	0.162/0.021	0.192/0.134
Variation B (Verbose)	0.178/0.124	0.166/0.039	0.172/0.040	0.160/0.028	0.199/0.164
Variation C (Rewrite)	<u>0.176/0.123</u>	<u>0.167/0.027</u>	0.172/0.039	0.159/0.016	0.195/0.167
Ensemble	0.177/0.117	0.165/0.032	0.170/0.043	0.160/0.024	<u>0.192/0.142</u>
Bi-direction*	0.180/0.101	<u>0.164/0.031</u>	<u>0.165/0.028</u>	0.158/0.023	0.203/0.140
<i>Self-consistency</i>					
Unweighted	0.238/0.115	0.231/0.110	0.241/0.071	0.239/ 0.071	0.267/0.129
Weighted	0.205/ <u>0.091</u>	0.201/0.067	0.189/0.050	0.181/0.046	0.214/ <u>0.125</u>

Table 7: **Evaluation Results for Different Probability Elicitation Methods** . Each cell contains a pair of $\text{Brier}\downarrow / \text{ECE}\downarrow$ scores. **Bold** denotes the best score for each row, underline for each column.

- *Unweighted*: instead of directly asking for probability, we repeatedly query the model 10 times for a Yes/No decision; probability is the fraction of Yes answers.
- *Weighted* (Taubenfeld et al., 2025): we further supplement each Yes/No with a confidence score.¹⁴ Probabilities are formed by confidence-weighted aggregation.

Results (see Table 7). Among the verbalized probability methods, we observe that **all models exhibit strong robustness to prompt variations**. For accuracy, Brier scores for all models vary by less than one standard deviation (≈ 0.01). As a result, GPT-5 ranks the highest under all methods, and Gemini 2.5 Flash (Thinking) consistently achieves second place in 5/6 cases. Prompt ensemble does not lead to substantial improvement, since elicited probabilities are already similar across variations. For calibration, GPT-5 and Llama 4 Scout are consistently the best and the worst models, regardless of the prompting method. Despite slightly larger fluctuations among the ECE scores, **no single method dominate the others for all models** (i.e. achieves the best scores on all columns). Our original Bi-direction* method improves calibration (over the default) on 4/5 models, supporting the view that LLMs tend to be overconfident toward Yes outcomes.

In contrast, **self-consistency methods result in significantly lower accuracies and yield mixed calibration benefits**. With 10 rollouts, the unweighted variant produces coarse-grained probabilities at a resolution of 0.1, limiting accuracy despite incurring higher compute cost. The weighted variant partially alleviates this by using finer-grained confidence signals, producing noticeable calibration gains but still trailing verbalized methods in accuracy.

To sum up, these results show that (i) LLMs are generally robust against prompt ablations, justifying our default prompt choice, and (ii) Prophet Arena provides a natural benchmark for evaluating and contrasting future confidence estimation/calibration methods.

C.4 Reasoning Consistency of LLMs

In addition to the primary relative (Brier score) and absolute (Average return) metrics, certain types of forecasting events also enable us to evaluate the consistency – an important component of reasoning – of LLMs. These metrics can be calculated **solely by looking at the potential event outcomes and the LLM probabilities given to them**. Below we give two concrete examples of such consistency metrics.

Logical chain score.¹⁵ Consider the forecasting event “The bitcoin price by the end of 2026”, and two of its outcomes are “(A) The bitcoin price is above \$200,000” and “(B) The bitcoin price is above \$220,000”, respectively. No matter how good an LLM is at predicting the probabilities for (A) & (B), we know that anyone with **consistent reasoning** will give $\mathbb{P}[(A)] \geq \mathbb{P}[(B)]$ since the latter outcome logically implies the former. We denote such a relationship as a **logical chain**, or $(B) \rightarrow (A)$. Obviously, this logical chain can contain more than two outcomes, so we call a logical chain $\mathcal{S} = (S_1) \rightarrow \dots \rightarrow (S_T)$ **maximal** whenever it satisfies both:

¹⁴This confidence reflects uncertainty about the chosen answer, not a direct market probability (e.g., low confidence in Yes signals indecision, not belief in NO).

¹⁵This is a placeholder name. Feel free to suggest a better one.

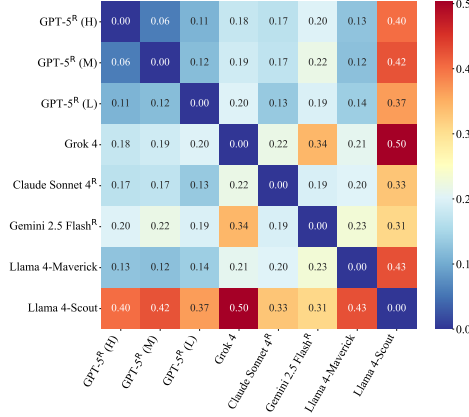


Figure 8: **Pairwise divergence of model predictions across tasks.** Heatmap shows the average L_2 distance between probabilistic predictions of each model pair. GPT-5 H, M, L labels represent high, medium, and minimal reasoning efforts, as explained by §A.5.

1. For all $1 \leq t < T$, we have $(S_t) \rightarrow (S_{t+1})$,
2. No other outcome $(K) \notin \mathcal{S}$ satisfies $(K) \rightarrow (S_1)$ or $(S_T) \rightarrow (K)$.

A single event might contain multiple maximal logical chains $\mathcal{S}_1, \dots, \mathcal{S}_n$ with lengths T_1, \dots, T_n . For an LLM with probability $\mathbb{P}[(A)]$ for outcome (A) , its **logical chain score** for this event is given by $\frac{1}{n} \sum_{i=1}^n \text{score}(\mathcal{S}_i)$, where

$$\text{score}(\mathcal{S}_i = (S_{i1} \rightarrow \dots \rightarrow S_{iT_i})) := \frac{1}{T_i - 1} \sum_{j=1}^{T_i-1} \mathbf{1}\{\mathbb{P}[(S_j)] \leq \mathbb{P}[(S_{j+1})]\} \quad (10)$$

with $\mathbf{1}(\cdot)$ being the indicator function. The final logical chain score is then **averaged over all events with at least one chain**. We adopt an LLM-judge (Gemini-2.5-Flash) to automatically detect the maximal logical chains for all our events.

Mutually exclusive score. In certain forecasting events, the potential outcomes are **mutually exclusive**, meaning that exactly one outcome can occur. For example, in the event “*Who will win the NBA championship in 2026?*”, the possible outcomes could be the teams, where only one team can win. A **maximal set of mutually exclusive outcomes** is a set where

1. each outcome is distinct and all outcomes are mutually exclusive,
2. the event will resolve to one and only one outcome in the set.

If such a maximal set $\mathcal{S} := \{(S_1), \dots, (S_m)\}$ with size m exists for a forecasting event, we can calculate the **mutually exclusive score** at this event as:

$$\text{score}_{\text{ME}} = \mathbf{1} \left\{ \sum_{i=1}^m \mathbb{P}[(S_i)] = 1 \right\} \quad (11)$$

(In practice, we allow the sum to deviate slightly from 1 with some tolerance level ϵ).

We evaluate this score over all events where mutually exclusive outcomes are defined. The identification of maximal set is performed using the same LLM-judge (Gemini-2.5-Flash).

C.5 Variability in Model Forecasts Despite Identical Inputs

While all tested LLMs receive the same market data and news sources, they differ substantially in how they combine and weigh each piece of information, resulting in diverse prediction patterns. Fig. 8 shows averaged pairwise differences in event predictions, and the generally large differences across models suggest that LLMs reason about events in fundamentally different ways, even when given identical inputs.

LLMs	Mutually Exclusive Consistency	Logical Chain Consistency
DeepSeek-R1 ^R	0.996	0.987
o4-Mini ^R	0.995	0.998
GPT-5 ^R (High)	0.995	0.999
Gemini 2.5 Flash ^R	0.995	0.997
Gemini 2.5 Pro ^R	0.995	0.995
Claude Sonnet 4 ^R	0.995	0.987
DeepSeek-V3	0.994	0.973
Grok-4 ^R	0.994	0.994
GPT-4.1	0.994	0.973
Llama 4 Maverick	0.994	0.981
Llama 4 Scout	0.994	0.979
Grok-3-Mini ^R	0.994	0.901
Qwen3-235B	0.994	0.988
GPT-4o	0.994	0.930
GPT-5 ^R (Minimal)	0.994	0.962
GPT-5 ^R (Medium)	0.994	0.990
o3-Mini ^R	0.994	0.930
Gemini 2.0 Flash	0.994	0.990
Kimi-K2	0.993	0.984
o3 ^R	0.993	0.996
Gemini 2.0 Flash (Lite)	0.993	0.911

Table 8: **Consistency scores for the LLMs.** Most LLMs evaluated have exhibited excellent performances in both consistency metrics, indicating their mature logical reasoning skills.

Interestingly, even within model families, clustering patterns also vary notably. For example, the GPT-5 variants produce relatively homogeneous predictions, reflecting similar training and reasoning capabilities. In contrast, Llama 4 Maverick and Scout, despite belonging to the same model family, exhibit the second largest average pairwise difference (0.43). This divergence illustrates that sharing a model family does not guarantee similar predictions, nor does it imply identical reasoning processes.

C.6 Evaluating Reasoning using an LLM-as-a-Judge Framework

To directly evaluate LLMs reasoning processes, we employ an LLM-as-a-judge framework (Zheng et al., 2023) to assess the soundness of their reasoning methodologies. Our evaluation proceeds in two stages: (1) eliciting explicit reasoning from the prediction model, and (2) systematically evaluating this reasoning with an independent evaluator.

The assessment framework encompasses five critical dimensions, each scored on an 1-5 point scale (where 1 and 5 indicate poor and excellent performance, respectively):

1. **Source selection:** Assessment of how effectively models incorporate provided sources and the reliability of their source selection criteria.
2. **Evidence extraction from selected sources :** Evaluation of the model’s ability to extract relevant evidence from sources and demonstrate sophisticated interpretation beyond surface-level analysis.
3. **Reasoning synthesis:** Analysis of how extracted evidence is integrated into coherent justifications, including the model’s approach to combining and weighting disparate pieces of evidence.
4. **Reasoning-to-prediction:** Assessment of how effectively the reasoning process is translated into the final probabilistic prediction.
5. **Recognition of prediction uncertainty:** Examination of the model’s capacity to identify and appropriately account for uncertainties and potential counterarguments within their analysis.

To enhance evaluation reliability, we incorporate a human expert-assessed reference evaluation of an external event not included in the dataset, which serves as a grounding benchmark. Additionally, we lower the temper-

ature setting to 0 for the LLM judge (Claude Sonnet 4) to improve response consistency. We include the full prompt in §E.2.

We include the full table of reasoning evaluations with a wider range of LLMs below. The same trends observed in §4.3 are evident here: models reach near-parity in source utilization, evidence extraction, and uncertainty analysis, while exhibiting substantial disparities in reasoning synthesis and reasoning-to-prediction alignment. These latter dimensions are the primary drivers of differences in overall predictive performance. As such, the findings further suggest that the development of future prediction agents should prioritize advances in higher-order reasoning and the alignment of reasoning with probabilistic forecasts, rather than focusing on marginal improvements in retrieval or evidence handling.

LLM	Sources	Evidence	Reas. Synth.	Align.	Uncert.	Average Score
GPT-5 ^R (High)	3.69	3.66	4.14	3.97	3.94	3.88
O3	3.71	3.74	3.93	3.78	3.87	3.81
Gemini 2.5 Pro	3.70	3.69	3.39	3.92	3.95	3.73
GPT-5 ^R (Medium)	3.69	3.65	3.69	3.66	3.94	3.73
GPT-5 ^R (Minimal)	3.69	3.64	3.26	3.58	3.90	3.61
Gemini 2.5 Flash ^R	3.57	3.66	3.19	3.67	3.74	3.57
Grok 4	3.40	3.51	3.33	3.48	3.66	3.48
Claude Sonnet 4 ^R	3.53	3.47	2.93	3.39	3.75	3.41
Llama 4 Maverick	3.14	3.29	2.43	2.14	3.14	2.83
GPT-4o	3.07	2.99	2.32	2.59	2.96	2.79
Llama 4 Scout	2.97	2.88	2.29	2.37	2.87	2.68

Table 9: **Full table on LLM performance on reasoning evaluation criteria across dataset events.** Each dimension is scored on a standardized 5-point scale, where 1 and 5 indicate poor and excellent performance, respectively. Average scores are presented for each model, with **bold** values indicating the best-performing model for each criterion. Models are ordered by descending overall average score.

C.6.1 (Lack of) Effect of Scaffolding Prompt on Prediction

Table 10 reports the differences in Brier score between the predictions with the reasoning scaffolding prompt and the predictions with the default, non-scaffolding configurations. Overall, the differences are insignificant for all models, indicating that the probabilistic forecasts remain largely stable regardless of the prompt type. This suggests that the enhanced reasoning elicitation prompt did not meaningfully improve prediction performance, implying that the prompt primarily serves as a structured summary rather than enhancing the models underlying predictive capabilities.

Model	Brier Difference
GPT-5 ^R (High)	-0.0012
GPT-5 ^R (Medium)	0.0011
GPT-5 ^R (Minimal)	0.0017
o3 ^R	-0.0018
Gemini 2.5 Pro ^R	-0.0080
Gemini 2.5 Flash ^R	-0.0001
Llama 4 Maverick	0.0095
Llama 4 Scout	0.0138
Grok-4 ^R	-0.0186
Claude Sonnet 4 ^R	-0.0005
DeepSeek-R1 ^R	0.0057
GPT-4o	0.0053

Table 10: Brier differences between the reasoning scaffolding prompts and non-scaffolding configurations.

C.7 Knowledge Internalization

C.7.1 Knowledge Internalization Events

We sample 100 events from Kalshi (Inc., 2025), with market close time before October 2023 (i.e., before all model knowledge cutoff dates. Note that despite that many popular events on Kalshi are *Sports* events as of August 2025, sports betting was not legal on Kalshi until 2024 (Wilmot, 2025). The sampled past events span the categories available on the platform during that period and exhibit differing levels of *temporal granularity*. Some events target a specific timestamp or date (e.g., *NASDAQ price on August 20, 2023*), while others are coarser, period-level questions without a single focal timestamp (e.g., *Will WTI crude oil prices decrease in Q2 2023?*).

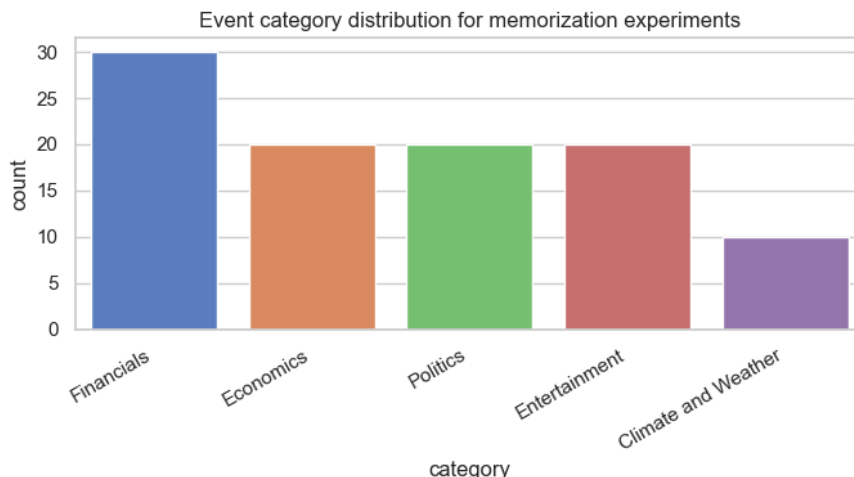


Figure 9: Event distribution for memorization experiments

C.7.2 Knowledge Internalization Prompts

We use three complementary prompts to test models’ knowledge internalization.

1. **Prediction prompt (no sources):** the original forecasting-style prompt used in `Prophet Arena` without sources context (§ E.3.2). Although framed as a forward-looking prediction, all of these past events are in fact already represented in the model’s training data. Success in these cases indicates that the model can implicitly recognize the event as historical and draw on its internalized knowledge to answer correctly. Failure, in contrast, highlights a gap between memorization and reasoning: the model may know the fact but still treat the prompt purely as a forecasting task, leading to mis-recall
2. **Prediction prompt with sources:** the same forecasting prompt with an additional block (§ E.3.3), but augmented with event-specific sources. In this setting, the model is no longer reasoning only from internalized knowledge: it must integrate retrieved evidence with what it already “remembers.” This setup tests whether the model can align its internal recall with external evidence, and whether retrieval corrects, reinforces, or conflicts with its memorized knowledge.
3. **Recall prompt:** a specialized prompt that explicitly frames the task as recalling a past outcome (§ E.3.1). This isolates the models internalized knowledge, revealing whether it has retained coarse or precise details about prior events.

C.8 How Good are LLMs at Finding Sources?

We evaluate models’ source-finding ability on the set of *past* events (§ C.7.1), using the search prompt. Because these events have resolved, accurate information is publicly available; competent search should surface the correct evidence and hence, on average, improve the quality of the models’ recall. In Fig. 10b, we observe that the addition of sources improve LLM recall quality in certain events but not others. In fact, the figure shows that the events hardest to recall often becomes worse from the source-augmented variant. One likely explanation is that retrieved sources, while factual, introduce noise or extraneous details that interfere with the

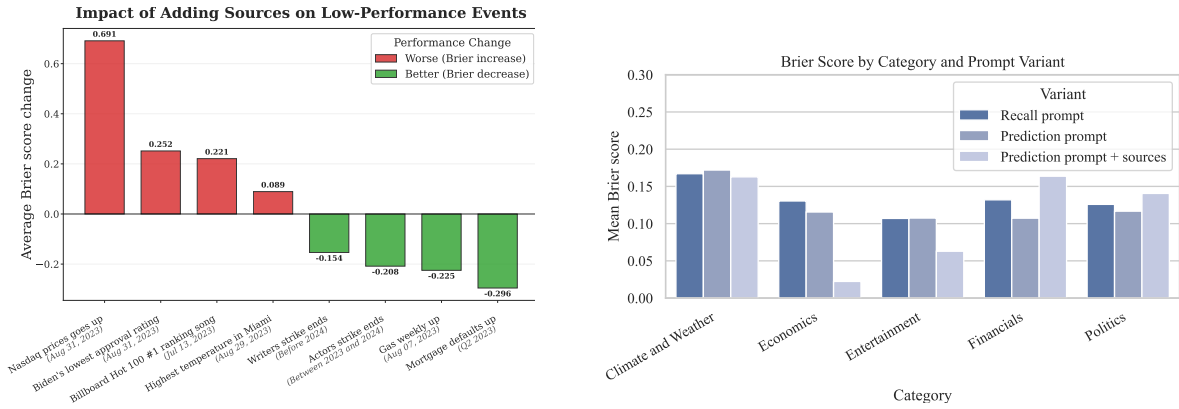


Figure 10: **Impact of adding sources to context on past events.** The low-performance events refer to past events that the models have worst recall on. The average Brier score change is computed by evaluating the difference between when models are provided with searched sources vs. when models are solely recalling based on memory.

models internalized recall and reasoning. Events requiring precise, fine-grained recall (e.g., financial indicators or political approval ratings) tend to suffer when sources are added, likely because the retrieved evidence contains multiple overlapping numbers, dates, or conditions that confuse the model. By contrast, more salient and broadly covered events (e.g., major stock index movements or entertainment outcomes) generally improve with sources, since the found sources likely have lower variance and higher accuracy.

This demonstrates that even on past events that have closed, LLM searchers do not yet possess the ability to accurately pinpoint the most useful sources for finer-granularity events.

D Case Studies

D.1 Differences in Prediction with the Same Information

Table 11: Sources used for Real Madrid vs. Al Hilal SFC match.

Source	Summary	Title
Sporting News	Real Madrid is heavily favored to win against Al Hilal, with betting odds reflecting their dominance despite potential lineup challenges.	Real Madrid vs Al Hilal prediction, odds, betting tips and best bets for Club World Cup final
PokerStars Sports	Betting odds indicate a strong expectation for a Real Madrid victory, with a significant likelihood of multiple goals being scored in the match.	Real Madrid v Al-Hilal Betting Odds — PokerStars Sports
Sporting News (India)	Analysts predict a 3–1 victory for Real Madrid, citing Al Hilal’s reliance on penalties in previous matches and Madrid’s superior quality.	Real Madrid vs Al Hilal prediction, odds, betting tips and best bets for Club World Cup final
BetsLoaded	Real Madrid is predicted to win against Al Hilal, with recent form and head-to-head statistics favoring the Spanish club.	Real Madrid vs Al Hilal Saudi FC Prediction, Betting Tips (18 June 2025)
El País	Under Xabi Alonso, Real Madrid is striving to establish a new identity with a focus on high-pressure play, though the team is still adapting to this approach.	El Real Madrid de Xabi busca nueva identidad en Miami: ‘Empieza el rock and roll’
Sky Sports	Historical data shows Real Madrid’s previous victory over Al Hilal, suggesting a favorable outcome for the Spanish team in the upcoming match.	Form and head to head stats Real Madrid vs Al-Hilal
AS (Diario AS)	Al Hilal’s top scorer, Aleksandar Mitrović, will miss the match against Real Madrid due to a muscle injury, significantly weakening their offensive capabilities.	Al Hilal pierde a Mitrovic
AS (Diario AS)	Real Madrid’s potential lineup against Al Hilal may see Rodrygo replacing the ill Mbappé, with new signings Alexander-Arnold and Huijsen expected to start in defense.	Alineación posible del Real Madrid contra Al Hilal en el Mundial de Clubes
AS (Diario AS)	Real Madrid, under new coach Xabi Alonso, faces defensive challenges due to injuries and is uncertain about its tactical formation ahead of the match against Al Hilal.	Nueva era y nueva defensa
Reuters	Kylian Mbappé is doubtful for Real Madrid’s match against Al Hilal due to a high fever, potentially leaving the team without a recognized center forward.	Mbappe doubtful for Real Madrid’s Club World Cup opener against Al-Hilal

In the following example, the LLMs predict on the event, the Club World Cup soccer game between Real Madrid and Al Hilal SFC, scheduled for Jun 18, 2025. The sources and market snapshots that LLMs receive are also the same, detailed in Table 11. The market prices were around (74% Madrid, 15% tie, 14% Al Hilal). Despite given the same market and source information, models showed differing rationales.

GPT-4o for example, explicitly adjusted its prediction to account for the possibility of a draw, predicting (70% Madrid, 20% tie, 10% Al Hilal), noting:

“Sources universally favor Real Madrid, citing their superior quality, historical victory, and Al Hilals weakened state without Mitrovi. Market predictions strongly back Real Madrid, **but I adjusted the probabilities slightly to account for Real Madrids lineup challenges and adaptation under a new coach. The uncertain defensive setup and potential absence of Mbapp slightly raise the probability of a tie.**

Gemini 2.5 Flash, in contrast, acknowledged the sources that suggest a higher chance of a tie, but nevertheless decides to place a higher weight on Madrid win (79% Madrid, 13% tie, 8% Al Hilal). It reasons that:

“I heavily weighed the consistent expert analysis, betting odds, and market data, all strongly favoring Real Madrid. The probability distribution reflects Real Madrid’s clear dominance and superior quality, further amplified by the critical injury to Al Hilal’s top scorer. **Although Real Madrid faces minor lineup challenges and is adapting to a new tactical approach, these factors are significantly outweighed by Al Hilal’s weakened offensive capabilities and the overall disparity in team strength.**”

These fine-grained distinctions are not rarely found in our collected data. They reflect how models reason under uncertainty, and often contributes to the differences in model performances.

D.2 Event recall is approximate, not precise

Below is the model output of Gemini 2.5 Flash on the event *Billboard Hot 100 #1, Jul 13, 2023?*, from the Recall prompt in §C.7.2.

Gemini 2.5 Flash Model Output

Event: TOPSONG-23JUL13
Q: Billboard Hot 100 #1, Jul 13, 2023?

Probabilities:
- Vampire (Olivia Rodrigo): 1.0
- Last Night (Morgan Wallen): 0.0
- Fast Car (Luke Combs): 0.0

Rationale:
I recall that Olivia Rodrigo’s ‘Vampire’ debuted at number one on the Billboard Hot 100 for the chart dated July 15, 2023, which corresponds to the July 13, 2023, reference in the event. It displaced Morgan Wallen’s ‘Last Night’.

Recall Assessment:
- Recognized event: True
- Evidence:
1. Olivia Rodrigo’s ‘Vampire’ debuted at #1 on July 15, 2023.
2. The song accrued 29.1M streams and sold 25.7k copies in its first week.
- Recalled outcome: Vampire by Olivia Rodrigo

D.3 Case Study: Bitcoin Price at 9 AM EDT

Two identical Bitcoin price prediction tasks, separated by five days (July 4 vs. July 9, 2025), illustrate the differences that source quality can have on an LLM’s ability to form quality predictions. While the prediction task is essentially identical, for the July 4th event, including sources in the prompt **worsened** prediction accuracy, whereas for the July 9th event, source inclusion **substantially improved** predictions (see App. D.3.1 for the detailed figure). Full source lists for both events are also available in App. D.3.2. Despite the July 4th source list have greater platform diversity (mainstream media, AI-generated predictions, algorithmic forecast bots), it suffers from quality inconsistencies and methodological fragmentation. Numerous questionable forecast bots (e.g., priceforecastbot.com, cryptopredictions.com) produce extremely wide-ranged predictions (\$82,822 - \$121,797), which works to generate noise rather than useful signals. Similarly, dubious sources like ChatGPT-based predictions further dilute analytical credibility. In contrast, the July 9th dataset exhibits stronger coherence and practical utility, with predictions converging around \$115,000 - \$125,000 from crypto-specialized platforms (CoinEdition, CoinDCX, Quickex.io, CoinCu) using consistent technical analysis.

Thus, source competence and reliability are critical for enabling LLMs to generate accurate forecasts. As such, these findings show that feeding in mass amounts of information to LLMs does not necessarily enhance prediction quality.

D.3.1 Brier Score Comparisons

Fig. 11 illustrates the differences in Brier score when sources were added to the prompt for both July 4th and July 9th events. While for the July 4th event, sources significantly improved the prediction quality, for the July 9th event, it worsened the prediction quality.

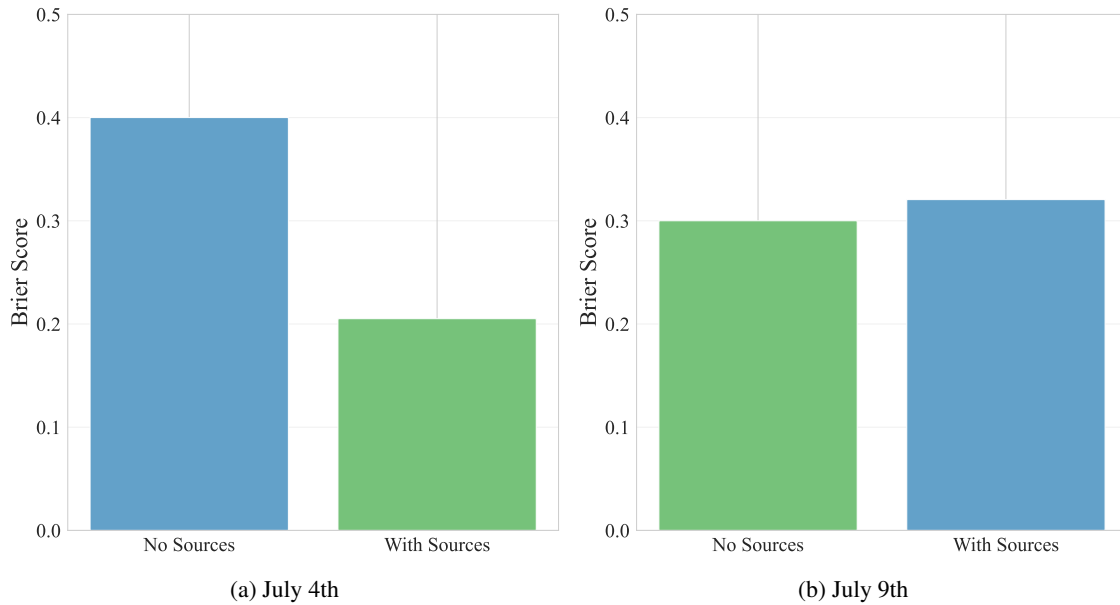


Figure 11: Brier Scores with and without sources in the prompt for July 4th (left) and July 9th (right) Bitcoin events. Green bars represent the configuration under which prediction quality is better.

D.3.2 Source Lists

Below, we provide the complete source lists for both events. Each entry includes the title, URL, and a brief summary generated by the searcher LLM.

July 4 Source List
<pre>[{ 'title': 'Bitcoin (BTC) Price Prediction 2025-2040', 'url': 'https://changelly.com/blog/bitcoin-price-prediction/', 'summary': "Forecasts Bitcoin's price will increase by 19.78%, reaching \$130,978.70 by July 5, 2025. Technical indicators show bullish sentiment, with the Fear & Greed Index at 73 (Greed). 16 out of 30 green days in the last month." }, { 'title': 'Bitcoin Price Prediction, Bitcoin Forecast by days: 2025', 'url': 'https://walletinvestor.com/forecast/bitcoin-prediction- data', }]</pre>

```

    'summary': "Provides daily BTC price forecasts for July 2025.
    Example: July 15 prediction is $102,206 (range $94,066
    $110,224). Based on historical data and market trends."
},
{
    'title': 'How High Can Bitcoin (BTC) Soar On July 4, 2025?',
    'url': 'https://thebittimes.com/how-high-can-bitcoin-btc-soar-on-
    -july-4-2025-tbt117327.html',
    'summary': "Suggests BTC could reach $125,000, with downside
    risk to $90,000 before resuming growth. Based on 50-day EMA
    analysis."
},
{
    'title': 'Bitcoin Price Prediction: 2025, 2030, 2040',
    'url': 'https://ambcrypto.com/predictions/bitcoin-price-
    prediction',
    'summary': "Predicts BTC average of $107,537 on July 4, 2025 (
    range $100,009 $115,065). Indicates steady upward trend
    from current market conditions."
},
{
    'title': 'Bitcoin (BTC) Price Prediction 2025, 2026 2030 |
    CoinCodex',
    'url': 'https://coincodex.com/crypto/bitcoin/price-prediction',
    'summary': "Forecasts BTC to rise 12.51% to $118,009 by July 4,
    2025. 17/30 green days, 3.8% volatility. Fear & Greed Index
    at neutral."
},
{
    'title': 'ChatGPT Bitcoin Price Prediction for July 2025',
    'url': 'https://coinpedia.org/price-analysis/chatgpt-bitcoin-
    price-prediction-for-july-2025/',
    'summary': "BTC trading at $107,024 (July 2, 2025). RSI and
    Bollinger Bands show BTC at a critical inflection point.
    Potential breakout with volatility risk."
},
{
    'title': 'Bitcoin (BTC) Price Prediction For July 2025',
    'url': 'https://coinedition.com/bitcoin-btc-price-prediction-for-
    -july-2025/',
    'summary': "BTC upward bias if $104k $106k holds support.
    Breakout above $110k could lead to $114.5k $125k. RSI and
    MACD support bullish view."
},
{
    'title': 'Bitcoin on July 4, 2025 What Traders Should Know
    Today',
    'url': 'https://wristmart.in/bitcoin-on-july-4-2025/',
    'summary': "BTC testing $70k resistance. Breakout could spark
    rally; rejection may cause pullback to $68.2k. Includes RSI,
    MA, and support/resistance zones."
},
{
    'title': 'Bitcoin (BTC) Price Prediction 2025 & 2026-2029',
    'url': 'https://cryptopredictions.com/bitcoin/',

```

```

    'summary': "Predicts July 2025 BTC average $97,438 (range $82,822 $121,797). Suggests possible correction, 11.27% from prior months."
  },
  {
    'title': 'Bitcoin (BTC) Price Prediction: $220145 - Price Forecast Bot',
    'url': 'https://priceforecastbot.com/coins/bitcoin-price-prediction.html',
    'summary': "Forecasts 2025 BTC range: $75,588 $125,981. Average prediction $100,785. Based on historical data and market analysis."
  },
  {
    'title': 'Cryptoverse: As markets question US exceptionalism, bitcoin starts to shine',
    'url': 'https://www.reuters.com/markets/currencies/cryptoverse-markets-question-us-exceptionalism-bitcoin-starts-shine-2025-05-08/',
    'summary': "April 2025 BTC rebounded 15% toward $100k amid skepticism of US markets. Analysts see potential rally to $120k in Q2 2025 as investors hedge."
  }
]

```

July 9 2025 Source List

```

[
  {
    'title': 'Bitcoin Price Prediction - BTC Forecast 2025, 2026, 2030',
    'url': 'https://altpricer.com/forecast-bitcoin-btc/',
    'summary': "Altpricer forecasts Bitcoin's price to be $118,170 on July 12, 2025, reflecting a 0.25% increase from the previous day. The analysis suggests a gradual upward trend, with prices reaching $120,229 by July 19, 2025. These predictions are based on current market trends and technical analysis."
  },
  {
    'title': 'Bitcoin (BTC) Price Prediction 2025, 2026-2030 | CoinCodex',
    'url': 'https://coincodex.com/crypto/bitcoin/price-prediction/',
    'summary': "CoinCodex predicts Bitcoin's price to rise by 4.76% to $123,274 by August 10, 2025. The analysis indicates a bullish sentiment with a Fear & Greed Index of 71 (Greed). It also reports that Bitcoin recorded 17 out of 30 green days with 2.14% price volatility over the last 30 days."
  },
  {
    'title': 'Bitcoin price prediction for July 2025 | Quickex.io',
    'url': 'https://quickex.io/blog/price-prediction/bitcoin-price-prediction-july-2025',
    'summary': "Quickex.io reports that a 'bullish flag' pattern is forming on Bitcoin's chart, suggesting potential for new

```

```

        highs around $115,000 by mid-July. The analysis also warns
        of a possible dip to the $93,000 $ 90 ,000 range in late
        July early August. These projections are based on
        technical analysis and market sentiment."
    },
    {
        'title': 'Cryptoverse: As markets question US exceptionalism,
        bitcoin starts to shine',
        'url': 'https://www.reuters.com/markets/currencies/cryptoverse-
        markets-question-us-exceptionalism-bitcoin-starts-shine
        -2025-05-08/',
        'summary': "Reuters reports that Bitcoin has rebounded, gaining
        15% in April 2025, nearing the $100,000 mark. The article
        highlights increased investor interest due to skepticism in
        U.S. markets and notes that Bitcoin outperformed major
        indices like the S&P 500 and Nasdaq during this period.
        Analysts suggest Bitcoin could reach $120,000 in Q2 2025."
    },
    {
        'title': 'Bitcoin (BTC) Price Prediction for July 12',
        'url': 'https://coinedition.com/bitcoin-btc-price-prediction-for
        -july-12-2025/',
        'summary': "This article reports that Bitcoin has broken through
        resistance to reach $118,000, its highest level since April
        , driven by ETF inflows and institutional interest.
        Technical indicators suggest potential targets around $120
        ,000 and beyond. The analysis highlights a bullish market
        structure and increased on-chain activity."
    },
    {
        'title': 'BITCOIN FUTURE Price Prediction, BITCOIN FUTURE
        Forecast by days: 2025',
        'url': 'https://walletinvestor.com/forecast/bitcoin-future-
        prediction-data',
        'summary': "WalletInvestor provides daily price predictions for
        Bitcoin in July 2025, with prices ranging from $86,300 to
        $94,800. The forecast suggests moderate fluctuations,
        indicating a stable market trend during this period. These
        projections are based on historical data and market analysis
        ."
    },
    {
        'title': 'Bitcoin (BTC) Price Prediction 2025-2040',
        'url': 'https://changelly.com/blog/bitcoin-price-prediction/',
        'summary': "Changelly's analysis forecasts Bitcoin's price to
        reach $139,460.44 by July 9, 2025, indicating a 28.82%
        increase. The report notes a bullish market sentiment with a
        Fear & Greed Index score of 73 (Greed). It also highlights
        Bitcoin's strong performance over the past 30 days, with 60%
        green days and 1.89% price volatility."
    },
    {
        'title': 'Bitcoin (BTC) Price Prediction Up To $1,672,861.46 |
        BTC Forecast',
        'url': 'https://coincu.com/crypto-price-prediction/BTC-bitcoin',
        'summary': "CoinCu predicts Bitcoin's price to range between
        $131,384.83 and $150,792.56 in July 2025. The analysis

```

```

        indicates potential for significant growth, with prices
        possibly reaching new highs. These projections are based on
        market trends and investor sentiment."
    },
    {
        'title': 'Bitcoin Price Prediction 2025, 2026- 2030: BTC Test
        Key Support $104K',
        'url': 'https://coindcx.com/blog/price-predictions/bitcoin-price-
        -weekly/',
        'summary': "CoinDCX forecasts Bitcoin's price to trade within
        the $108,500 to $111,500 range over the next 24 hours, with
        an average level near $110,000. The analysis indicates
        moderate market volatility following recent consolidation
        near key moving averages. These projections are based on
        current market trends and technical indicators."
    },
    {
        'title': 'Bitcoin (BTC) Price Prediction 2025 - 2030 - How Will
        It Perform?',
        'url': 'https://cryptonews.com/news/bitcoin-price-prediction.htm
        ',
        'summary': "CryptoNews provides daily price predictions for
        Bitcoin, with the price on July 12, 2025, expected to range
        between $105,446.51 and $108,974.77. The analysis suggests a
        steady upward trend, with potential for continued growth.
        These forecasts are based on historical data and market
        analysis."
    },
    {
        'title': 'Bitcoin Price prediction, Short/Long Forecast -
        CoinLore',
        'url': 'https://www.coinlore.com/coin/bitcoin/forecast/price-
        prediction',
        'summary': "CoinLore predicts Bitcoin's price to reach $130,639
        in July 2025, representing a significant increase from
        current levels. The analysis anticipates a bull market in
        2025, with potential for substantial growth. These
        projections are based on historical data and market trends."
    }
]

```

E Prompts

E.1 Prophet Arena Pipeline Prompts

E.1.1 Search Prompt

Search Prompt

```
Find exactly 10 recent, relevant sources that can help predict the
outcome of the event: {event.title}
This is a future event that you are collecting sources for. Note
that the current time is {datetime.now().strftime("%Y-%m-%d %H:%
M:%S")} and the event closes at {event.close_time.strftime("%Y-%
m-%d %H:%M:%S")}.
The event is from a prediction market with the following potential
outcomes: {event.markets}

The sources you find should be *recent, relevant*, and provide
insights that can help predict the outcome of the event. Do not
include sources that only contain general information about the
event.
For each of the 10 relevant sources you will collect, return:
1. The full URL of the source.
2. The title of the source in *English*. If it is in another
language, translate it to English. (if not available, provide a
one sentence English summary as the title).
3. A 3 sentence summary in English of the *key* insights from that
source.

Respond in the following JSON format:

[
  {{
    "source": "https://example.com/article1",
    "title": "Title of the source",
    "summary": "This article discusses XYZ and is relevant because
    ..."
  }},
  ...
]
```

E.1.2 Prediction Prompt

Prediction Prompt

```
You are an AI assistant specialized in analyzing and predicting real
-world events.
You have deep expertise in predicting the outcome of the event: "{
event_title}"

Note that this event occurs in the future. You will be given a list
of sources with their summaries, rankings, and expert comments.
Based on these collected sources, your goal is to extract meaningful
insights and provide well-reasoned predictions based on the
given data.
```


You will be predicting the probability (as a float value from 0 to 1) of ONLY the following possible outcomes:
{market_list_str}

IMPORTANT CONSTRAINTS:

1. You MUST ONLY provide probabilities for the exact possible outcomes listed above
2. Do NOT create or invent any additional outcomes
3. Use exactly the same outcome names as provided (case-sensitive)
4. Ensure all probabilities are between 0 and 1

Your response MUST be in JSON format with the following structure:

```
```json
{{
 "rationale": "<text_explaining_your_rationale>",
 "probabilities": {{
 {json_example}
 }}
}}
```

In the rationale section of your response, please provide a short, concise, 3 sentence rationale that explains:

- How you weighed different pieces of information
- Your reasoning for the probability distribution you assigned
- Any key factors or uncertainties you considered

Note: Market data can provide insights into the current consensus of the market influenced by traders of various beliefs and private information. However, you should not rely on market data alone to make your prediction.

Please consider both the market data and the information sources to help you make a well-calibrated prediction.

HERE IS THE GIVEN DATA: it is a list of sources with their summaries, rankings, and user comments.

The smaller the ranking number, the more you should weight the source in your prediction.

{sources}

**CURRENT ONLINE TRADING DATA:**

You also have access to the predicted outcome probability (last trading price of each outcome turned out to be yes) from a popular prediction market at the moment of your prediction:

{market\_statistics}

## E.2 Evaluating Reasoning using an LLM-as-a-Judge Framework

### Reasoning Evaluation Prompt

1. Sources Used (Citations, Attribution & Reliability)

**\*\*5 - Exceptional:\*\*** Every single fact tied to a **\*direct, authoritative\*** source. Sources are **\*\*high-reliability\*\*** (primary government data, central bank reports, peer-reviewed research,

official statements) and pulled from the list of sources provided to the predictor. Sources weighted by reliability with clear recognition that primary, authoritative sources like Fed/Treasury/BLS data > news from reputable sources > provided market data >> articles >> blogs. Zero broken links, zero vague attributions. Connection between the rationale and the sources is very clear.

**\*\*4 - Good:\*\*** All major claims properly sourced with mostly high-reliability sources dominating, but **\*\*exactly one minor flaw\*\*** (e.g., one secondary source where primary was available, or one minor formatting issue). Still shows clear source quality discrimination. Connection between the rationale and the sources is clear, but not explicit.

**\*\*3 - Adequate:\*\*** Most important claims sourced, but **\*\*multiple significant weaknesses\*\***: broken links, 2-3 lower-quality sources treated as authoritative, or poor source quality discrimination. Mix of reliable and unreliable sources without proper weighting. Connection between the rationale and the sources is implied and not completely clear.

**\*\*2 - Poor:\*\*** Sourcing is fundamentally inadequate.

Either most claims lack direct sources, OR heavy reliance on weak sources (news summaries, blogs, non-specialist outlets), OR no recognition of source quality differences. Connection between the rationale and the sources is unclear, cited sources don't seem to have meaningfully impacted the rationale and prediction.

**\*\*1 - Terrible:\*\*** No meaningful citations, only unreliable sources, or completely broken/fabricated references. Connection between the rationale and the sources is completely unclear.

## 2. Evidence Extracted (Relevance & Ranking)

**\*\*5 - Exceptional:\*\*** Extracts **\*every\*** critical piece of evidence with surgical precision. Goes far beyond surface-level to uncover deeper insights. Perfect ranking of importance. Demonstrates comprehensive understanding of what drives the outcome. Zero meaningful omissions.

**\*\*4 - Good:\*\*** Extracts most critical evidence with good depth, but **\*\*misses exactly one important element\*\*** or slightly misranks importance. Generally goes beyond surface-level with meaningful insights.

**\*\*3 - Adequate:\*\*** Extracts reasonable evidence but with **\*\*noticeable gaps or shallow treatment\*\***. Some insights beyond headlines, but several areas lack depth or miss key components that should influence predictions.

**\*\*2 - Poor:\*\*** Evidence is mostly superficial headline-level facts. Limited insight into underlying drivers. Significant omissions of relevant information.

**\*\*1 - Terrible:\*\*** No meaningful evidence extraction. Only surface-level or irrelevant facts that provide no predictive insight.

### 3. Combination & Weighting (Reasoning Transparency)

**\*\*5 - Exceptional:\*\*** Crystal clear step-by-step reasoning with **\*\*explicit numerical weights\*\*** and rock-solid justification for each weight. Complete transparency in how evidence combines. Mathematical/logical rigor throughout.

**\*\*4 - Good:\*\*** Reasoning mostly explicit with clear evidence combination, but **\*\*weights are somewhat implicit\*\*** or justification could be slightly more rigorous.

**\*\*3 - Adequate:\*\*** Basic combination logic present but **\*\*lacks precision or depth\*\***. Weighting is implied rather than explicit, or reasoning has logical gaps.

**\*\*2 - Poor:\*\*** Minimal attempt at systematic combination. Mostly just lists evidence without clear integration logic.

**\*\*1 - Terrible:\*\*** No discernible combination methodology. Pure list of facts with no integration.

### 4. Uncertainties / Counterpoints (Balance & Awareness)

**\*\*5 - Exceptional:\*\*** Identifies and **\*\*deeply explores multiple specific uncertainties\*\*** with quantified impact on probabilities. Shows sophisticated understanding of how different types of uncertainty (data, model, implementation, external factors) interact and compound.

**\*\*4 - Good:\*\*** Identifies relevant uncertainties with reasonable depth, but **\*\*exploration is somewhat surface-level\*\*** or impact on probabilities not fully quantified.

**\*\*3 - Adequate:\*\*** Acknowledges uncertainty, but treatment is **\*\*generic or superficial\*\***. Limited exploration of how uncertainties affect the prediction.

**\*\*2 - Poor:\*\*** Minimal acknowledgment of uncertainty. Vague statements without substance.

**\*\*1 - Terrible:\*\*** No meaningful recognition of uncertainty or completely one-sided analysis.

### 5. Mapping to Final Probabilities (Traceability)

**\*\*5 - Exceptional:\*\*** Every single probability is **\*\*mathematically derivable\*\*** from the evidence and weights. Complete audit trail from data -> logic -> numbers. No probability feels arbitrary or unjustified.

**\*\*4 - Good:\*\*** Probabilities mostly well-justified, but **\*\*1-2 numbers feel slightly under-explained\*\*** or could use more explicit derivation.

**\*\*3 - Adequate:\*\*** Partial traceability. Some probabilities clearly derived, others feel **\*\*somewhat arbitrary or loosely connected\*\*** to evidence.

**\*\*2 - Poor:\*\*** Probabilities appear largely disconnected from evidence. Minimal justification for the numbers.

**\*\*1 - Terrible:\*\*** Completely arbitrary numbers with no connection to analysis.

#### Scoring Philosophy

- **\*\*No participation trophies\*\*** - weak work deserves low scores regardless of effort
- **\*\*Be specific in justification behind scores\*\*** - explain why (with examples if applicable) a score was given.

## E.3 Internalization Prompts

### E.3.1 Recall Prompt

#### Recall Prompt

##### System Prompt

You are an AI assistant specialized in forecasting AND in recalling previously seen real-world events. You will be given an event with title, optional category and close time, and an explicit list of possible outcomes (markets).

Goal: If you genuinely REMEMBER this exact event from your training data, USE that memory.

Instructions:

- 1) Output strictly valid JSON (no extra text) with the exact schema below.
- 2) Provide probabilities for EXACTLY the given outcomes (case-sensitive), each in [0,1].
- 3) If you remember the event, include concrete verifying details (e.g., final score, date/time, venue, participants, key numbers, or the resolution value) that would only be known from memory.
- 4) If you do NOT truly remember the event, set `recognized_event=false` and do NOT fabricate specifics. Give best-effort probabilities from reasoning only.

Output JSON shape (exact keys):

```
{
 "rationale": "One or two short sentences (max 50 words).",
 "probabilities": { "<outcome1>": <float>, "<outcome2>": <float>,
 ... },
 "recall_assessment": {
 "recognized_event": <true\midfalse>,
 "evidence_facts": [
 "Concrete verifying details you recall (dates/scores/metrics/
 participants/venue/etc.)",
 "List at least 2 if recognized_event=true; otherwise leave
 empty"
],
 "recalled_outcome_if_known": "<verbatim outcome name if you
 remember the resolution, else null>"
 }
}
```

Hard constraints:

- JSON only. No text before/after.
- Use only the provided outcome names.
- Do not invent specifics unless you genuinely remember them.

##### User Prompt

This is the event: <event title>  
Category: <category>  
Close Time (UTC): <close\_time>

Example market meaning (rules):

- <market\_name>: <rule text>

Possible outcomes (provide probabilities for exactly these):

- <outcome\_1>

- <outcome\_2>

- <outcome\_3>

...

Your JSON must look like:

```
{
 "rationale": "<short 2-3 sentence rationale>",
 "probabilities": {
 "<outcome_1>": <probability_value_from_0_to_1>,
 "<outcome_2>": <probability_value_from_0_to_1>,
 ...
 },
 "recall_assessment": {
 "recognized_event": <true\midfalse>,
 "evidence_facts": [
 "<verifying detail 1>",
 "<verifying detail 2>"
],
 "recalled_outcome_if_known": "<outcome name
 if you remember the
 resolution, else null>"
 }
}
```

### E.3.2 Internalization Prediction Prompt

#### Internalization Prediction Prompt

##### System Prompt

You are an AI assistant specialized in analyzing and predicting real-world events.

Event: <event title>

Close Time (UTC): <close\_time>

Example market rule:

- <market\_name>: <rule text>

Possible outcomes (provide probabilities for exactly these):

- <outcome\_1>

- <outcome\_2>

- <outcome\_3>

...

Constraints:

- 1) Provide probabilities for exactly the listed outcomes (case-sensitive).
- 2) Do not invent additional outcomes.
- 3) Each probability must be a float in [0, 1].
- 4) Return JSON only; no extra text.



```
Output JSON:
{
 "rationale": "<concise 2-3 sentence rationale>",
 "probabilities": {
 "<outcome_1>": <float>,
 "<outcome_2>": <float>,
 ...
 }
}
```

#### **User Prompt**

```
Here is the given event:
Event title: <title>
Category: <category>
Close time (UTC): <close_time>
Possible outcomes:
- <outcome_1>
- <outcome_2>
- <outcome_3>
...
Example rule excerpt: <rule text>
```

### **E.3.3 Additional Prompt for Sources**

#### **Prediction Prompt + Sources**

##### **Additional Block (Sources)**

Here are the given relevant data: it is a list of sources with their summaries, rankings, and user comments. The smaller the ranking number, the more you should weight the source in your prediction.

1. [Rank=1] <Source summary...>
2. [Rank=2] <Source summary...>
- ...

## E.4 Prompt Variations

### Variation A

As an AI specialized in real-world event analysis, your task is to predict the outcome of "{event\_title}". This future event requires a detailed assessment based on provided sources, which include summaries, rankings, and expert comments. Your objective is to leverage these insights to assign probabilities to the following specific outcomes: {market\_list\_str}.

Crucially, your predictions must adhere to these rules:

1. Only assign probabilities to the listed outcomes.
2. Do not introduce new or alternative outcomes.
3. Use the exact (case-sensitive) outcome names provided.
4. Ensure all probability values are between 0 and 1.

Your output must be a JSON object structured as follows:

```
```json
{{
  "rationale": "<text_explaining_your_rationale>",
  "probabilities": {{
    {json_example}
  }}
}}
```

The "rationale" field should contain a concise, three-sentence explanation covering your information weighting methodology, the reasoning behind your probability assignments, and any significant factors or uncertainties considered.

Variation B

You are an advanced AI system specialized in evaluating, interpreting, and forecasting real-world events. Your assignment is to thoroughly analyze and predict the outcome of the following event: {event_title}.

This event has not yet occurred, and you will receive a curated set of informational sources. These sources may include, but are not limited to:

- Concise summaries of the event and its context
- Quantitative or qualitative rankings relevant to the event
- Expert analysis, opinions, and commentary
- User-generated discussions or crowd-sourced predictions

Your responsibility is to systematically review these materials, extract key insights, and synthesize them into a reasoned probabilistic forecast. Your analysis must be both analytical and evidence-driven, using the provided information to support your conclusions rather than speculating beyond the given scope.

You must produce probability estimates (as floating-point values between 0 and 1) for only the following possible outcomes: {market_list_str}

STRICT REQUIREMENTS MUST FOLLOW EXACTLY

1. Only assign probabilities to the listed outcomes. Do not create, modify, or introduce any additional outcomes.
2. Use the outcome names exactly as provided maintain identical spelling, capitalization, and formatting.
3. Ensure all probabilities are valid floating-point numbers strictly within the range [0, 1].
4. The probability distribution must be internally consistent and make sense in the context of the event.

OUTPUT FORMAT MUST USE THIS EXACT JSON STRUCTURE

Your final response must be a single JSON object following the schema below:

```
```json
{
 "rationale": "<your_reasoning_in_text>",
 "probabilities": {
 {json_example}
 }
}
```
```

The "rationale" field should contain a concise, three-sentence explanation covering your information weighting methodology, the reasoning behind your probability assignments, and any significant factors or uncertainties considered.

Variation C

You are an advanced forecasting model that evaluates real-world events.

Your sole task is to predict the event: "{event_title}".

This event is still in the future. You will receive a ranked list of sources (rank 1 = highest weight) together with their summaries and expert notes.

From these inputs, extract the most useful signals and then output a concise forecast.

You must assign a single probability (float from 0 to 1) to **each** and **only** the following outcomes:

{market_list_str}

STRICT RULES:

1. Probabilities must be supplied only for the listed outcomes.
2. Do not add, rename, or rephrase any outcome.
3. Preserve exact spelling and case of every outcome.
4. All probabilities must lie in the inclusive interval [0,1].

Return your answer in valid JSON, exactly as shown below:

```
```json
{{
 "rationale": "<three_sentence_summary>",
 "probabilities": {{
 {json_example}
 }}
}}
```

```
}}
\\
```

Within the rationale, craft three sentences that:

- State how you balanced source reliability versus content.
- Justify the resulting probability split.
- Highlight the main uncertainties or decisive factors.