LEAF: A LIVING BENCHMARK FOR EVENT-AUGMENTED FORECASTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Large Language Models (LLMs) are increasingly applied to real-world forecasting tasks, yet evaluating their true predictive capability remains compromised by pre-training data contamination and look-ahead leakage in automated search. Existing benchmarks either rely on static contexts, restrict evaluations to narrow environments, or fail to audit auxiliary textual events for future information leakage. To establish a rigorous evaluation paradigm, we propose LEAF, the first living benchmark for event-augmented forecasting tasks, including trend, event, and time series forecasting. LEAF couples a recursive retrieval agent system with dual-agent cross-validation to gather comprehensive, relevant, and temporally aligned auxiliary context. A comprehensive audit across 500 tasks by 47 domain specialists demonstrates that our pipeline suppresses future information leakage from 8.6% to 1.6%. Across extensive evaluations of 16 frontier proprietary and open-weight models on our benchmark, we show that LLMs effectively extract signals from verified events to boost trend and event forecasting[1].

1 INTRODUCTION

As Large Language Models (LLMs) have demonstrated increasingly sophisticated reasoning capabilities (Google DeepMind, 2026; Anthropic, 2026b; OpenAI, 2026), researchers have explored their potential in decision-making and forecasting tasks, including financial trading (Xiao et al., 2025; Fan et al., 2025), event forecasting (Karger et al., 2025; Halawi et al., 2024), and multimodal time series forecasting (Merrill et al., 2024; Williams et al., 2025). Unlike deterministic reasoning benchmarks, forecasting requires predicting uncertain futures rather than extracting established facts. This probabilistic task requires interpreting weak signals within dynamic environments, serving as a challenging testbed to evaluate LLMs' temporal reasoning and decision-making under uncertainty.

While numerous benchmarks have been introduced to evaluate the predictive abilities of LLMs, existing frameworks suffer from two fundamental vulnerabilities: i) *pre-training contamination in static benchmarks* and ii) *look-ahead leakage in living benchmarks*. Static benchmarks (Chen et al., 2025; Yu et al., 2026; Liu et al., 2024; Li et al., 2024) relying on fixed historical sequences or manually curated context (Williams et al., 2025) fall within the pre-training corpora of newer LLMs. As we demonstrate, evaluating on data preceding training cutoffs artificially inflates apparent forecasting accuracy by up to 147% (as detailed in §3.4), measuring training memorization rather than true predictive reasoning. To avoid contamination, recent "living" benchmarks continuously update test sets (Tan et al., 2024; Xu et al., 2025; Fan et al., 2025). However, dynamically retrieving external news via automated search introduces severe look-ahead leakage, such as post-event summaries, retrospective analyses, or misdated articles that reveal post-cutoff outcomes to the model. Furthermore, current living benchmarks remain restricted to narrow domains or lack verified contextual diversity (Karger et al., 2025; Li et al., 2025).

To address these challenges, we introduce LEAF, a living benchmark for event-augmented forecasting across diverse event-driven domains, including politico-economic events, weather, electricity, stock markets, and commodities. As illustrated in Figure 2, LEAF mitigates pre-training data contamination via a continuously updated "living" test set, employing a retrieval and dual-agent validation system to ensure event breadth and temporal alignment while filtering out fabricated information. To

[1]The code and data are provided here: `https://anonymous.4open.science/r/LEAF-4565`

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

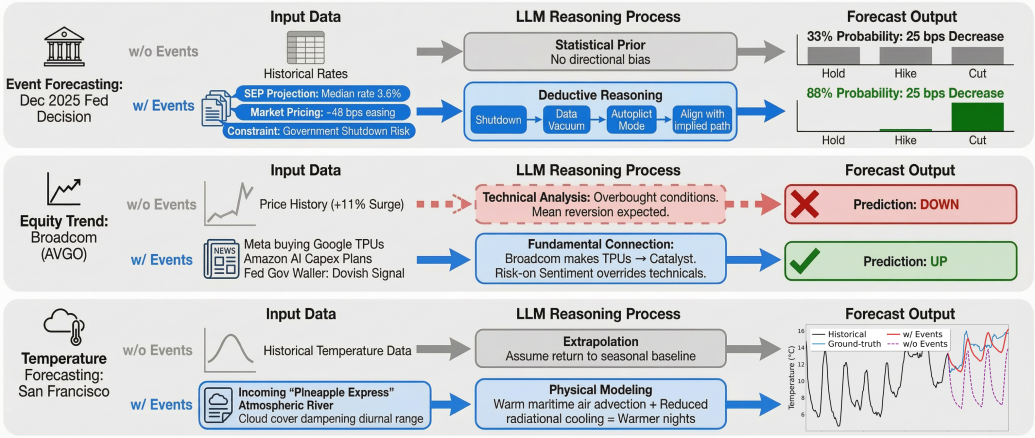

Figure 1: **Example evaluation instances from LEAF** across its three core tasks: discrete event forecasting (top), trend forecasting (middle), and time-series forecasting (bottom). LLMs struggle to make effective predictions without event inputs (w/o Event); however, their performance is enhanced with events provided by our retrieval agents. Further details are in Appendix E.3.

validate benchmark integrity, we conducted an exhaustive evaluation across 500 tasks with 47 professional domain specialists (averaging 13.6 years of professional experience across finance, meteorology, power systems, and geopolitics). Our human-expert audit reveals that while unverified automated search exhibits an 8.6% future look-ahead leakage rate, our dual-agent cross-validation pipeline suppresses future information leakage down to 1.6%, with over 60% of the remaining minor discrepancies bounded by a single day.

Through evaluating sixteen LLMs, including proprietary (Gemini-, Grok-, Claude-Family) and open-weight models (Qwen-, Gemma-Family), we find that proprietary models effectively extract signals from verified events, achieving relative gains of 30.03% in trend forecasting and 25.02% in event probability forecasting over event-free baselines (§4.2), and illustrative examples are provided in Figure 1. Our main contributions are:

- We introduce LEAF, a living benchmark to evaluate LLMs' ability to leverage multidimensional signals for event-driven forecasting in complex environments (§3.1).
- We propose a multi-agent retrieval system to collect events, augmented by a dual-agent cross-validation pipeline that ensures authenticity and temporal alignment (§3.2), which significantly improves event date accuracy and mitigates look-ahead leakage (§3.4).
- Evaluating 16 LLMs reveals fundamental limitations: even with informative text, most models perform merely on par with time-series foundation models (TSFMs) and fail to outperform naive baselines in market price forecasting, empirically confirming the latter task's infeasibility (§4.3).
- In market trend predictability tests, despite the Efficient Market Hypothesis (EMH) (Fama, 1970), we observe that LLMs exhibit capacity to distinguish between strong signals and those already priced in, yielding forecasting improvements in high-confidence scenarios (§4.4).

2 Related Work

Forecasting with LLMs. Many studies employ LLMs for time series forecasting based solely on historical numerical data (Gruver et al., 2023; Zhou et al., 2023; Jin et al., 2024), or incorporate textual information (Wang et al., 2024b; Luo et al., 2025). For instance, Williams et al. (2025) uses manually crafted expected future events to assist LLMs in predicting time series trends and Merrill et al. (2024) uses synthetic text and time series data for context-aided forecasting. Other studies combine LLMs with news to predict the probability of future events (Halawi et al., 2024; Yang et al., 2026). Karger et al. (2025) leverages odds from prediction platform like Polymarket (Polymarket) to aid LLM forecasting. Yang et al. (2026) utilizes GPT-4o to retrieve the top-k news articles to aid LLM live predictions. Additionally, some works directly deploy LLMs for forecasting and trading in markets (Nof1.ai, 2026; Fan et al., 2025; Li et al., 2025). However, the textual information provided

Table 1: Comparison of forecasting benchmarks. Our benchmark includes **three forecast tasks**: event (EF), time series (TSF), and trend forecasting. Instead of relying on APIs or static datasets for text conditions, we employ a **multi-agent system** to ensure comprehensive text coverage. We further compare the benchmarks based on Diversity (**Div**): whether the event incorporates various sentiments and from multiple sources; Quality Control (**QC**): whether textual relevance is validated; and whether the benchmark is **Living**.

Benchmark	Forecast	Context Source	Div	QC	Living
Karger et al. (2025)	EF	API (e.g., Polymarket)	✗	✗	✓
Fan et al. (2025)	Trend	API (e.g., Jina AI (2026))	✓	✗	✓
Nof1.ai (2026)	Trend	None (Season 1 & 2)	✗	✗	✓
Xu et al. (2025)	TSF	Weather API	✓	✗	✓
Yu et al. (2026)	TSF	Records & Synthetic	✓	✗	✗
Chen et al. (2025)	TSF & Trend	Records & Dataset	✗	✓	✗
Williams et al. (2025)	TSF	Manually Crafted	✗	✓	✗
Tan et al. (2025)	EF & TSF	API	✗	✓	✓
LEAF (ours)	EF & TSF & Trend	Multi-Agent System	✓	✓	✓

in these works is often synthesized or constrained by narrow data sources (e.g., specific APIs), which struggles to guarantee comprehensiveness. Moreover, some existing forecasting benchmarks are static, making it difficult to prevent data contamination (Chen et al., 2025; Liu et al., 2024; Yu et al., 2026). To evaluate LLMs' ability to extract informative signals from complex contexts for event-driven forecasting, we propose a living benchmark featuring comprehensive event data.

Living Benchmarks. To mitigate data contamination, several living benchmarks dynamically construct test sets to evaluate general model performance. For instance, LMSYS (Zheng et al., 2024) continuously ranks LLMs via crowdsourced human pairwise evaluations, while LiveBench (White et al., 2025) regularly updates its questions using recent objective data to prevent memorization. Extending this dynamic paradigm, other benchmarks focus on specific capabilities, such as reasoning (Fan et al., 2023; Glazer et al., 2024), coding (Jain et al., 2024; Zheng et al., 2025), and multimodal reasoning (Zou et al., 2024; Bitton et al., 2023; Yang et al., 2024). Living benchmarks and evaluations, as summarized in Table 1, have also been proposed for forecasting tasks (Nof1.ai, 2026; Karger et al., 2025; Li et al., 2025; Tan et al., 2025). For instance, Xu et al. (2025) evaluates LLMs in predicting power generation or traffic flow using weather forecasts and synthetic sensor data. Fan et al. (2025) collects real-time news via APIs for market forecasting. Additionally, Williams et al. (2025) incorporates manually crafted text for time-series forecasting, relying on human annotation to ensure relevance. However, constructing comprehensive context-aided test sets in complex open environments remains challenging due to data scarcity. To bridge this gap, we introduce a real-world dataset and empirically validate the auxiliary text's relevance.

3 LEAF: A Living Benchmark

3.1 Problem Definition

Let $X_{\leq T} = \{x_1, x_2, \ldots, x_T\}$ denote a real-valued historical time series observed up to the present timestamp T. In traditional time series analysis, forecasting models often predict future states by relying solely on historical numerical sequences. Our living event-augmented forecasting benchmark evaluates LLMs' ability to extract informative signals from historical values and events for forecasting. Specifically, the models are conditioned on real-valued inputs $X_{\leq T}$ (when available) and historical events $\mathcal{C}_{\leq T}$ to perform three primary prediction tasks:

- **Event Forecasting (EF)** evaluates the LLM's ability to predict the probability of a discrete future event $P(E_{T+h})$, scheduled to resolve at a future timestamp $T + h$, using available priors.
- **Time Series Forecasting (TSF)** evaluates the ability to predict exact future values $X_{T+1:T+h}$ over a horizon h, conditioned on both numerical and textual histories $(X_{\leq T}, \mathcal{C}_{\leq T})$.
- **Trend Forecasting (TF)** converts exact numerical prediction into a categorical task, given that LLMs excel at logical classification over precise regression. The model predicts the future move-

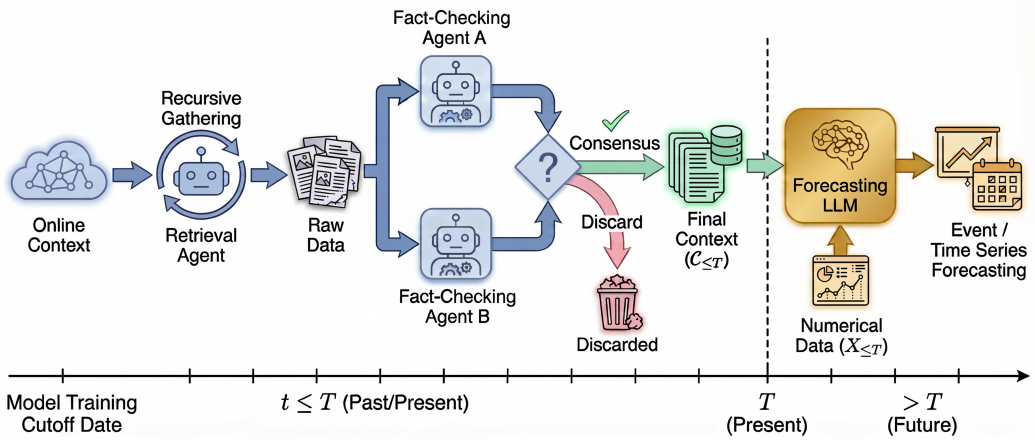

Figure 2: The data retrieval and cross-validation framework enables predictive evaluation in realistic and complex environments. To provide events relevant to the forecasting target, an LLM-based *Retrieval Agent* recursively gathers temporally constrained online context ($t \leq T$). A dual-agent *Fact-Checking Module* then filters this data, discarding unverified information and retaining only events with a strict consensus on both publication time and factual authenticity. The test data is restricted to periods after the models' training cutoffs to prevent pre-training data leakage.

ment direction y_{T+1} as a 3-class or 5-class classification using the prior $(X_{\leq T}, \mathcal{C}_{\leq T})$, where the ground truth is derived from the differential $\Delta x = x_{T+1} - x_T$. Details refer to Appendix D.1.

Table 2: Evaluation data statistics in LEAF across domains, entities, and forecasting tasks. We report the number of raw events prior to filtering, the filtering rate of the dual-agent fact-checking module, the number of days with active events (# Days), the evaluation frequency, and the evaluation horizon.

Domain	Tasks	# Entities	# Events	Filter Rate	# Days	Forecast Freq	Horizon
Polymarket	EF	66	12,162	39.5%	2,796	10%	04.2026–09.2026
Large-cap Stocks	Trend, TSF	30	37,343	41.9%	3,959	Daily	04.2026–09.2026
Mid-cap Stocks	Trend, TSF	40	13,427	51.7%	3,136	Monthly	03.2026–08.2026
Commodities	Trend	10	4,240	72.8%	509	Daily	10.2025–02.2026
Electricity	TSF	3	1,518	26.3%	353	Hourly/Daily	12.2025–04.2026
Weather	TSF	10	5,271	24.0%	1,177	Hourly/Daily	10.2025–02.2026

3.2 METHODOLOGY

Retrieval Agent and Fact-checking Module. In our benchmark, the textual context is gathered recursively by an LLM-based retrieval agent. This agent is augmented with Gemini Grounded Search[2]. As illustrated in Figure 2, for each specified time interval (e.g., daily or weekly), we collect all published information relevant to the forecasting target. To ensure comprehensiveness and avoid duplication, we conduct multiple search iterations, appending previously retrieved event as context for subsequent rounds. Beyond detailed *event descriptions* and *reference date*, the retrieval agent outputs a "*causality*" rationale explaining why the event affects the target, a "*sentiment*" label indicating the impact direction (Positive or Negative), and an "*impact_type*" classification (Direct or Indirect). Furthermore, to ensure data authenticity and temporal alignment, we utilize a dual-agent cross-validation mechanism. Retrieved events are retained to form the final context if and only if two independent fact-checking agents reach a strict consensus regarding their authenticity and public release time; any discordant information is discarded. For the detailed searching and fact-checking prompt, refer to Appendix D.3 (Figure 9 and 10).

Human Expert Data Annotation. We engaged 47 domain experts (averaging 13.6 years of experience across finance, meteorology, energy, geopolitics, etc.) to annotate 500 events. With the

[2] https://ai.google.dev/gemini-api/docs/google-search

implementation of the fact-checking agent, only 1.6% (8 events) exhibited future information leakage, and five of these cases preceded the actual date by only a single day, which has limited impact given our 3-day and 7-day forecasting horizons. In contrast, without the fact-checking agent, 43 events suffered from future information leakage. Additionally, experts identified 1.6% of the events as hallucinations. For further details on the expert annotations, refer to Appendix C.

Stock Market Problem Formulation. Stock market forecasting is notoriously challenging under the Efficient Market Hypothesis (EMH) (Fama, 1970), as most public events are rapidly priced in. Nevertheless, we include equities in LEAF, hypothesizing that rare, strong signals persist that are not immediately absorbed or require time to unfold. To determine whether LLMs can distinguish between these strong signals and priced-in events, we present the models with all stocks at each prediction step, conditioned on $(X_{\leq T}, \mathcal{C}_{\leq T})$. The LLMs rate the signal strength of these 30 (40) stocks on 5 levels. We observe that higher LLM confidence correlates with better performance (§4.4). This improvement is evident in trend forecasting for short-term (1 or 2 days) large-cap stocks and long-term (3 or 6 months) mid-cap stocks, though accurately predicting specific prices remains infeasible (§4.3). Detailed prompts and classification criteria are provided in Figure 22 in the Appendix D.3.

Retrieval Events Validations. To validate our event retrieval pipeline, we construct a "consistent" textual context to assist predictions during evaluation, similar to Chen et al. (2025) in stock domain. Since the ground-truth change $\Delta x = x_{T+h} - x_T$ is known during benchmark formulation, we use it to systematically select events $e \in \mathcal{C}_{\leq T}$ whose expected "*sentiment*", from retrieval agent, aligns with the actual future trajectory. Specifically, we control the context mixture using a consistency ratio $\alpha \in [0, 1]$. For a future downward trend ($\Delta x < 0$), we mix an α proportion of "consistent" (negative sentiment) events with $1 - \alpha$ "inconsistent" (positive sentiment) events. We then condition the LLM on this controlled context to perform the forecasting. A monotonic decrease in forecasting error as α increases empirically validates that the retrieved events genuinely correlate with the time series dynamics, and that our *sentiment* annotations are accurate. Testing $\alpha \in \{0, 0.2, 0.4, 0.6, 0.8, 1.0\}$ reveals a strong correlation between performance and α (Further details are provided in §4.2).

3.3 Domains and Retrieval

Domains. Our living benchmark focuses on event-driven forecasting targets that are inherently volatile and continuously shaped by real-world occurrences (Karger et al., 2025; Chen et al., 2025; Fan et al., 2025; Li et al., 2025). Specifically, as listed in Table 2, we cover **1) Stock Market**, tracking both large-cap (e.g., AAPL, NVDA) and mid-cap equities (e.g., HOOD, NET); and **2) Commodities**, monitoring raw materials influenced by global supply chains and geopolitics (e.g., GC=F, CL=F). Furthermore, we include **3) Electricity** and **4) Meteorological** time series for major cities (e.g., New York, Chicago) to reflect complex physical systems, alongside **5) Event Forecasting** for real-world predictive tasks spanning economics, politics, and technology (e.g., elections, interest rate decisions). For detailed descriptions and examples of each domain, please refer to Appendix D.2 and E.4.

Data Retrieval. LEAF gathers at least one quarter of event data for each domain. Specifically, as detailed in Table 2, we collect target-relevant events daily for daily forecasting domains, and weekly for both monthly and event forecasting tasks. For time series data, we utilize the Yahoo Finance API (Aroussi, 2026) for real-time stock and commodity pricing, the Meteostat API (Meteostat, 2026) for weather observations, the Grid Status API (Kanter & Grid Status, 2026) for electricity load data and the Polymarket API (Polymarket) to track trending predictive events and outcomes. Finally, our retrieval agent is powered by Gemini 3.1 Pro (Google DeepMind, 2026), while the dual fact-checking agents are built upon Gemini 3.1 Pro and Claude 4.6 Sonnet (Anthropic, 2026a).

3.4 Benchmark Integrity and Expert Verification

A core vulnerability of static and automated forecasting benchmarks is evaluation failure driven by either pre-training memorization or search-induced look-ahead information leakage.

Quantifying Pre-Training Contamination. Evaluating models on data predating their training cut-offs (Jan 2023–Jun 2024) reveals severe performance inflation due to pre-training data leakage. Across equity trend tasks, models exhibit an average 147% relative performance gain on historical data compared to out-of-sample living sequences, underscoring that static benchmarks reward memorization over genuine predictive reasoning (detailed breakdowns in Appendix F.2, Table 11).

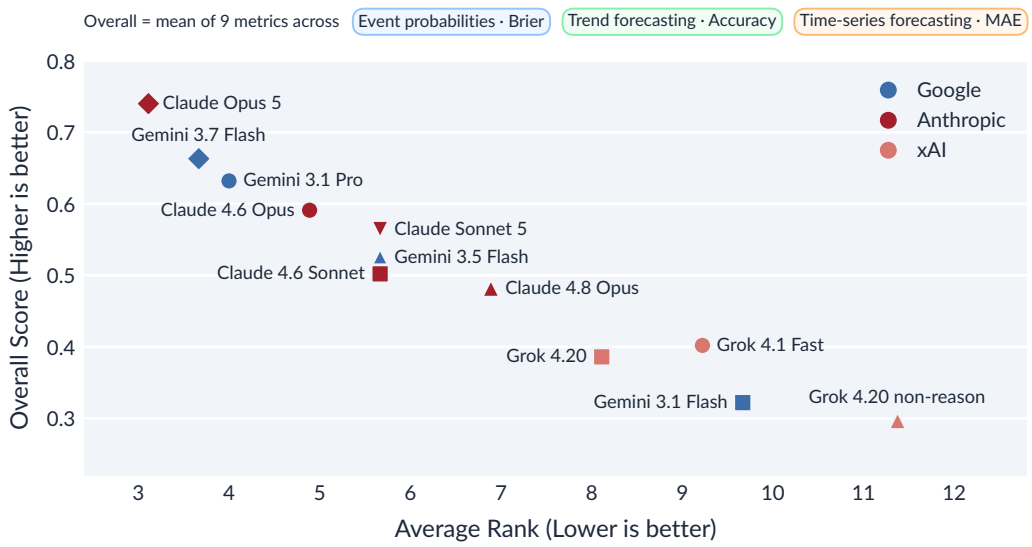

Figure 4: Overall performance of the evaluated LLMs. The x-axis denotes the average rank across tasks, while the y-axis represents the overall score. The two metrics demonstrate a strong correlation.

Mitigating Look-Ahead Information Leakage. Automated search often surfaces retrospective analyses or misdated articles that reveal post-cutoff outcomes, causing severe look-ahead information leakage. To audit benchmark hygiene, we engaged 47 vetted domain specialists (averaging 13.6 years of professional industry experience across finance, meteorology, power systems, and geopolitics) to evaluate 500 tasks across all five domains (100 per domain).

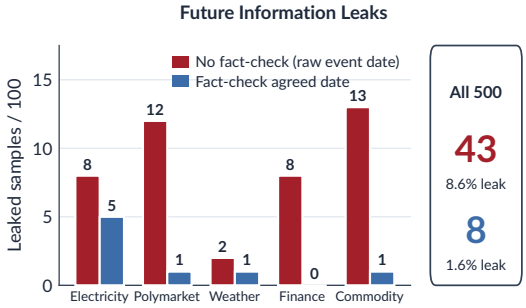

Figure 3: Human-expert-evaluated look-ahead information leakage rates across domains. Without fact-checking, 8.6% of events exhibit leakage; applying the module suppresses this to 1.6%.

Without our dual-agent cross-validation mechanism, automated scrapers matched verified publication dates in only 305/500 tasks (61.0%) and introduced look-ahead information leakage in 43/500 tasks (8.6%). Implementing our cross-validation module reconciled date accuracy to 405/500 tasks (81.0%) and suppressed look-ahead leakage down to just 8/500 tasks (1.6%). Crucially, as shown in Figure 3, leakage was heavily concentrated in the Electricity domain, where 5 of the 8 cases deviated by only a single calendar day, a bounded error with negligible impact over our 3-day and 7-day forecasting horizons.

Factuality and Semantic Grounding. Expert verification confirmed that only 1.6% of cross-validated events contained factual fabrications, with a 0% false negative rate (indicating no valid events were erroneously excluded). In semantic extraction, automated "Sentiment" and "Impact Type" achieved 89.8% and 89.1% expert agreement, respectively. Conversely, "Causality" achieved only 17.3% agreement, as models frequently make unsupported assumptions linking macro news to micro assets (e.g., global wafer expansions affecting DDR5 prices). These semantic attributes are excluded from our inputs, ensuring models are tested purely on their ability to filter noise and infer causal dynamics. Expert demographics, rubrics, and the failure taxonomy are detailed in Appendix C.

4 Experiments and Results

In this section, we evaluate LLMs on our benchmark to investigate three research questions: **RQ1**: What is the overall performance of LLMs on forecasting tasks, and can they leverage events to enhance forecasting performance? (§4.2) **RQ2**: Can LLMs outperform baselines in time series forecasting? (§4.3) **RQ3**: Are market trends predictable, and are events relevant to market forecasting? (§4.4)

Table 3: Accuracy comparison for 3-class trend forecasting across domains (weather, electricity, and long-term mid-cap / short-term large-cap stocks), using only time series (TS) vs. incorporating events (w/ Evts). Green (Red) indicates improvement (degradation) over the TS-only baseline. Complete results (quinary targets, 95% CIs, additional baselines) are in Appendix F.2 (Table 16, 12, 13).

Model	Weather				Electricity			
	Horizon: 3 Days		Horizon: 7 Days		Horizon: 3 Days		Horizon: 7 Days	
	TS	w/ Evts	TS	w/ Evts	TS	w/ Evts	TS	w/ Evts
Gemini 3.1 Pro	49.25	68.30	71.18	77.84	51.76	61.76	62.80	70.73
Gemini 3.7 Flash	51.67	69.81	70.96	78.65	51.76	59.41	71.34	70.73
Claude Opus 4.6	47.22	61.30	59.23	66.92	44.71	60.59	54.88	60.37
Claude Opus 5	45.93	69.26	59.42	69.23	50.59	64.71	68.90	77.44
Grok 4.20	37.78	62.59	36.92	65.96	40.59	51.76	44.51	59.76
Gemma3-27B-it	31.13	52.26	30.59	57.84	36.47	47.06	45.73	59.15
Gemma3-12B-it	35.09	52.45	42.75	63.33	33.53	41.76	43.29	60.98

Model	Mid-Cap Stocks (Long-Term)				Large-Cap Stocks (Short-Term)			
	Horizon: 3 Months		Horizon: 6 Months		Horizon: 1 Day		Horizon: 2 Days	
	TS	w/ Evts	TS	w/ Evts	TS	w/ Evts	TS	w/ Evts
Gemini 3.1 Pro	33.76	35.90	40.17	49.15	38.72	40.85	42.34	44.89
Gemini 3.7 Flash	32.91	35.04	34.19	38.03	37.09	39.01	43.26	46.24
Claude Opus 4.6	31.20	32.05	26.92	28.63	36.03	40.21	41.77	45.53
Claude Opus 5	41.45	42.31	54.70	59.83	39.72	38.94	43.97	45.74
Grok 4.20	14.10	20.51	6.41	26.92	38.79	41.13	39.57	42.27
Gemma3-27B-it	26.07	33.76	20.09	33.33	33.48	38.79	34.61	39.36
Gemma3-12B-it	17.95	31.20	13.25	32.48	33.26	39.72	35.74	37.87

4.1 Experimental Setup

Models and Baseline. We evaluate a total of 16 models, including both the latest proprietary models (e.g., the Claude 4.6 and 5 families, Gemini 3.x series, and Grok 4 variants) and prominent open-weight models (e.g., Qwen3.5-27B/-9B (Qwen Team, 2026), Gemma-3-27B-it/-12B-it (Gemma Team et al., 2025)). We employ two time-series foundation models, TimesFM 3.0 (Das et al., 2024) and Chronos-2 (Ansari et al., 2024), alongside naive time series baselines: the Last Value Baseline (LVB) and Historical Mean Baseline (HMB). Depending on the domain, LVB repeats the most recent observation or 24-hour sequence, while HMB predicts the overall or 14-day rolling average (see Appendix E.1 for details).

Statistics and Evaluation Settings. For time-series forecasting, we use a 14-day look-back window to predict 3- and 7-day horizons. Large-cap stocks are evaluated on daily closing prices. For weather and electricity, we evaluate two distinct targets: hourly values and daily max/min values. Consequently, the test sample counts for the 3- and 7-day horizons are: commodities (340/320), stocks (1,380/1,320), electricity (170×2 / 164×2), and weather (540×2 / 520×2), where the $\times 2$ accounts for these dual targets. For trend forecasting, we add 3- and 6-month horizons for mid-cap stocks using a 3-month look-back (234/234 samples), and 1- and 2-day horizons for large-cap stocks (1,410/1,410 samples) to reflect rapid information pricing. Finally, we allocate an 8,192-token budget per model, discarding any baseline that yields over 33% invalid outputs to ensure fair comparison.

Metrics. For market price forecasting with n timestamps, we evaluate performance using Mean Absolute Percentage Error (MAPE, $\frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i-\hat{y}_i}{y_i}\right|$). For weather forecasting, we employ Mean Absolute Error (MAE) and Mean Squared Error (MSE) calculated on z-score normalized data. For event forecasting, we evaluate predictions at 6 distinct points across the forecasting window, ranging from 100% down to 10% of the time-to-resolution. This window spans from the prediction start time, aligned with Polymarket settings, to the final event resolution. In the evaluation, as mentioned in Section 3.1, we prompt the model for the predicted probability of the actual ground-truth outcome. For evaluation metric, we utilize the average Brier Score of m forecasting cutoffs ($\frac{1}{m}\sum_{t=1}^{m}(f_t-o_t)^2$), where f_t is the predicted probability and o_t is the outcome[3].

We define an **Overall Score** as the unweighted mean of normalized $[0, 1]$ performance metrics across 9 available tasks for each model. This aggregate metric combines scaled accuracy for trend

[3] https://en.wikipedia.org/wiki/Brier_score

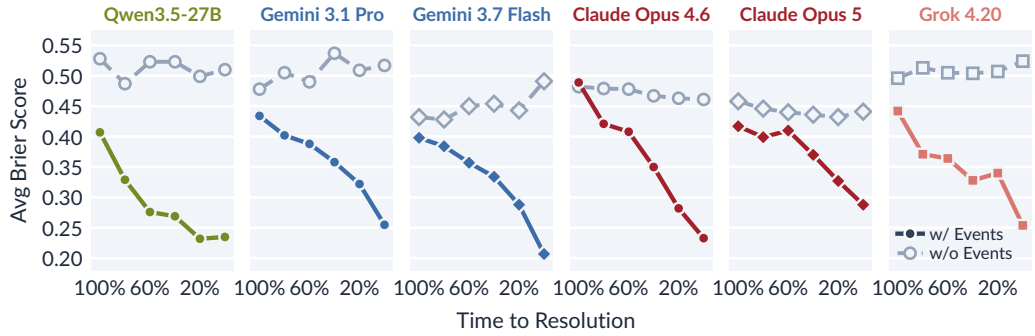

Figure 5: Event forecasting performance across six time-to-resolution cut-offs from 100% to 10% based on the timeframes in Polymarket. The prediction error of context-augmented models consistently decreases as the resolution date approaches. Details can be seen in Table 15 in Appendix F.2.

Table 4: Forecasting error (MAE on z-score normalized data) using time series only (TS) vs. incorporating events (w/ Evts): hourly point forecasting and daily Max/Min forecasting of the same series. The best value per column is underlined; shaded cells mark numerical time-series models that cannot ingest text. Complete MSE metrics and baselines are in Table 17 and 18 in Appendix F.2.

Domain	Model	Point Forecasting (Hourly)				Daily Max/Min Forecasting			
		Horizon: 3 Days		Horizon: 7 Days		Horizon: 3 Days		Horizon: 7 Days	
		TS	w/ Evts	TS	w/ Evts	TS	w/ Evts	TS	w/ Evts
Weather	Gemini 3.1 Pro	0.810	0.780	0.992	0.924	1.036	0.769	1.163	0.899
	Gemini 3.7 Flash	0.917	0.865	1.049	1.006	0.989	0.675	1.119	0.818
	Claude Opus 4.6	0.847	0.735	0.988	0.855	1.006	0.764	1.143	0.894
	Claude Opus 5	0.808	0.639	0.973	0.774	1.016	0.707	1.149	0.819
	Grok 4.20	1.077	1.000	1.170	1.098	1.020	0.827	1.182	0.950
	TimesFM 3.0	0.699	–	0.858	–	1.035	–	1.228	–
	Chronos-2	0.693	–	0.859	–	1.032	–	1.216	–
Electricity	Gemini 3.1 Pro	0.571	0.486	0.695	0.590	1.057	0.815	1.178	0.955
	Gemini 3.7 Flash	0.584	0.563	0.695	0.650	0.957	0.766	1.151	0.932
	Claude Opus 4.6	0.738	0.675	0.842	0.774	1.017	0.831	1.206	1.010
	Claude Opus 5	0.484	0.423	0.575	0.496	0.926	0.753	1.089	0.897
	Grok 4.20	0.836	0.772	0.860	0.842	1.080	1.044	1.226	1.177
	TimesFM 3.0	0.491	–	0.570	–	1.015	–	1.164	–
	Chronos-2	0.474	–	0.563	–	1.014	–	1.158	–

classification, 1 − Brier score for event forecasting, and inverted normalized MAE for time series point forecasting. Full calculation details are provided in Appendix E.2.

4.2 Event and Trend Forecasting with Context

Figure 4 illustrates the overall performance of the evaluated LLMs. The average rank reflects the models' relative standing across diverse tasks, whereas the overall score quantifies the specific performance gaps between them. Generally, the forecasting capability of an LLM correlates with its general-purpose abilities; for instance, Claude 5 Opus outperforms Claude 4.6 Opus, which in turn surpasses Claude 4.6 Sonnet. Since smaller-parameter models still lack the capacity to perform certain tasks, they are excluded from these aggregate results.

In addition, Table 3 shows trend forecasting accuracy for weather, electricity, and stock. Text-augmented proprietary models consistently outperform historical-data-only baselines across all settings, proving the benefit of auxiliary events. For event forecasting (Figure 5), textual context (solid lines) enhances performance over event-free inference (dashed lines) across all models, with prediction errors decreasing near resolution. Detailed prompts and complete results for all 16 models, including 95% confidence intervals, are provided in Appendices D.3 and F.2, confirming the universal benefit of textual context in forecasting. Notably, open-weight models such as Qwen are excluded due to frequent invalid outputs in certain tasks (e.g., weather and electricity trend forecasting), whereas proprietary models yield a 0% failure rate.

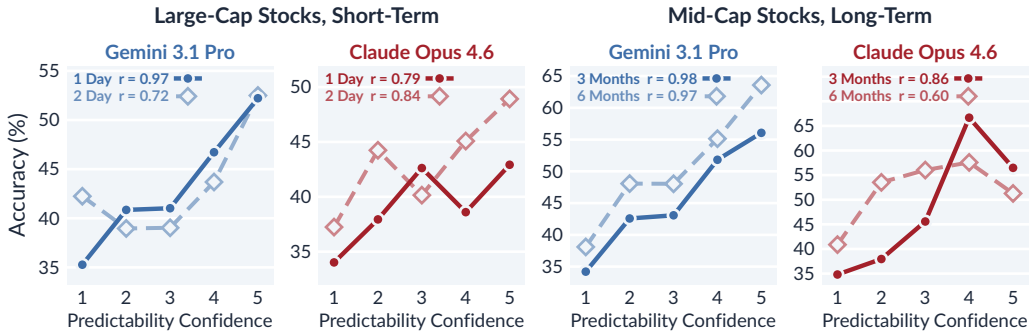

Figure 6: Forecasting performance of large-cap and mid-cap stocks across varying levels of predictability confidence of Gemini-3.1 Pro (Gemini) and Claude Opus 4.6 (Claude). Higher forecasting accuracy is achieved on stocks assigned higher predictability confidence by the LLMs.

4.3 Time Series Forecasting with Context

Given that time series forecasting is notoriously difficult for LLMs (Merrill et al., 2024; Tan et al., 2024), our evaluation primarily focuses on advanced proprietary models. As Table 4 illustrates, while these models effectively leverage event data to enhance forecasting performance, most still perform only on par with specialized time series foundation models like TimesFM 3.0 (Das et al., 2024), with the exception of Claude Opus 5. Furthermore, while market price prediction is frequently included in time series forecasting tasks (Chen et al., 2025; Wang et al., 2024a; Dong et al., 2024), our results (Table 9 in Appendix F.2) reveal that a naive Baseline Last Value (BLV) outperforms all other methods. Given the Efficient Market Hypothesis (EMH), we conclude that this task is inherently infeasible. Detailed explanations are provided in Appendix F.1.

4.4 Identifying Forecastable Signals and Events Validations

We use stock trend forecasting as a benchmark task, hypothesizing that strong, unpriced signals persist (§3.2). We prompt LLMs to classify the daily context (events and historical prices) of 40 mid-cap or 30 large-cap stocks into five predictability confidence levels, and aggregate the daily accuracy for each level. **We observe a strong correlation between accuracy and confidence**; for example, Gemini 3.1 Pro achieves correlations of 0.97 and 0.98 in short- and long-term stock price forecasting, respectively. The main results are presented in Figure 6, with additional details and ablation studies provided in Table 14 (Appendix F.2). To further validate whether our events correlate with prediction targets, we introduce "sentiment consistent" text (Chen et al., 2025) to aid predictions (§3.2). By controlling the proportion (α) of this text, Figure 8 (Appendix F.2) shows that forecasting performance correlates strongly with α when target-relevant text is used. Specifically, evaluating 15 out of 30 instances in the market domain, target-relevant text yields a correlation of -0.967, whereas using text from unrelated companies yields only -0.537.

5 Conclusions and Future Works

In summary, we introduce LEAF, a novel living benchmark for event-augmented LLM forecasting. To construct this benchmark, we propose a multi-agent retrieval and cross-validation system. Our human expert evaluation confirms that this system guarantees authentic and temporally aligned events, successfully suppressing future information leakage to just 1.6%, with over 60% of these remaining discrepancies bounded by a single day. Our evaluation of 16 LLMs demonstrates that incorporating these events consistently enhances forecasting performance across diverse domains. Additionally, we find that market trends exhibit predictability and correlate well with external events. Furthermore, we reveal that evaluating on outdated data causes severe performance inflation (e.g., a 147% relative improvement in market forecasting), underscoring the critical necessity of live testing.

Future Work. We aim to expand the benchmark with a broader spectrum of event-driven forecasting targets and train agents or predictive models within its dynamic environment.

AI USE STATEMENT

In this work, we used generative AI tools for the following tasks with required disclosure:

- **Data Retrieval and Processing Pipeline:** We employed LLMs as core operational agents within our benchmark construction pipeline. Specifically, an LLM agent powered by Gemini 3.1 Pro (augmented with search grounding) was used to recursively gather candidate event summaries, timestamps, and attribution descriptions.
- **Automated Fact-Checking and Temporal Alignment:** We implemented a dual-agent consensus mechanism using Gemini 3.1 Pro and Claude 4.6 Sonnet to verify the factual authenticity and anchor the first-public realization date of retrieved events prior to human expert auditing.
- **Benchmarked Subjects:** We evaluated 16 frontier proprietary and open-weight LLMs as experimental subjects across event probability, market trend, and time-series forecasting tasks.

Generative AI tools were used to assist with grammatical refinement, and readability improvements of the written manuscript. We have *not* used generative AI tools to conceptualize the benchmark taxonomy, formulate mathematical definitions or metric normalization frameworks, conduct human-expert audits (which were performed independently by 47 professional domain specialists), write non-agent evaluation code, or interpret empirical findings. The authors have thoroughly reviewed all AI-generated prompts, retrieval traces, and textual revisions, and take full responsibility for the contents of this work.

REPRODUCIBILITY STATEMENT

To ensure complete reproducibility, we provide the complete codebase, evaluation scripts, and dataset artifacts via an anonymous public repository and within the supplementary material. Method and dataset details are provided in §3. Details on full experimental setup including hyperparameters, hardware, model variants, prompts used, etc. are provided in §4, App. D, and App. E. Complete demographics, hourly compensation rates, rubrics, and failure taxonomy for the 47-specialist expert validation panel are reported in §3.4 and App. C.

REFERENCES

Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, Jasper Zschiegner, Danielle C. Maddix, Hao Wang, Michael W. Mahoney, Kari Torkkola, Andrew Gordon Wilson, Michael Bohlke-Schneider, and Yuyang Wang. Chronos: Learning the language of time series. *arXiv preprint arXiv:2403.07815*, 2024.

Anthropic. Claude sonnet 4.5 system card. Technical report, Anthropic, 2026a. URL https://assets.anthropic.com/m/12f214efcc2f457a/original/Claude-Sonnet-4-5-System-Card.pdf.

Anthropic. Claude opus 4.6 system card. Technical report, Anthropic, 2 2026b. URL https://www-cdn.anthropic.com/c788cbc0a3da9135112f97cdf6dcd06f2c16cee2.pdf.

Ran Aroussi. yfinance: Yahoo! finance market data downloader. https://github.com/ranaroussi/yfinance, 2026. Accessed: 2026-03-11.

Yonatan Bitton, Hritik Bansal, Jack Hessel, Rulin Shao, Wanrong Zhu, Anas Awadalla, Josh Gardner, Rohan Taori, and Ludwig Schimdt. Visit-bench: A benchmark for vision-language instruction following inspired by real-world use, 2023.

Jialin Chen, Aosong Feng, Ziyu Zhao, Juan Garza, Gaukhar Nurbek, Cheng Qin, Ali Maatouk, Leandros Tassiulas, Yifeng Gao, and Rex Ying. Mtbench: A multimodal time series benchmark for temporal reasoning and question answering. *arXiv preprint arXiv:2503.16858*, 2025.

Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting. In *Forty-first International Conference on Machine Learning*, 2024.

Zihan Dong, Xinyu Fan, and Zhiyuan Peng. Fnspid: A comprehensive financial news dataset in time series, 2024.

Eugene F Fama. The behavior of stock-market prices. *The journal of business*, 38(1):34–105, 1965.

Eugene F Fama. Efficient capital markets: A review of theory and empirical work. *The Journal of Finance*, 25(2):383–417, 1970.

Lizhou Fan, Wenyue Hua, Lingyao Li, Haoyang Ling, and Yongfeng Zhang. Nphardeval: Dynamic benchmark on reasoning ability of large language models via complexity classes. *arXiv preprint arXiv:2312.14890*, 2023.

Tianyu Fan, Yuhao Yang, Yangqin Jiang, Yifei Zhang, Yuxuan Chen, and Chao Huang. Ai-trader: Benchmarking autonomous agents in real-time financial markets. *arXiv preprint arXiv:2512.10971*, 2025.

Gemma Team, Aishwarya Kamath, Johan Ferret, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

Elliot Glazer, Ege Erdil, Tamay Besiroglu, Diego Chicharro, Evan Chen, Alex Gunning, Caroline Falkman Olsson, Jean-Stanislas Denain, Anson Ho, Emily de Oliveira Santos, et al. Frontiermath: A benchmark for evaluating advanced mathematical reasoning in ai. *arXiv preprint arXiv:2411.04872*, 2024.

Google DeepMind. Gemini 3.1 pro. `https://deepmind.google/models/gemini/pro/`, 2026. Accessed: 2026-03-11.

Nate Gruver, Marc Finzi, Shikai Qiu, and Andrew Gordon Wilson. Large Language Models Are Zero Shot Time Series Forecasters. In *Advances in Neural Information Processing Systems*, 2023.

Danny Halawi, Fred Zhang, Chen Yueh-Han, and Jacob Steinhardt. Approaching human-level forecasting with language models. *Advances in Neural Information Processing Systems*, 37: 50426–50468, 2024.

Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. Livecodebench: Holistic and contamination free evaluation of large language models for code. *arXiv preprint arXiv:2403.07974*, 2024.

Ming Jin, Shiyu Wang, Lintao Ma, Zhixuan Chu, James Y Zhang, Xiaoming Shi, Pin-Yu Chen, Yuxuan Liang, Yuan-Fang Li, Shirui Pan, et al. Time-llm: Time series forecasting by reprogramming large language models. In *ICLR*, 2024.

Jina AI. Jina Reader and Search API: Extract and search the web for LLMs. `https://jina.ai/`, 2026. Accessed: 2026-03-15.

Max Kanter and Grid Status. gridstatus: Extract data from ISOs and other energy grid sources. `https://github.com/gridstatus/gridstatus`, 2026. Accessed: 2026-04-28.

Ezra Karger, Houtan Bastani, Chen Yueh-Han, Zachary Jacobs, Danny Halawi, Fred Zhang, and Philip E. Tetlock. Forecastbench: A dynamic benchmark of ai forecasting capabilities. In *International Conference on Learning Representations (ICLR)*, 2025. URL `https://iclr.cc/virtual/2025/poster/28507`.

Changlun Li, Yao Shi, Chen Wang, Qiqi Duan, Runke Ruan, Weijie Huang, Haonan Long, Lijun Huang, Nan Tang, and Yuyu Luo. Time travel is cheating: Going live with deepfund for real-time fund investment benchmarking. *arXiv preprint arXiv:2505.11065*, 2025.

Yuan Li, Bingqiao Luo, Qian Wang, Nuo Chen, Xu Liu, and Bingsheng He. Cryptotrade: A reflective llm-based agent to guide zero-shot cryptocurrency trading. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 1094–1106, 2024.

Haoxin Liu, Shangqing Xu, Zhiyuan Zhao, Lingkai Kong, Harshavardhan Kamarthi, Aditya B Sasanur, Megha Sharma, Jiaming Cui, Qingsong Wen, Chao Zhang, et al. Time-mmd: Multi-domain multimodal dataset for time series analysis. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024.

Yucong Luo, Yitong Zhou, Mingyue Cheng, Jiahao Wang, Daoyu Wang, Tingyue Pan, and Jintao Zhang. Time series forecasting as reasoning: A slow-thinking approach with reinforced llms. *arXiv preprint arXiv:2506.10630*, 2025.

Mike A Merrill, Mingtian Tan, Vinayak Gupta, Thomas Hartvigsen, and Tim Althoff. Language models still struggle to zero-shot reason about time series. In *Findings of EMNLP*, 2024.

Meteostat. Meteostat Python: Historical weather data. `https://github.com/meteostat/meteostat-python`, 2026. Accessed: 2026-03-11.

Nof1.ai. Alpha arena leaderboard. `https://nof1.ai/leaderboard`, 2026. Accessed: 2026-03-11.

OpenAI. Introducing gpt-5.4. `https://openai.com/index/introducing-gpt-5-4/`, 3 2026. Accessed: 2026-03-11.

Polymarket. Polymarket api documentation. `https://docs.polymarket.com/`. Accessed: 2026-03-11.

Qwen Team. Qwen3.5: Towards native multimodal agents. `https://qwen.ai/blog?id=qwen3.5`, February 2026.

Mingtian Tan, Mike A Merrill, Vinayak Gupta, Tim Althoff, and Thomas Hartvigsen. Are language models actually useful for time series forecasting? In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Mingtian Tan, Mike A Merrill, Zack Gottesman, Tim Althoff, David Evans, and Tom Hartvigsen. Inferring events from time series using language models. *arXiv preprint arXiv:2503.14190*, 2025.

Xinlei Wang, Maike Feng, Jing Qiu, Jinjin Gu, and Junhua Zhao. From news to forecast: Integrating event analysis in llm-based time series forecasting with reflection. *arXiv preprint arXiv:2409.17515*, 2024a.

Xinlei Wang, Maike Feng, Jing Qiu, Jinjin Gu, and Junhua Zhao. From news to forecast: Integrating event analysis in llm-based time series forecasting with reflection. In *Neural Information Processing Systems*, 2024b.

Colin White, Samuel Dooley, Manley Roberts, Arka Pal, Benjamin Feuer, Siddhartha Jain, Ravid Shwartz-Ziv, Neel Jain, Khalid Saifullah, Sreemanti Dey, Shubh-Agrawal, Sandeep Singh Sandha, Siddartha Venkat Naidu, Chinmay Hegde, Yann LeCun, Tom Goldstein, Willie Neiswanger, and Micah Goldblum. Livebench: A challenging, contamination-free LLM benchmark. In *The Thirteenth International Conference on Learning Representations*, 2025.

Andrew Robert Williams, Arjun Ashok, Étienne Marcotte, Valentina Zantedeschi, Jithendaraa Subramanian, Roland Riachi, James Requeima, Alexandre Lacoste, Irina Rish, Nicolas Chapados, and Alexandre Drouin. Context is key: A benchmark for forecasting with essential textual information. In *Forty-second International Conference on Machine Learning*, 2025. URL `https://openreview.net/forum?id=ih2WuBT1Fn`.

Yijia Xiao, Edward Sun, Di Luo, and Wei Wang. Tradingagents: Multi-agents llm financial trading framework, 2025. URL `https://arxiv.org/abs/2412.20138`.

Zhijian Xu, Wanxu Cai, Xilin Dai, Zhaorong Deng, and Qiang Xu. Fidel-ts: A high-fidelity benchmark for multimodal time series forecasting. *arXiv preprint arXiv:2509.24789*, 2025.

Qingchuan Yang, Simon Mahns, Sida Li, Anri Gu, Jibang Wu, and Haifeng Xu. Llm-as-a-prophet: Understanding predictive intelligence with prophet arena. In *International Conference on Learning Representations (ICLR)*, 2026.

Yue Yang, Shuibai Zhang, Wenqi Shao, Kaipeng Zhang, Yi Bin, Yu Wang, and Ping Luo. Dynamic multimodal evaluation with flexible complexity by vision-language bootstrapping. *arXiv preprint arXiv:2410.08695*, 2024.

Fangxu Yu, Xingang Guo, Lingzhi Yuan, Haoqiang Kang, Hongyu Zhao, Lianhui Qin, Furong Huang, Bin Hu, and Tianyi Zhou. Tsrbench: A comprehensive multi-task multi-modal time series reasoning benchmark for generalist models. *arXiv preprint arXiv:2601.18744*, 2026.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Tianle Li, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zhuohan Li, Zi Lin, Eric P Xing, et al. Lmsys-chat-1m: A large-scale real-world llm conversation dataset, 2024. *URL https://arxiv. org/abs/2309.11998*, 1, 2024.

Zihan Zheng, Zerui Cheng, Zeyu Shen, Shang Zhou, Kaiyuan Liu, Hansen He, Dongruixuan Li, Stanley Wei, Hangyi Hao, Jianzhu Yao, et al. Livecodebench pro: How do olympiad medalists judge llms in competitive programming? *arXiv preprint arXiv:2506.11928*, 2025.

Tian Zhou, Peisong Niu, Liang Sun, Rong Jin, et al. One fits all: Power general time series analysis by pretrained lm. In *NeurIPS*, 2023.

Chengke Zou, Xingang Guo, Rui Yang, Junyu Zhang, Bin Hu, and Huan Zhang. Dynamath: A dynamic visual benchmark for evaluating mathematical reasoning robustness of vision language models, 2024. URL https://arxiv.org/abs/2411.00836.

Table of Contents for Appendix

A Limitations **15**

B Broader Impacts. **15**

C Human Expert Data Annotation. **15**

C.1 Expert Pool Demographics 15

C.2 Temporal Alignment and Leakage Mitigation. 15

C.3 Event Authenticity and Semantic Attributes 16

C.4 Fact-Check Module Case Study. 16

C.5 Evaluation of the Fact-Check Agent 16

D Methodology Supplementary **18**

D.1 Trend Boundary 18

D.2 Forecasting Domains. 18

D.3 Prompts 20

E Experiment Supplementary **21**

E.1 Models and Computation Resources. 21

E.2 Overall Score. 21

E.3 Case Studies 22

E.4 Dataset Taxonomy 22

F Additional Experiment Results. **23**

F.1 Forecast Accurate Market Price is Infeasible 23

F.2 Detailed Results 23

A LIMITATIONS

Here, we discuss primary limitations of our paper. First, given our living benchmark requires retrieval agents to dynamically collect auxiliary events for LLM evaluation. The associated API costs constrain our domain coverage, preventing an exhaustive assessment across all forecasting fields. Second, our event forecasting setup does not incorporate historical multimodal data (e.g., time-series values). Due to data collection constraints, we rely solely on textual events as auxiliary inputs. Evaluations of other forecasting tasks in the paper incorporate both textual events and time-series values.

B BROADER IMPACTS

Our paper introduces a living benchmark to evaluate LLM forecasting capabilities across complex domains like market trends, weather, and electricity loads. By utilizing dynamically updated test data, this work addresses pre-training data contamination in LLM evaluation. Furthermore, enhancing forecasting in physical domains can positively impact societal infrastructure, such as optimizing energy grids and environmental preparedness. However, exploring financial market predictions carries the risk of unintended misuse, as users might over-trust models for algorithmic trading or personal decisions, risking financial loss. To mitigate this, we explicitly state that our research does not constitute financial advice or endorse commercial deployment, remaining strictly intended for academic evaluation.

C HUMAN EXPERT DATA ANNOTATION

To ensure the high quality and factual authenticity of our benchmark, we conducted an exhaustive human-expert audit across 500 randomly sampled event-driven forecasting instances (100 per domain across Finance, Meteorology, Energy/Electricity, Geopolitics, and Hardware/Commodities).

C.1 EXPERT POOL DEMOGRAPHICS

The annotation pool comprised 47 veteran specialists across seven operational domains[4]. These specialists possess extensive professional backgrounds, with domain experience ranging from 2 to 35 years (averaging 13.6 years, with a median of 10 years; 30 specialists have $\geq$10 years experience). To ensure fair compensation commensurate with domain expertise, annotators were paid an average rate of $140 per hour. The vast majority (68%) hold advanced postgraduate degrees (8 Doctorates, 24 Master's degrees) and serve in senior roles such as Chief Meteorologists, Quantitative Modellers, Institutional Credit Underwriters, Energy Systems Data Scientists, and Geopolitical Risk Consultants:

- **Finance & Markets (15 specialists):** CFAs, institutional credit underwriters, quantitative modellers, and long/short equity research associates (avg. 12.5 years of experience).
- **Weather & Meteorology (10 specialists):** Operational forecasters, chief meteorologists, Doppler radar analysts, and numerical weather prediction (NWP) researchers (avg. 18.2 years of experience).
- **Energy & Electricity (8 specialists):** Power systems engineers, grid contingency and transmission modelers, and commercial power pricing analysts (avg. 9.3 years of experience).
- **Geopolitics (4 specialists):** International political economists, trade sanction and export control consultants, and open-source intelligence (OSINT) researchers (avg. 11.8 years of experience).
- **Infrastructure, Sales & Mobility (10 specialists):** Cloud observability engineers, transport demand modelers, and enterprise S&OP planning directors (avg. 10.7 years of experience).

C.2 TEMPORAL ALIGNMENT AND LEAKAGE MITIGATION

As outlined in Section 3.2, expert evaluation reveals a major discrepancy between raw automated news retrieval and actual event occurrence dates. As illustrated in Figure 3, raw retrieval dates suffer from look-ahead information leakage in 43 out of 500 cases (8.6%), with only 305 out of 500 dates matching the true first-public release. After applying our dual-agent fact-checking module, verified

[4]Annotation guidelines are provided in `https://anonymous.4open.science/r/LEAF-4565`

publication dates match ground truth in 405 out of 500 tasks (81.0%), while future information leakage drops to just 8 out of 500 cases (1.6%).

The exact profiles of these 8 residual leakage instances are documented in Table 5. Notably, 5 of these 8 cases originate in the Electricity domain, and 5 exhibit a temporal slippage of only a single calendar day (e.g., an outlook filed on day T quoting a day $T+1$ morning schedule)—representing a bounded discrepancy with minimal impact given our 3-day and 7-day forecast horizons.

C.3 Event Authenticity and Semantic Attributes

Regarding factual authenticity, expert inspection confirmed that only 8 events (1.6%) were completely fabricated or directly contradicted cited primary sources after passing through the dual fact-checking filters. As detailed in Table 6, these residual errors were predominantly concentrated in the Weather domain, where fine-grained localized forecasts are often poorly archived or superseded online.

Finally, as shown in Table 7, experts audited extracted semantic labels ("*causality*", "*sentiment*", and "*impact_type*"). While models accurately captured "*sentiment*" (89.5% acceptance) and "*impact_type*" (89.1% acceptance), the "*causality*" labels achieved only 17.3% expert acceptance (52/300 eligible tasks). Reviewers noted that models routinely make unsupported causal leaps (70.8% of tasks), such as assuming broad industry investments immediately dictate micro-asset price movements. Consequently, **these noisy causality labels were excluded from downstream model inputs**, ensuring LLMs must independently deduce event impacts.

C.4 Fact-Check Module Case Study

To explicitly illustrate the consequences of omitting the fact-checking module, Figure 7 presents an example of severe performance inflation in Gemini-3.1-Pro caused by look-ahead information leakage. In the event forecasting task "the first leader to depart by 2025-12-31" (cut-off date: 2025-06-17), an unfiltered search query surfaced a leaked retrospective article (dated 2025-09-07), artificially jumping the predicted probability from 0% to 92%. In contrast, the fact-check agent identifies the chronological mismatch and suppresses the leaked record, preserving true out-of-sample evaluation hygiene.

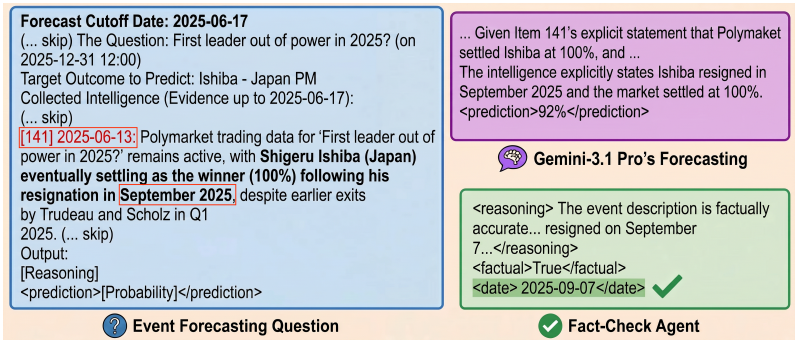

Figure 7: Example of event forecasting without Fact-Check Agents. Event #141 leaks future information (published in 09/2025), inflating the 2025-06-13 prediction to 92%, compared to 0% when the event is filtered. The bottom-right panel shows the Fact-Check Agent's filtering analysis triggered by this chronological discrepancy.

C.5 Evaluation of the Fact-Check Agent

We define "false negative cases" for the fact-check agent as invalid instances that bypass the filter, potentially allowing the LLM to gain an unfair advantage during prediction and inflate performance. Based on human expert evaluation of 500 examples, this rate is 3.2% in total (1.6% future information leakage and 1.6% fabricated events). Conversely, we define "false positive cases" as authentic events collected by Gemini-3.1-Pro with correct raw dates that are mistakenly discarded by the fact-check agent. Human experts manually inspected 40 randomly selected events filtered out by the mechanism and found no such instances, resulting in a false positive rate of 0%.

Table 5: Future-information leaks relative to the fact-check-agreed `date_val`: 8 of 500 expert-annotated samples (1.6%). A sample is listed when the expert's annotation states that a specific part of the event description was first published after `date_val`. "Gap" is the number of days between `date_val` and the first public release of the leaked content.

#	Task	date_val	Leaked element	Leaked content in description	First public	Gap
1	Electricity_1_San_Francisco	2026-01-21	quotes the Ascend Analytics weekly market outlook, posted the next day	"...(Source: CAISO / Ascend Analytics Market Outlook) ... solar output projected to strictly eclipse 10 GW beginning on Friday, January 23 ..."	2026-01-22	1 d
2	Electricity_2_New_York	2026-01-22	attendance figure taken from the post-event press release	"... gathered over 1,100 real estate, government, and civic leaders ..."	2026-01-23	1 d
3	Electricity_3_San_Francisco	2026-03-17	observed midnight temperature, first published in the 02:00 NWS discussion of the next day	"... downtown San Francisco recording 70°F around midnight ..."	2026-03-18	1 d
4	Electricity_4_New_York	2025-12-09	content of an Amperon white paper whose earliest verifiable publication is 14 days later	"... Amperon published a NYISO-specific reliability report ... highlights shrinking reserve margins and winter fuel-security constraints ..."	2025-12-23	14 d
5	Electricity_5_San_Francisco	2025-12-23	forecast wording that first appears in the next day's NWS discussion	"... Saturated soils and stressed trees ... are forecast to cause compounding tree failures ... on December 25-26 ..."	2025-12-24	1 d
6	Weather_1_Washington_DC	2025-12-04	sounding / freezing-drizzle sentence from the advisory issued the next day	"... Forecast soundings indicate a transition to patchy freezing drizzle by the afternoon ..."	2025-12-05	1 d
7	Event Forecasting_1	2025-12-06	incident first reported in IAEA Update 332, five days after the event	"... ZNPP lost all off-site power for 30 minutes, marking the 11th such blackout ..."	2025-12-11	5 d
8	Commodity_1	2025-07-15	export licences asserted as granted; they were issued 24 days later	"The U.S. government approved the sale of NVIDIA's H20 GPUs and AMD's MI308 AI chips to China ..."	2025-08-08	24 d

Table 6: Expert-identified hallucinations where the core event was completely fabricated or the central claim directly contradicted the source material (8 of 500 samples, 1.6%; 7 in the Weather domain). The dual fact-check agents failed to verify these 8 cases as factual. Both text columns provide verbatim excerpts from the raw dataset; [...] denotes elided text.

#	Task	Date	Description (verbatim excerpt)	Expert reviewer (verbatim excerpt)
1	Weather_1_New_York	2025-12-23	"Official National Weather Service (NWS) winter storm warnings from December 23 establish a cold thermal baseline [...] recording an exact overnight low temperature forecast of -5°C (23°F)."	"This is a completely fabricated event. There were no winter storm warnings issued on December 23 [...] temperatures were actually warming [...] Thus, the 23 degree F temperature in the explanation is wrong."
2	Weather_2_Denver	2025-11-10	"NWS Boulder issued a Red Flag Warning for the Interstate 25 urban corridor [...] valid for Tuesday, November 11."	"This is a made up event by the model. The NWS Boulder office did not issue any Red Flag warnings [...] valid for Tuesday, November 11th."
3	Weather_3_Denver	2025-12-13	"The December 17 cold front will lack deep moisture [...] Northern Colorado will be positioned in the left-exit region of the upper-level jet [...] producing localized snow squalls"	"This is a completely fabricated event [...] There is zero mention of a cold front on December 17th [...] there is no mention of northern Colorado being in the left exit region of the jet".
4	Weather_4_Seattle	2025-11-21	"[...] The thermal outlook calculates a 33% to 50% probability of above-normal temperatures across the Seattle region [...] The precipitation outlook concurrently projects a 33% to 50% probability of above-normal precipitation"	"This event was far from authentic. [...] the temperature outlook [...] for the Seattle area was below average temperatures and the precipitation was below avearge. This is actually opposite of what was stated in the event."
5	Weather_5_New_York	2026-01-11	"[...] Published on January 11, 2026, the long-range hazards assessment highlights a 'Slight Risk' contour [...] for heavy snow accumulation across the Northeast region, including New York [...]"	"[...] the CPC hazard map from the 11th does not include the slight risk of heavy snowfall to New York [...] Eventually this event happened but not through the process described in this event."
6	Weather_6_San_Francisco	2025-10-16	"For the 6-10 day period [...] the CPC forecasts temperatures to average near climatological normals for the West Coast [...] The 8-14 day temperature outlook [...] indicates a slight probability for above-average thermal anomalies across California."	"[...] the temperature forecasts are incorrect. The 6-10 day outlook issed that day has temperatures mostly above normal [...] The 8-14 day was near average. It's like the event and models flipped the forecasts [...]"
7	Weather_7_Boston	2025-12-23	"[...] 8-14 day extended-range forecast (ERF) indicating a macro-scale thermal anomaly characterized by strongly above-normal temperatures building across the eastern United States [...] as a larger cyclonic system approaches the Northeast"	"The 8-14 day outlook [...] was actually near to below normal right along the East Coast. [...] There also did not end up being a large cylonic system along the east coast [...]"
8	Electricity_1_Boston	2026-02-19	"ISO New England (ISO-NE) published its Participant Readiness Project Outlook, detailing the upcoming implementation of the Inventoried Energy Program (IEP) and new Day-Ahead Ancillary Services (DASC) reserve requirements [...]"	"This event has been mixed up with the previous year's news. [...] The Inventoried Energy Program [...] was not extended into 2026, so there was no upcoming IEP implementation to report. [...] the record contains no citation with ID 48."

Table 7: Expert acceptance of generated event attributes by domain. Each value denotes the share of samples whose attribute was accepted by expert reviewers, computed over eligible samples carrying the attribute; "–" marks domains whose schema does not evaluate that field.

Attribute	RAM / Comm.	Electricity	Geopolitical	Meteorology	Finance	Overall Pass
Evaluated Tasks	100	100	100	100	100	500
Date Validity Pass	83.0%	46.0%	91.0%	92.0%	93.0%	81.0% (405/500)
Sentiment	84.0%	–	–	–	95.9%	89.5% (179/200)
Impact Type	87.0%	–	–	–	92.8%	89.1% (179/201)
Causality	13.0%	–	23.0%	–	14.4%	17.3% (52/300)

D Methodology Supplementary

D.1 Trend Boundary

To define the trend boundaries for stocks and commodities, we follow an approach similar to (Chen et al., 2025) by categorizing based on the Rate of Change (ROC), calculated as ROC $= \frac{x_{T+N}-x_T}{x_T}$, where x_T denotes the price on the last observed day T, and x_{T+N} represents the price on the N-th day in the future. We adopt the boundary thresholds detailed in Table 8 to ensure a more balanced class distribution. Due to the long forecasting horizon and elevated volatility of mid-cap stocks, their trend boundaries are widened relative to short-term predictions. Specifically, the neutral and strong boundaries are expanded from $\pm 0.5\%$ to $\pm 5\%$ and from $\pm 5\%$ to $\pm 15\%$, respectively.

For Weather trend classification, we avoid the rate-of-change method (Chen et al., 2025) since our dataset includes winter temperatures near 0 °C, where percentage changes become ill-defined. Instead, we use absolute temperature differences based on human perception. Defining the difference between the minimum temperature over the future N days (T_{target}) and the previous day's minimum temperature ($T_{reference}$) as $\Delta T = T_{target} - T_{reference}$ (measured in °C), we categorize the labels as follows: Strong Up ($\Delta T > 3$), Up ($1.5 < \Delta T \leq 3$), Neutral ($-1.5 \leq \Delta T \leq 1.5$), Down ($-3 \leq \Delta T < -1.5$), and Strong Down ($\Delta T < -3$).

For electricity load forecasting, we evaluate the absolute load change ΔL (MW) by categorizing the labels based on their observed distribution, ensuring symmetric thresholds while maintaining a relatively balanced class distribution: Strong Up ($\Delta L > 1000$, 36.70%), Up ($400 < \Delta L \leq 1000$, 17.32%), Neutral ($-400 \leq \Delta L \leq 400$, 25.98%), Down ($-1000 \leq \Delta L < -400$, 13.40%), and Strong Down ($\Delta L < -1000$, 6.60%). Notably, ΔL is computed based on daily peak electricity loads, as accurate peak demand forecasting is essential for optimizing grid stability, managing energy resources efficiently, and preventing system overloads.

D.2 Forecasting Domains

This section provides a comprehensive breakdown of the specific targets continuously monitored within our living benchmark, categorized across four distinct domains.

D.2.1 Stock Market

We track 30 **large-cap equities** representing seven critical sectors to capture diverse market dynamics, corporate announcements, and macroeconomic shifts. The selected tickers are as follows:

- **Tech Giants (Magnificent Seven):** Meta (META), Apple (AAPL), NVIDIA (NVDA), Microsoft (MSFT), Alphabet (GOOGL), Amazon (AMZN), Tesla (TSLA).
- **Semiconductors:** AMD (AMD), Intel (INTC), TSMC (TSM), Broadcom (AVGO), Qualcomm (QCOM).
- **Software & Media:** Netflix (NFLX), Disney (DIS), Adobe (ADBE), Salesforce (CRM), Oracle (ORCL).
- **Finance & Payments:** JPMorgan Chase (JPM), Bank of America (BAC), Visa (V), Mastercard (MA).

Table 8: Comparison of trend label distribution and variance between our boundary and MT-Bench Chen et al. (2025).

Benchmark	Trend Label	Boundary	Large-Cap Stock	Commodities
			Proportion (%)	Proportion (%)
Ours	Strong Up	$> 2\%$	23.94	27.69
	Up	$0.5\% \sim 2\%$	16.23	16.93
	Neutral	$-0.5\% \sim 0.5\%$	18.05	15.74
	Down	$-2\% \sim -0.5\%$	20.59	13.15
	Strong Down	$< -2\%$	21.20	26.49
	Variance	-	*7.06*	*35.17*
MTBench	Strong Up	$> 4\%$	8.62	14.71
	Up	$2\% \sim 4\%$	15.31	15.59
	Neutral	$-2\% \sim 2\%$	54.87	48.53
	Down	$-4\% \sim -2\%$	11.97	13.82
	Strong Down	$< -4\%$	9.23	7.35
	Variance	-	*309.55*	*211.90*

- **Consumer & Retail:** Walmart (WMT), Costco (COST), Coca-Cola (KO), PepsiCo (PEP), Procter & Gamble (PG).
- **Healthcare:** Johnson & Johnson (JNJ), Pfizer (PFE), Eli Lilly (LLY), UnitedHealth Group (UNH).

We extend our benchmark to include 40 mid-cap equities, characterized by higher volatility and concentration in emerging technology sectors. The tickers are categorized as follows:

- **Cloud, AI & Data Analytics:** Snowflake (SNOW), MongoDB (MDB), Datadog (DDOG), Palantir (PLTR), Super Micro Computer (SMCI), Vertiv (VRT).
- **Software, Security & Interactive Media:** Cloudflare (NET), Zscaler (ZS), SentinelOne (S), AppLovin (APP), Duolingo (DUOL), Roblox (RBLX), Unity (U), Reddit (RDDT), UiPath (PATH).
- **Semiconductors & Industrial Automation:** Marvell (MRVL), Astera Labs (ALAB), Coherent (COHR), Symbotic (SYM).
- **Fintech & Digital Assets:** Robinhood (HOOD), SoFi (SOFI), Affirm (AFRM), Coinbase (COIN), MicroStrategy (MSTR), Toast (TOST).
- **Consumer, E-commerce & Lifestyle:** Carvana (CVNA), Coupang (CPNG), Sea Limited (SE), e.l.f. Beauty (ELF), On Holding (ONON), CAVA Group (CAVA), Celsius (CELH), DraftKings (DKNG).
- **Frontier Tech, Biotech & Energy:** AST SpaceMobile (ASTS), Rocket Lab (RKLB), Joby Aviation (JOBY), Rivian (RIVN), CRISPR Therapeutics (CRSP), Viking Therapeutics (VKTX), Constellation Energy (CEG).

D.2.2 COMMODITIES

We dynamically monitor 10 major commodities, encompassing energy, metals, and agricultural products, to reflect vulnerabilities to global supply chains and geopolitical events:

- **Energy:** Crude Oil WTI (CL=F), Brent Crude (BZ=F), Natural Gas (NG=F).
- **Metals:** Gold (GC=F), Silver (SI=F), Copper (HG=F).
- **Agriculture:** Corn (ZC=F), Soybeans (ZS=F), Wheat (ZW=F), Coffee (KC=F).

D.2.3 METEOROLOGICAL TIME SERIES

To evaluate predictions on complex physical systems, we collect time-series weather data from 10 major U.S. cities, selected for their diverse geographical and climatic profiles. The locations (with

latitude and longitude) include: Chicago (41.8781, -87.6298), New York (40.7128, -74.0060), San Francisco (37.7749, -122.4194), Washington D.C. (38.9072, -77.0369), Miami (25.7617, -80.1918), Los Angeles (34.0522, -118.2437), Houston (29.7604, -95.3698), Seattle (47.6062, -122.3321), Boston (42.3601, -71.0589), and Denver (39.7392, -104.9903).

D.2.4 ELECTRICITY TIME SERIES

To assess forecasting capabilities on critical energy infrastructure, we retrieve electricity load data via the Grid Status API (Kanter & Grid Status, 2026). The dataset includes load measurements from three major metropolitan areas: New York, Boston, and San Francisco. The data is collected at a 5-minute frequency, providing a high-resolution record of power demand fluctuations across diverse urban environments. While the raw data is sampled at 5-minute intervals, we aggregate it into hourly time series for the forecasting task, where each hourly value represents the mean of all observations within that hour.

D.2.5 EVENT FORECASTING

Our benchmark incorporates, trending 30 diverse, real-world predictive tasks to test LLMs on reasoning over dynamic, high-uncertainty scenarios, sourcing from Polymarket. Rather than relying on static question-answering, we carefully curate these events across three core categories to evaluate distinct facets of real-world decision-making:

- **Tech & AI:** Aimed at testing the models' responsiveness to rapid technological advancements and digital culture. This involves predicting industry milestones such as AI model supremacy, major tech-focused public addresses, and broader internet trends (e.g., top search volumes and best AI Models).
- **Economy & Business:** To assess the models' grasp of macroeconomic policies and market-driving fiscal shifts, this category tracks critical financial indicators. Specific targets include consecutive Federal Reserve interest rate decisions and the regulatory rollout of major prediction platforms.
- **Politics & World Affairs:** Designed to evaluate reasoning capabilities within evolving international landscapes. This category encompasses diverse global electoral outcomes across multiple jurisdictions, key diplomatic developments impacting regional stability, and notable administrative shifts in global governance.

D.3 PROMPTS

This section details the specific prompt templates utilized across our diverse evaluation tasks. We categorize these prompts into two primary groups based on their forecasting objectives:

- **Time Series and Event Forecasting:** The prompt designs for exact point and probability predictions are presented in Figure 11 (weather), Figure 12 (commodities and stocks price forecasting), and Figure 13 (events forecasting).
- **Trend Prediction:** The templates designed for directional classification are illustrated in Figure 14 (3-class market trend), Figure 15 (5-class market trend), and Figure 16 (weather trend).
- **Retrieval Agent and Predictability Prompt:** The Search and Fact-Check agent prompt for stock events and Predictability Tests prompt are illustrated in Figure 9, 21 and 22.

Note that redundant or similar prompts are omitted here. The complete set of prompts is available in our repository[5].

[5] https://anonymous.4open.science/r/LEAF-4565

E EXPERIMENT SUPPLEMENTARY

E.1 MODELS AND COMPUTATION RESOURCES

Our study evaluates a total of 16 large language models (LLMs), ranging from cutting-edge proprietary models to prominent open-weight models, alongside two time-series foundation models (TSFMs) as baselines.

- **Proprietary Models (13 in total):** We evaluate a comprehensive suite of the latest proprietary models, including the Claude family (Claude 4.6 Sonnet, Claude 5 Sonnet, Claude 4.6 Opus, Claude 4.8 Opus, Claude 5 Opus), the Gemini series (Gemini, Gemini 3.1 Flash, Gemini 3.5 Flash, Gemini 3.7 Flash), the Grok 4 variants (Grok 4.1, Grok 4.20, Grok 4.20 Non-Reasoning). These state-of-the-art models exhibit frontier performance in reasoning and zero-shot tasks.
- **Open-weight Models (4 in total):** To examine the effectiveness of publicly accessible architectures, we include four prominent models: the Qwen 3.5 series (Qwen3.5-27B and Qwen3.5-9B) (Qwen Team, 2026), alongside the Gemma 3 series (Gemma-3-27B-it and Gemma-3-12B-it) (Gemma Team et al., 2025).
- **Time-Series Foundation Models (TSFMs):** To compare the text-driven forecasting capabilities of LLMs against purely data-driven quantitative approaches, we establish robust baselines using two prominent TSFMs: Chronos-2 Ansari et al. (2024) and TimesFM 3.0 Das et al. (2024), a decoder-only foundation model pre-trained on massive datasets.

The retrieval agent is powered by Gemini 3.1 Pro, while the fact-checking agents utilize Gemini 3.1 Pro and Claude 4.6 Sonnet. Inference for proprietary models is conducted via their respective APIs. The primary token consumption depends on the domain and temporal coverage; on average, processing a single event consumes approximately 3,864 tokens, inclusive of the fact-checking process. The open-weight models are deployed locally on NVIDIA A100 GPUs. Further implementation details and configurations are available in our repository.

E.2 OVERALL SCORE

The Overall score is a single scalar in $[0, 1]$ that aggregates per-domain metrics onto a common "higher is better" scale. We evaluate models across nine metrics, denoted by index k: Stock H1D, Stock H3M, Polymarket, Electricity 3-Class, Electricity TS, Electricity Max/Min, Weather 3-Class, Weather TS, and Weather Max/Min. Given the inherent difficulty of real-world forecasting, our current evaluation focuses on these specific tasks as they exhibit relatively strong predictability.

Let $x_k(m)$ represent the raw performance value of model m on metric k. We first map each raw metric to a normalized score, $\text{score}_k(m) \in [0, 1]$, based on its evaluation type:

$$\text{score}_k(m) = \begin{cases} \frac{x_k(m)}{100}, & \text{if } k \text{ is an accuracy metric;} \\ 1 - x_k(m), & \text{if } k \text{ is the Polymarket Brier score;} \\ 1 - \frac{x_k(m) - \min_{m'} x_k(m')}{\max_{m'} x_k(m') - \min_{m'} x_k(m')}, & \text{if } k \text{ is an MAE metric.} \end{cases} \tag{1}$$

Specifically, the accuracy metrics include Stock H1D, Stock H3M, Electricity 3-Class, and Weather 3-Class. The MAE metrics consist of Electricity TS, Electricity Max/Min, Weather TS, and Weather Max/Min. For MAE metrics, values are min-max normalized across all evaluated models m' and subsequently inverted.

Finally, the Overall score for model m is computed as the unweighted mean of its available normalized metric scores:

$$\text{Overall}(m) = \frac{1}{|K(m)|} \sum_{k \in K(m)} \text{score}_k(m), \tag{2}$$

where $K(m)$ denotes the subset of metrics evaluated for model m (excluding missing or unavailable tasks). A higher Overall score indicates better aggregate performance.

Table 9: The market price forecasting performance of LLMs. The results demonstrate that while LLMs achieve some forecasting improvements by incorporating events (w/ Evts), the Baseline Last Value (BLV) still significantly outperforms all other methods. The best value per column is underlined.

Domain	Model	Horizon: 3 Days				Horizon: 7 Days			
		TS	w/ Evts	TS	w/ Evts	TS	w/ Evts	TS	w/ Evts
Stock	*Metrics*	*MAPE (%)*		*DLV*		*MAPE (%)*		*DLV*	
	Gemini 3.1 Pro	2.748	2.873	1.125	1.481	3.865	4.028	1.419	2.060
	Claude 4.6 Opus	2.679	2.646	0.825	0.945	3.787	3.796	1.115	1.309
	Claude 4.8 Opus	2.691	2.640	0.911	0.897	3.813	3.768	1.171	1.214
	Claude Sonnet 5	2.711	2.606	0.956	0.910	3.819	3.751	1.295	1.217
	Grok 4.1 Fast	2.970	2.698	1.620	1.135	4.159	3.903	2.244	1.800
	TimesFM 3.0	2.909		1.192		4.157		1.865	
	Chronos-2	2.691		0.708		3.934		1.279	
	Base: Hist Mean	4.865				5.619			
	Base: Last Value	2.539				3.619			
Commodity	*Metrics*	*MAPE (%)*		*DLV*		*MAPE (%)*		*DLV*	
	Gemini 3.1 Pro	2.711	2.679	0.885	1.301	3.411	3.515	1.161	1.706
	Claude 4.6 Opus	2.697	2.620	0.791	0.968	3.377	3.285	1.031	1.258
	Claude 4.8 Opus	2.646	2.644	0.872	1.142	3.384	3.387	1.214	1.536
	Claude Sonnet 5	2.670	2.671	0.845	1.086	3.371	3.361	1.159	1.500
	Grok 4.1 Fast	2.687	2.673	1.091	1.191	3.470	3.513	1.231	1.719
	TimesFM 3.0	2.840		1.029		3.741		1.585	
	Chronos-2	2.734		0.648		3.630		1.117	
	Base: Hist Mean	4.325				4.763			
	Base: Last Value	2.540				3.236			

E.3 Case Studies

To display the reasoning process of how LLMs extract signals from complex events to aid forecasting, we present cases here. First, Figure 17 presents an event forecasting scenario where the model predicts the Federal Reserve's interest rate decision by synthesizing policy projections, labor market data, and macroeconomic constraints. Second, Figure 18 demonstrates a market trend prediction for Broadcom Inc., highlighting the model's ability to extract bullish signals from AI hardware partnerships and broader market sentiment. Finally, Figure 19 provides an example of meteorological time series forecasting guided by relevant weather events, such as "Pineapple Express". To further illustrate the necessity of our dual-agent cross-validation mechanism, we present a qualitative case study demonstrating how information leakage occurs when the fact-check module is disabled. In the event forecasting task shown in Figure 20, the strict temporal cut-off is set to **2025-06-17**, while the retrieved evidence "[141]" claims a valid timestamp, its actual information originates from post-cutoff events on 2025-09-07. The fact-checking process for this event, which successfully detects and filters out this temporal leakage, is detailed in Figure 21.

E.4 Dataset Taxonomy

We provide a detailed breakdown of our dataset taxonomy. Listing 1 and 2 illustrates examples of the event-augmented inputs across the four evaluated domains: Event Forecasting, Weather, Stock, and Commodities. These inputs combine textual event descriptions with historical numerical time-series data to form comprehensive multimodal prompts for the forecasting models. Due to space constraints, the complete data structures are omitted here.

F Additional Experiment Results

F.1 Forecast Accurate Market Price is Infeasible

Grounded in the Semi-Strong Efficient Market Hypothesis (EMH) (Fama, 1970), all publicly available information is rapidly incorporated into asset prices. Consequently, future price movements are driven almost exclusively by unforeseeable future events. Accurately predicting exact future prices would essentially require forecasting all unknown future occurrences and precisely quantifying their market impact. It is an inherently impossible endeavor. Under the Random Walk Hypothesis Fama (1965), a direct mathematical corollary of the EMH, future price changes are essentially zero-mean random noise driven by unrevealed information. Consequently, attempting to predict exact future prices largely equates to forecasting this unpredictable noise. Therefore, as shown in Table 9, in highly traded domains such as **stocks and commodities**, the current price naturally serves as the optimal, unbiased estimator. This supports the empirical finding that the Last Value Baseline (LVB) dominates exact point forecasting.

However, market efficiency is rarely absolute. Because markets can exhibit delayed responses (underreaction) to highly complex or impactful news, strong signals may not be fully absorbed within a single trading day. Consequently, market *trend forecasting* remains a viable endeavor. Additionally, our current focus on highly liquid targets likely amplifies the effects of the Efficient Market Hypothesis, leading to highly efficient signal absorption and inherently greater forecasting difficulty. To better modulate the benchmark's difficulty, we will explore alternative forecasting targets in future work.

F.2 Detailed Results

F.2.1 Stock Experiments

Additional stock trend forecasting results are provided in Tables 12 and 13. As observed in Table 13, proprietary models frequently underperform the random guess baseline in the 5-way long-term mid-cap stock trend forecasting task. This degradation primarily stems from a severe mismatch between the models' conservative prediction behaviors and a heavy-tailed ground-truth distribution.

As detailed in Table 10, frontier models systematically avoid predicting the extreme classes ("Strong Up" and "Strong Down"). Conversely, these two categories constitute the vast majority of the actual ground truth, leading to significant deviations from reality. Under our current formulation, the price change boundaries are defined as $< 5\%$ for "Neutral" and $> 15\%$ for "Strong". While raising the "Strong" threshold could balance the label distribution and artificially bring results closer to the baseline, such tuning would simply overfit to the volatility of this specific evaluation period.

Instead of retroactively manipulating thresholds, we choose to maintain our fixed definitions. We highlight this conservative prediction bias as an inherent **limitation of both current LLMs and static-threshold experimental setups** in trend forecasting, and present this analysis to provide deeper insights into model behavior. Ultimately, a highly capable forecasting LLM should remain robust against such degrees of class imbalance.

Table 10: Prediction distributions across the 5-way classification categories for long-term mid-cap stock forecasting. Proprietary models disproportionately favor moderate and neutral classes, whereas the ground truth is dominated by extreme price movements.

Model	Strong Up (%)	Up (%)	Neutral (%)	Down (%)	Strong Down (%)
Gemini 3.1 Pro (w/ Evts)	10.87	25.28	22.42	36.11	5.32
Gemini 3.1 Pro (TS)	3.57	23.61	40.44	24.40	7.98
Gemini 3.1 Flash (w/ Evts)	0.08	40.36	37.98	21.55	0.04
Gemini 3.1 Flash (TS)	0.00	1.35	84.48	13.93	0.24
Claude 4.6 Opus (w/ Evts)	0.40	68.89	22.34	8.29	0.08
Claude 4.6 Opus (TS)	0.04	55.63	31.75	12.58	0.00
Claude 4.6 Sonnet (w/ Evts)	0.67	49.72	21.79	27.70	0.12
Claude 4.6 Sonnet (TS)	0.12	37.06	32.50	30.28	0.04
Ground Truth	**43.69**	**9.05**	**9.17**	**8.69**	**29.40**

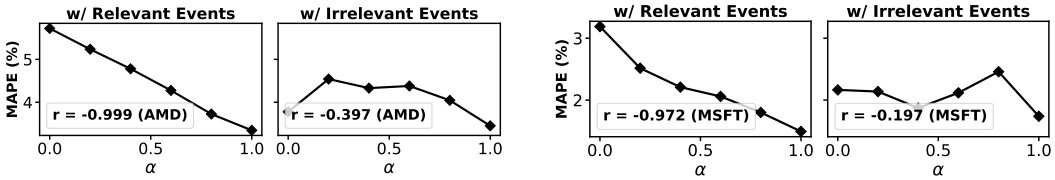

Figure 8: Examples comparing forecasting performance with (w/) relevant or irrelevant events as the proportion (α, x-axis) of "consistent" text increases. The correlation coefficient is denoted by r.

Figure 8 presents two examples of data validation, as mentioned in Section 4.4. The results indicate that when target-relevant events are utilized to assist forecasting, the variations in prediction outcomes strongly correlate with the trend of α. Conversely, no significant correlation is observed when irrelevant events are used. Furthermore, Table 11 illustrates the impact of outdated data on evaluation performance, as mentioned in Section 3.4. Specifically, utilizing historical data significantly inflates the models' apparent capabilities, highlighting a evaluation data contamination problem.

Table 14 details the relationship between predictability levels and forecasting performance, as discussed in §4.4. As observed, models achieve higher forecasting accuracy on samples assigned high predictability confidence by the models themselves. However, this correlation vanishes when events are excluded, indicating that textual events are a prerequisite for models to make reliable, high-confidence predictions.

F.2.2 Other Results

This section provides supplementary evaluation results across different tasks. Table 15 details the event forecasting performance at varying time-to-resolution cut-offs. For the weather and electricity domains, Table 16 reports the comprehensive trend classification results, while Table 17 and Table 18 present the fine-grained numerical errors for hourly point forecasting and daily extremum (Max/Min) predictions, respectively.

Table 11: **Temporal Ablation on "Living" Data:** Accuracy (%) comparison for 3-class and 5-class stock predictions using pure time-series inputs. The "Before" columns denote a historical test period (Jan 2, 2023 – Jun 2, 2024) preceding the general training cut-offs, where models exhibit artificially inflated accuracy due to data contamination. The "After" columns represent strictly unseen, out-of-sample "living" data (Oct 1, 2025 – Apr 5, 2026). Green shading highlights instances where historical performance is strictly greater than the true out-of-sample performance.

		Large-Cap Stocks Short-Term				**Mid-Cap Stocks Long-Term**			
		Horizon: 1D		Horizon: 2D		Horizon: 3M		Horizon: 6M	
Model (TS Only)		After	Before	After	Before	After	Before	After	Before
3-class	Gemini 3.1 Pro	37.33	66.90	41.27	73.10	40.76	86.93	46.02	92.36
	Claude 4.6 Opus	37.81	44.73	37.97	47.29	34.86	74.74	42.23	82.97
	Baseline (Random Guess)	33.33				33.33			
5-Class	Gemini 3.1 Pro	30.51	65.32	23.63	62.66	10.90	26.14	13.24	22.91
	Claude 4.6 Opus	29.49	35.76	21.34	28.87	9.86	49.93	7.87	68.43
	Baseline (Random Guess)	20.00				20.00			

Table 12: Accuracy comparison for ternary (3-class) classification between time-series-only (TS) and event-inclusive (w/ Evts) forecasting for long-term mid-cap and short-term large-cap stocks, for the completed 16 models. Each cell is the mean over the pooled samples ± the half-width of its 95 % confidence interval.

Model	**Mid-Cap Stocks Long-Term**				**Large-Cap Stocks Short-Term**			
	Horizon: 3 Months		Horizon: 6 Months		Horizon: 1 Day		Horizon: 2 Days	
	TS	w/ Evts	TS	w/ Evts	TS	w/ Evts	TS	w/ Evts
Gemini 3.1 Pro	33.8 ±6.1	35.9 ±6.2	40.2 ±6.3	49.1 ±6.4	38.7 ±2.5	40.9 ±2.6	42.3 ±2.6	44.9 ±2.6
Gemini 3.1 Flash	23.5 ±5.4	29.5 ±5.9	19.7 ±5.1	33.3 ±6.1	33.8 ±2.5	37.9 ±2.5	32.7 ±2.4	36.7 ±2.5
Gemini 3.5 Flash	28.0 ±5.8	42.1 ±6.4	27.3 ±5.8	41.2 ±6.3	38.2 ±2.5	39.8 ±2.6	43.5 ±2.6	43.4 ±2.6
Gemini 3.7 Flash	32.9 ±6.0	35.0 ±6.1	34.2 ±6.1	38.0 ±6.2	37.1 ±2.5	39.0 ±2.5	43.3 ±2.6	46.2 ±2.6
Claude 4.6 Opus	31.2 ±5.9	32.1 ±6.0	26.9 ±5.7	28.6 ±5.8	36.0 ±2.5	40.2 ±2.6	41.8 ±2.6	45.5 ±2.6
Claude 4.6 Sonnet	37.2 ±6.2	37.2 ±6.2	34.6 ±6.1	37.2 ±6.2	36.7 ±2.5	44.1 ±2.6	40.1 ±2.6	44.3 ±2.6
Claude 4.8 Opus	27.8 ±5.8	30.3 ±5.9	32.9 ±6.0	34.6 ±6.1	37.6 ±2.5	40.9 ±2.6	40.3 ±2.6	42.8 ±2.6
Claude Opus 5	41.5 ±6.3	42.3 ±6.3	54.7 ±6.4	59.8 ±6.3	39.7 ±2.6	38.9 ±2.5	44.0 ±2.6	45.7 ±2.6
Claude Sonnet 5	29.1 ±5.8	32.5 ±6.0	31.2 ±5.9	41.0 ±6.3	34.5 ±2.5	38.6 ±2.5	33.0 ±2.5	39.7 ±2.6
Grok 4.1 Fast	19.2 ±5.1	24.8 ±5.5	11.5 ±4.1	27.4 ±5.7	37.3 ±2.5	36.7 ±2.5	36.2 ±2.5	39.9 ±2.6
Grok 4.20	14.1 ±4.5	20.5 ±5.2	6.4 ±3.1	26.9 ±5.7	38.8 ±2.5	41.1 ±2.6	39.6 ±2.6	42.3 ±2.6
Grok 4.20 non-reason	17.1 ±4.8	14.1 ±4.5	13.2 ±4.4	15.0 ±4.6	34.0 ±2.5	38.4 ±2.5	32.6 ±2.4	35.6 ±2.5
Gemma3-27B-it	26.1 ±5.6	33.8 ±6.1	20.1 ±5.1	33.3 ±6.1	33.5 ±2.5	38.8 ±2.5	34.6 ±2.5	39.4 ±2.6
Gemma3-12B-it	17.9 ±4.9	31.2 ±5.9	13.2 ±4.4	32.5 ±6.0	33.3 ±2.5	39.7 ±2.6	35.7 ±2.5	37.9 ±2.5
Qwen3.5-27B	17.9 ±4.9	26.5 ±5.7	11.1 ±4.0	33.8 ±6.1	36.7 ±2.5	41.6 ±2.6	38.8 ±2.5	43.5 ±2.6
Qwen3.5-9B	15.9 ±4.7	17.5 ±4.9	9.5 ±3.8	19.5 ±5.1	30.8 ±2.4	39.3 ±3.0	33.4 ±2.5	42.0 ±3.0

Table 13: Accuracy comparison for quinary (5-class) classification between time-series-only (TS) and event-inclusive (w/ Evts) forecasting for long-term mid-cap and short-term large-cap stocks, for the proprietary models. Each cell is the mean over the pooled samples ± the half-width of its 95 % confidence interval.

Model	**Mid-Cap Stocks Long-Term**				**Large-Cap Stocks Short-Term**			
	Horizon: 3 Months		Horizon: 6 Months		Horizon: 1 Day		Horizon: 2 Days	
	TS	w/ Evts	TS	w/ Evts	TS	w/ Evts	TS	w/ Evts
Gemini 3.1 Pro	13.7 ±4.4	14.5 ±4.5	12.4 ±4.2	16.7 ±4.8	23.9 ±2.2	27.6 ±2.3	22.3 ±2.2	29.0 ±2.4
Gemini 3.1 Flash	10.3 ±3.9	10.7 ±4.0	7.3 ±3.3	9.8 ±3.8	23.5 ±2.2	26.5 ±2.3	16.5 ±1.9	20.2 ±2.1
Gemini 3.5 Flash	14.1 ±4.5	15.0 ±4.6	10.7 ±4.0	9.4 ±3.7	22.6 ±2.2	26.5 ±2.3	26.0 ±2.3	26.0 ±2.3
Gemini 3.7 Flash	14.1 ±4.5	12.8 ±4.3	15.4 ±4.6	13.7 ±4.4	23.0 ±2.2	24.3 ±2.2	21.0 ±2.1	25.8 ±2.3
Claude 4.6 Opus	11.5 ±4.1	12.0 ±4.2	6.4 ±3.1	5.6 ±2.9	22.8 ±2.2	26.5 ±2.3	18.9 ±2.0	21.5 ±2.1
Claude 4.6 Sonnet	11.5 ±4.1	10.3 ±3.9	8.5 ±3.6	6.8 ±3.2	23.9 ±2.2	24.4 ±2.2	21.0 ±2.1	25.7 ±2.3
Claude 4.8 Opus	10.3 ±3.9	12.8 ±4.3	9.0 ±3.7	6.0 ±3.0	24.3 ±2.2	24.3 ±2.2	20.2 ±2.1	21.3 ±2.1
Claude Opus 5	14.1 ±4.5	13.7 ±4.4	31.2 ±5.9	37.6 ±6.2	25.6 ±2.3	27.3 ±2.3	23.0 ±2.2	24.1 ±2.2
Claude Sonnet 5	9.0 ±3.7	6.8 ±3.2	9.4 ±3.7	9.0 ±3.7	24.0 ±2.2	25.0 ±2.3	18.5 ±2.0	20.6 ±2.1
Grok 4.1 Fast	12.8 ±4.3	6.4 ±3.1	6.0 ±3.0	8.5 ±3.6	23.3 ±2.2	23.6 ±2.2	18.9 ±2.0	22.0 ±2.2
Grok 4.20	14.1 ±4.5	9.8 ±3.8	6.4 ±3.1	7.7 ±3.4	22.9 ±2.2	23.5 ±2.2	18.7 ±2.0	22.1 ±2.2
Grok 4.20 non-reason	12.8 ±4.3	12.0 ±4.2	9.4 ±3.7	6.0 ±3.0	24.8 ±2.3	24.9 ±2.3	17.0 ±2.0	18.9 ±2.0
Gemma3-27B-it	12.8 ±4.3	9.4 ±3.7	6.8 ±3.2	10.7 ±4.0	24.8 ±2.3	25.7 ±2.3	20.6 ±2.1	22.6 ±2.2
Gemma3-12B-it	12.4 ±4.2	7.7 ±3.4	6.0 ±3.0	6.8 ±3.2	24.5 ±2.2	25.7 ±2.3	20.9 ±2.1	21.2 ±2.1
Qwen3.5-27B	12.4 ±4.2	9.0 ±3.7	7.7 ±3.4	8.5 ±3.6	23.3 ±2.2	26.4 ±2.4	18.6 ±2.0	24.0 ±2.3
Qwen3.5-9B	13.2 ±4.4	-	-	9.0 ±3.7	23.1 ±2.2	-	17.0 ±2.0	-

Table 14: Accuracy (%) and correlation analysis across different Predictability Levels for Large-Cap and Mid-Cap stocks and both Gemini 3.1 Pro and Claude 4.6 Opus.

		Large-Cap Stocks Short-Term				**Mid-Cap Stocks Long-Term**			
		Horizon: 1		Horizon: 2		Horizon: 3 Months		Horizon: 6 Months	
Metrics		TS	w/ Evts	TS	w/ Evts	TS	w/ Evts	TS	w/ Evts
Gemini 3.1 Pro	Level 1	39.32	35.27	41.85	42.24	35.96	34.21	48.09	38.10
	Level 2	34.93	40.85	42.16	38.96	42.38	42.58	44.79	48.07
	Level 3	37.37	41.02	39.19	39.03	40.84	43.08	41.67	48.04
	Level 4	40.85	46.72	40.96	43.69	46.05	51.82	49.43	55.14
	Level 5	34.80	52.22	42.58	52.50	46.97	56.06	54.50	63.58
	Correlation	-0.184	0.974	0.031	0.720	0.919	0.978	0.569	0.970
Claude 4.6 Opus	Level 1	38.13	34.02	36.51	37.24	26.87	34.81	29.82	40.88
	Level 2	38.03	37.94	39.03	44.24	30.97	37.96	48.54	53.53
	Level 3	37.66	42.62	37.50	40.15	38.37	45.58	45.83	56.03
	Level 4	40.85	38.60	40.91	45.08	50.82	66.67	45.30	57.58
	Level 5	29.22	42.92	32.17	48.92	75.83	56.48	58.10	51.28
	Correlation	-0.538	0.792	-0.328	0.845	0.945	0.861	0.829	0.596

Table 15: Event forecasting performance across six time-to-resolution cut-offs from 100% to 10% based on the timeframes in Polymarket. The prediction error of context-augmented models consistently decreases as the resolution date approaches; each cell is the mean Brier score over the resolved markets ± the half-width of its 95 % confidence interval (1.96 standard errors). Green (red) marks a w/ Evts mean that improves on (degrades from) its event-free baseline. The prediction error of context-augmented models consistently decreases as the resolution date approaches.

Model	Metric	Avg Brier Score					
	Time to Resolution	100%	80%	60%	40%	20%	10%
Gemini 3.1 Pro	baseline	0.478 ±0.084	0.505 ±0.087	0.490 ±0.086	0.537 ±0.091	0.509 ±0.090	0.517 ±0.090
	w/ Evts	0.434 ±0.091	0.401 ±0.095	0.388 ±0.091	0.358 ±0.094	0.322 ±0.096	0.255 ±0.096
Gemini 3.1 Flash	baseline	0.461 ±0.091	0.474 ±0.094	0.460 ±0.098	0.496 ±0.098	0.493 ±0.100	0.442 ±0.099
	w/ Evts	0.412 ±0.085	0.381 ±0.088	0.336 ±0.088	0.310 ±0.087	0.306 ±0.090	0.271 ±0.095
Gemini 3.5 Flash	baseline	0.410 ±0.087	0.395 ±0.086	0.427 ±0.090	0.430 ±0.094	0.462 ±0.101	0.429 ±0.096
	w/ Evts	0.414 ±0.089	0.367 ±0.091	0.321 ±0.087	0.274 ±0.087	0.265 ±0.091	0.212 ±0.090
Gemini 3.7 Flash	baseline	0.432 ±0.081	0.428 ±0.082	0.450 ±0.087	0.454 ±0.089	0.443 ±0.093	0.491 ±0.083
	w/ Evts	0.398 ±0.082	0.384 ±0.086	0.357 ±0.087	0.334 ±0.087	0.289 ±0.091	0.207 ±0.085
Claude 4.6 Opus	baseline	0.482 ±0.089	0.479 ±0.091	0.478 ±0.090	0.467 ±0.091	0.463 ±0.094	0.461 ±0.090
	w/ Evts	0.489 ±0.092	0.422 ±0.093	0.408 ±0.090	0.350 ±0.090	0.282 ±0.089	0.233 ±0.082
Claude 4.6 Sonnet	baseline	0.486 ±0.083	0.470 ±0.085	0.453 ±0.082	0.456 ±0.086	0.444 ±0.088	0.454 ±0.079
	w/ Evts	0.502 ±0.088	0.420 ±0.092	0.380 ±0.089	0.338 ±0.087	0.289 ±0.088	0.224 ±0.084
Claude 4.8 Opus	baseline	0.449 ±0.081	0.481 ±0.085	0.438 ±0.087	0.446 ±0.088	0.447 ±0.091	0.423 ±0.086
	w/ Evts	0.416 ±0.085	0.413 ±0.090	0.375 ±0.088	0.369 ±0.086	0.306 ±0.089	0.255 ±0.091
Claude Opus 5	baseline	0.458 ±0.087	0.446 ±0.088	0.440 ±0.090	0.436 ±0.093	0.432 ±0.094	0.440 ±0.092
	w/ Evts	0.417 ±0.089	0.399 ±0.088	0.410 ±0.091	0.370 ±0.090	0.327 ±0.090	0.288 ±0.090
Claude Sonnet 5	baseline	0.461 ±0.085	0.454 ±0.084	0.446 ±0.086	0.465 ±0.089	0.475 ±0.092	0.452 ±0.086
	w/ Evts	0.408 ±0.085	0.407 ±0.090	0.384 ±0.086	0.350 ±0.084	0.292 ±0.089	0.247 ±0.090
Grok 4.1 Fast	baseline	0.590 ±0.089	0.581 ±0.090	0.617 ±0.090	0.632 ±0.089	0.630 ±0.091	0.629 ±0.095
	w/ Evts	0.451 ±0.092	0.406 ±0.091	0.333 ±0.081	0.317 ±0.081	0.316 ±0.089	0.289 ±0.097
Grok 4.20	baseline	0.496 ±0.068	0.513 ±0.068	0.505 ±0.072	0.504 ±0.074	0.507 ±0.080	0.524 ±0.074
	w/ Evts	0.442 ±0.075	0.372 ±0.077	0.364 ±0.072	0.328 ±0.076	0.340 ±0.085	0.254 ±0.082
Grok 4.20 non-reason	baseline	0.567 ±0.081	0.574 ±0.084	0.583 ±0.082	0.571 ±0.084	0.571 ±0.091	0.592 ±0.087
	w/ Evts	0.524 ±0.085	0.475 ±0.092	0.470 ±0.088	0.438 ±0.092	0.414 ±0.096	0.381 ±0.096
Gemma3-27B-it	baseline	0.527 ±0.089	0.516 ±0.092	0.535 ±0.093	0.514 ±0.092	0.513 ±0.089	0.537 ±0.095
	w/ Evts	0.451 ±0.088	0.357 ±0.083	0.390 ±0.083	0.374 ±0.081	0.388 ±0.090	0.392 ±0.095
Gemma3-12B-it	baseline	0.598 ±0.088	0.602 ±0.089	0.580 ±0.089	0.609 ±0.092	0.575 ±0.097	0.621 ±0.096
	w/ Evts	0.506 ±0.090	0.396 ±0.082	0.383 ±0.076	0.379 ±0.080	0.367 ±0.080	0.446 ±0.089
Qwen3.5-27B	baseline	0.528 ±0.086	0.487 ±0.082	0.523 ±0.084	0.523 ±0.085	0.499 ±0.087	0.510 ±0.080
	w/ Evts	0.407 ±0.091	0.329 ±0.088	0.276 ±0.083	0.269 ±0.080	0.232 ±0.085	0.235 ±0.091
Qwen3.5-9B	baseline	0.486 ±0.087	0.521 ±0.087	0.505 ±0.089	0.506 ±0.090	0.496 ±0.091	0.523 ±0.089
	w/ Evts	0.442 ±0.090	0.392 ±0.092	0.377 ±0.091	0.333 ±0.087	0.277 ±0.087	0.264 ±0.091

Table 16: Accuracy comparison between using only time series (TS) and incorporating text events (w/ Evts) for ternary and quinary classifications in Electricity and Weather forecasting. Each cell is the mean over the pooled samples ± the half-width of its 95 % confidence interval. Green (red) marks a w/ Evts mean that improves on (degrades from) the TS-only mean. Certain models, such as Qwen, are excluded from the results because they failed to produce valid outputs for over one-third of the samples.

Domain	Model	Ternary Classification (3-Class)				Quinary Classification (5-Class)			
		Horizon: 3 Days		Horizon: 7 Days		Horizon: 3 Days		Horizon: 7 Days	
		TS	w/ Evts	TS	w/ Evts	TS	w/ Evts	TS	w/ Evts
Weather	Gemini 3.1 Pro	49.2 ±4.3	68.3 ±4.0	71.2 ±3.9	77.8 ±3.6	40.9 ±4.2	56.4 ±4.2	54.1 ±4.3	64.3 ±4.2
	Gemini 3.1 Flash	38.5 ±4.1	61.5 ±4.1	25.7 ±3.8	56.3 ±4.3	35.5 ±4.1	47.7 ±4.3	20.6 ±3.5	44.5 ±4.3
	Gemini 3.5 Flash	42.2 ±4.2	61.5 ±4.1	48.1 ±4.3	65.0 ±4.1	31.7 ±3.9	49.1 ±4.2	37.9 ±4.2	54.8 ±4.3
	Gemini 3.7 Flash	51.7 ±4.2	69.8 ±3.9	71.0 ±3.9	78.7 ±3.5	37.6 ±4.1	55.7 ±4.2	55.6 ±4.3	66.0 ±4.1
	Claude 4.6 Opus	47.2 ±4.2	61.3 ±4.1	59.2 ±4.2	66.9 ±4.0	28.5 ±3.8	47.8 ±4.2	32.1 ±4.0	52.3 ±4.3
	Claude 4.6 Sonnet	39.1 ±4.1	59.8 ±4.1	42.9 ±4.3	66.5 ±4.1	29.1 ±3.8	44.4 ±4.2	25.6 ±3.8	49.8 ±4.3
	Claude 4.8 Opus	34.6 ±4.0	63.0 ±4.1	26.2 ±3.8	57.1 ±4.3	30.2 ±3.9	51.3 ±4.2	16.3 ±3.2	46.0 ±4.3
	Claude Opus 5	45.9 ±4.2	69.3 ±3.9	59.4 ±4.2	69.2 ±4.0	36.1 ±4.1	56.9 ±4.2	41.7 ±4.2	60.2 ±4.2
	Claude Sonnet 5	37.4 ±4.1	63.7 ±4.1	48.3 ±4.3	68.7 ±4.0	26.7 ±3.7	47.2 ±4.2	29.2 ±3.9	52.1 ±4.3
	Grok 4.1 Fast	38.5 ±4.1	62.2 ±4.1	27.9 ±3.9	66.7 ±4.1	30.2 ±3.9	46.7 ±4.2	19.8 ±3.4	52.9 ±4.3
	Grok 4.20	37.8 ±4.1	62.6 ±4.1	36.9 ±4.2	66.0 ±4.1	26.3 ±3.7	47.8 ±4.2	19.8 ±3.4	53.7 ±4.3
	Grok 4.20 non-reason	35.7 ±4.0	59.6 ±4.1	24.8 ±3.7	62.5 ±4.2	33.0 ±4.0	43.0 ±4.2	20.6 ±3.5	47.5 ±4.3
	Gemma3-27B-it	32.0 ±3.9	53.3 ±4.2	37.5 ±4.2	58.5 ±4.2	14.6 ±3.0	35.2 ±4.0	15.4 ±3.1	46.2 ±4.3
	Gemma3-12B-it	36.7 ±4.1	51.9 ±4.2	51.3 ±4.3	63.8 ±4.1	15.7 ±3.1	34.1 ±4.0	15.6 ±3.1	44.8 ±4.3
Electricity	Gemini 3.1 Pro	51.8 ±7.5	61.8 ±7.3	62.8 ±7.4	70.7 ±7.0	38.2 ±7.3	47.1 ±7.5	44.5 ±7.6	51.8 ±7.7
	Gemini 3.1 Flash	32.4 ±7.1	52.4 ±7.5	27.4 ±6.9	54.3 ±7.6	20.6 ±6.1	32.4 ±7.1	21.3 ±6.3	27.4 ±6.9
	Gemini 3.5 Flash	50.6 ±7.5	52.9 ±7.5	59.1 ±7.5	57.9 ±7.6	38.2 ±7.3	40.0 ±7.4	45.7 ±7.6	36.6 ±7.4
	Gemini 3.7 Flash	51.8 ±7.5	59.4 ±7.4	71.3 ±6.9	70.7 ±7.0	37.6 ±7.3	50.0 ±7.5	51.8 ±7.7	56.7 ±7.6
	Claude 4.6 Opus	44.7 ±7.5	60.6 ±7.4	54.9 ±7.6	60.4 ±7.5	36.5 ±7.3	45.9 ±7.5	32.9 ±7.2	51.2 ±7.7
	Claude 4.6 Sonnet	45.9 ±7.5	54.1 ±7.5	55.5 ±7.6	61.0 ±7.5	34.7 ±7.2	40.6 ±7.4	31.7 ±7.1	42.1 ±7.6
	Claude 4.8 Opus	39.4 ±7.4	52.9 ±7.5	36.6 ±7.4	61.0 ±7.5	25.3 ±6.6	35.9 ±7.2	25.6 ±6.7	47.6 ±7.7
	Claude Opus 5	50.6 ±7.5	64.7 ±7.2	68.9 ±7.1	77.4 ±6.4	40.6 ±7.4	55.3 ±7.5	47.6 ±7.7	57.9 ±7.6
	Claude Sonnet 5	48.8 ±7.5	55.3 ±7.5	62.8 ±7.4	67.1 ±7.2	31.2 ±7.0	44.1 ±7.5	43.3 ±7.6	51.8 ±7.7
	Grok 4.1 Fast	32.9 ±7.1	51.2 ±7.5	26.8 ±6.8	50.0 ±7.7	26.5 ±6.7	42.9 ±7.5	20.1 ±6.2	41.5 ±7.6
	Grok 4.20	40.6 ±7.4	51.8 ±7.5	44.5 ±7.6	59.8 ±7.5	27.1 ±6.7	38.2 ±7.3	39.0 ±7.5	48.8 ±7.7
	Grok 4.20 non-reason	29.8 ±6.9	40.6 ±7.4	28.0 ±6.9	47.6 ±7.7	21.0 ±6.2	21.2 ±6.2	17.1 ±5.8	28.0 ±6.9
	Gemma3-27B-it	33.5 ±7.1	45.3 ±7.5	54.9 ±7.6	61.6 ±7.5	22.9 ±6.3	29.4 ±6.9	25.0 ±6.6	34.1 ±7.3
	Gemma3-12B-it	33.5 ±7.1	49.4 ±7.5	44.5 ±7.6	59.1 ±7.5	22.9 ±6.3	25.3 ±6.6	25.6 ±6.7	29.9 ±7.0

Table 17: Forecasting performance using only time series (TS) versus incorporating events (w/ Evts) for proprietary models on hourly point forecasting (72-hour and 168-hour horizons). MAE and MSE are computed on z-score normalized data (lower is better). As mentioned in Section 4.3, most small open-weight models are excluded from time series forecasting tasks. Each cell is the mean over all forecasts of the per-forecast MAE or MSE ± the half-width of its 95 % confidence interval. Green (red) marks a w/ Evts mean that improves on (degrades from) the TS-only mean. The best value per column is underlined, and solid cells denote models that cannot process text.

Domain	Model	Horizon: 3 Days				Horizon: 7 Days			
		TS	w/ Evts	TS	w/ Evts	TS	w/ Evts	TS	w/ Evts
Weather	*Metrics*	*MAE*		*MSE*		*MAE*		*MSE*	
	Gemini 3.1 Pro	0.810 ±0.038	0.780 ±0.032	1.264 ±0.128	1.081 ±0.088	0.992 ±0.038	0.924 ±0.034	1.756 ±0.131	1.480 ±0.106
	Gemini 3.1 Flash	1.120 ±0.033	1.097 ±0.033	2.050 ±0.131	1.945 ±0.117	1.207 ±0.031	1.109 ±0.030	2.338 ±0.126	2.027 ±0.108
	Gemini 3.5 Flash	0.949 ±0.033	0.924 ±0.030	1.576 ±0.116	1.456 ±0.096	1.105 ±0.032	1.028 ±0.028	2.016 ±0.114	1.721 ±0.094
	Gemini 3.7 Flash	0.917 ±0.032	0.864 ±0.027	1.488 ±0.111	1.276 ±0.082	1.049 ±0.030	1.006 ±0.027	1.847 ±0.104	1.640 ±0.088
	Claude 4.6 Opus	0.847 ±0.035	0.735 ±0.024	1.308 ±0.115	0.956 ±0.067	0.988 ±0.033	0.855 ±0.026	1.681 ±0.110	1.263 ±0.081
	Claude 4.6 Sonnet	0.811 ±0.034	0.718 ±0.027	1.198 ±0.105	0.959 ±0.083	0.974 ±0.034	0.850 ±0.028	1.657 ±0.116	1.304 ±0.096
	Claude 4.8 Opus	0.926 ±0.034	0.910 ±0.030	1.518 ±0.120	1.421 ±0.106	1.061 ±0.032	1.002 ±0.026	1.897 ±0.117	1.646 ±0.090
	Claude Opus 5	0.808 ±0.038	0.639 ±0.025	1.252 ±0.125	0.739 ±0.059	0.973 ±0.034	0.774 ±0.027	1.680 ±0.114	1.042 ±0.076
	Claude Sonnet 5	0.801 ±0.036	0.737 ±0.029	1.195 ±0.116	0.965 ±0.078	0.967 ±0.036	0.825 ±0.031	1.646 ±0.121	1.211 ±0.100
	Grok 4.1 Fast	1.004 ±0.040	0.862 ±0.031	1.789 ±0.147	1.328 ±0.109	1.141 ±0.039	0.973 ±0.029	2.223 ±0.160	1.614 ±0.102
	Grok 4.20	1.077 ±0.033	1.000 ±0.032	1.959 ±0.127	1.697 ±0.126	1.170 ±0.032	1.098 ±0.029	2.257 ±0.129	1.962 ±0.108
	Grok 4.20 non-reason	-	-	-	-	-	-	-	-
	TimesFM 3.0	0.699 ±0.038		0.956 ±0.115		0.858 ±0.039		1.349 ±0.128	
	Chronos-2	0.693 ±0.037		0.942 ±0.111		0.859 ±0.038		1.346 ±0.125	
	Base: Hist Mean	0.877 ±0.042		1.296 ±0.125		0.954 ±0.042		1.545 ±0.135	
	Base: Last Value	0.848 ±0.043		1.328 ±0.141		0.988 ±0.044		1.737 ±0.160	
Electricity	*Metrics*	*MAE*		*MSE*		*MAE*		*MSE*	
	Gemini 3.1 Pro	0.571 ±0.042	0.486 ±0.031	0.619 ±0.088	0.459 ±0.055	0.695 ±0.041	0.590 ±0.031	0.847 ±0.098	0.626 ±0.067
	Gemini 3.1 Flash	0.945 ±0.052	0.961 ±0.049	1.498 ±0.155	1.513 ±0.149	1.024 ±0.052	1.148 ±0.068	1.756 ±0.166	2.164 ±0.273
	Gemini 3.5 Flash	0.687 ±0.047	0.676 ±0.043	0.883 ±0.114	0.856 ±0.108	0.803 ±0.047	0.764 ±0.043	1.127 ±0.125	1.027 ±0.112
	Gemini 3.7 Flash	0.584 ±0.042	0.563 ±0.037	0.642 ±0.101	0.603 ±0.080	0.695 ±0.040	0.650 ±0.034	0.857 ±0.097	0.761 ±0.077
	Claude 4.6 Opus	0.738 ±0.044	0.675 ±0.038	0.950 ±0.109	0.799 ±0.088	0.842 ±0.044	0.774 ±0.038	1.215 ±0.121	1.012 ±0.095
	Claude 4.6 Sonnet	0.743 ±0.041	0.864 ±0.197	0.984 ±0.107	2.824 ±2.238	0.840 ±0.039	0.892 ±0.134	1.210 ±0.111	1.987 ±1.202
	Claude 4.8 Opus	0.781 ±0.044	0.736 ±0.044	1.098 ±0.119	1.009 ±0.119	0.890 ±0.041	0.831 ±0.041	1.340 ±0.120	1.192 ±0.109
	Claude Opus 5	0.484 ±0.040	0.423 ±0.029	0.465 ±0.073	0.360 ±0.050	0.575 ±0.040	0.496 ±0.030	0.615 ±0.079	0.469 ±0.057
	Claude Sonnet 5	0.603 ±0.043	0.511 ±0.033	0.677 ±0.093	0.502 ±0.066	0.686 ±0.040	0.577 ±0.030	0.831 ±0.091	0.592 ±0.059
	Grok 4.1 Fast	0.753 ±0.047	0.872 ±0.045	1.077 ±0.135	1.332 ±0.132	0.810 ±0.043	0.988 ±0.050	1.173 ±0.120	1.716 ±0.170
	Grok 4.20	0.836 ±0.046	0.772 ±0.043	1.251 ±0.131	1.103 ±0.118	0.860 ±0.045	0.842 ±0.041	1.281 ±0.126	1.213 ±0.113
	Grok 4.20 non-reason	0.784 ±0.045	0.821 ±0.051	1.065 ±0.121	1.144 ±0.138	0.849 ±0.049	-	1.245 ±0.154	-
	TimesFM 3.0	0.491 ±0.037		0.457 ±0.065		0.570 ±0.040		0.592 ±0.078	
	Chronos-2	0.474 ±0.035		0.424 ±0.057		0.563 ±0.038		0.569 ±0.072	
	Base: Hist Mean	0.645 ±0.046		0.690 ±0.090		0.691 ±0.047		0.800 ±0.096	
	Base: Last Value	0.613 ±0.039		0.671 ±0.080		0.674 ±0.040		0.808 ±0.093	

Table 18: Daily Max/Min forecasting performance using only time series (TS) versus incorporating events (w/ Evts): the model forecasts the daily high and the daily low of the target series over the horizon; the two series are pooled and each is normalised by its own look-back. MAE and MSE, lower is better; intervals, colouring and underlining as in Table 17. Baseline rows are zero-shot TimesFM 3.0 and Chronos-2 given the same 14-day max/min look-back as a two-variate series, plus the naive baselines adapted to this setting: Last Value repeats the most recent day's max and min over the horizon, Hist Mean predicts the mean daily max and min of the preceding 14 days

Domain	Model	Horizon: 3 Days				Horizon: 7 Days			
		TS	w/ Evts	TS	w/ Evts	TS	w/ Evts	TS	w/ Evts
Weather	*Metrics*	*MAE*		*MSE*		*MAE*		*MSE*	
	Gemini 3.1 Pro	1.036 ±0.049	0.769 ±0.035	1.919 ±0.198	1.119 ±0.115	1.163 ±0.046	0.899 ±0.034	2.384 ±0.224	1.497 ±0.124
	Gemini 3.1 Flash	1.003 ±0.053	0.832 ±0.039	1.904 ±0.229	1.309 ±0.132	1.143 ±0.050	0.930 ±0.036	2.424 ±0.254	1.624 ±0.148
	Gemini 3.5 Flash	1.052 ±0.051	0.779 ±0.037	1.983 ±0.208	1.173 ±0.126	1.175 ±0.047	0.906 ±0.035	2.465 ±0.229	1.552 ±0.145
	Gemini 3.7 Flash	0.989 ±0.048	0.675 ±0.030	1.761 ±0.200	0.871 ±0.091	1.119 ±0.044	0.818 ±0.029	2.257 ±0.213	1.240 ±0.102
	Claude 4.6 Opus	1.006 ±0.051	0.764 ±0.035	1.867 ±0.217	1.111 ±0.121	1.143 ±0.048	0.894 ±0.035	2.379 ±0.240	1.497 ±0.139
	Claude 4.6 Sonnet	1.020 ±0.049	0.827 ±0.039	1.865 ±0.188	1.309 ±0.137	1.165 ±0.047	0.938 ±0.039	2.418 ±0.235	1.664 ±0.167
	Claude 4.8 Opus	0.994 ±0.053	0.770 ±0.036	1.870 ±0.233	1.133 ±0.124	1.136 ±0.049	0.878 ±0.035	2.388 ±0.264	1.463 ±0.150
	Claude Opus 5	1.016 ±0.049	0.707 ±0.031	1.883 ±0.201	0.932 ±0.094	1.149 ±0.045	0.819 ±0.031	2.354 ±0.223	1.253 ±0.111
	Claude Sonnet 5	1.022 ±0.054	0.812 ±0.040	1.952 ±0.223	1.274 ±0.157	1.162 ±0.051	0.931 ±0.040	2.496 ±0.267	1.660 ±0.186
	Grok 4.1 Fast	1.014 ±0.054	0.829 ±0.042	1.948 ±0.230	1.343 ±0.157	1.172 ±0.051	0.952 ±0.039	2.545 ±0.271	1.693 ±0.161
	Grok 4.20	1.020 ±0.053	0.827 ±0.042	1.965 ±0.234	1.349 ±0.176	1.182 ±0.049	0.950 ±0.040	2.572 ±0.260	1.717 ±0.186
	Grok 4.20 non-reason	1.062 ±0.054	0.873 ±0.042	2.077 ±0.239	1.448 ±0.153	1.206 ±0.050	1.019 ±0.041	2.633 ±0.260	1.918 ±0.177
	TimesFM 3.0	1.035 ±0.055		2.011 ±0.246		1.228 ±0.055		2.782 ±0.294	
	Chronos-2	1.032 ±0.054		1.988 ±0.238		1.216 ±0.053		2.701 ±0.276	
	Base: Hist Mean	1.183 ±0.057		2.372 ±0.260		1.287 ±0.057		2.901 ±0.302	
	Base: Last Value	1.067 ±0.057		2.145 ±0.253		1.284 ±0.059		3.023 ±0.300	
Electricity	*Metrics*	*MAE*		*MSE*		*MAE*		*MSE*	
	Gemini 3.1 Pro	1.057 ±0.096	0.815 ±0.070	1.952 ±0.395	1.167 ±0.183	1.178 ±0.104	0.955 ±0.066	2.544 ±0.486	1.589 ±0.243
	Gemini 3.1 Flash	1.072 ±0.098	1.063 ±0.097	1.997 ±0.397	1.965 ±0.359	1.195 ±0.101	1.246 ±0.103	2.584 ±0.479	2.755 ±0.472
	Gemini 3.5 Flash	0.975 ±0.085	0.891 ±0.078	1.667 ±0.315	1.364 ±0.248	1.236 ±0.102	1.059 ±0.081	2.738 ±0.509	2.024 ±0.350
	Gemini 3.7 Flash	0.957 ±0.087	0.766 ±0.062	1.618 ±0.341	1.017 ±0.170	1.151 ±0.101	0.932 ±0.072	2.415 ±0.465	1.589 ±0.291
	Claude 4.6 Opus	1.017 ±0.092	0.831 ±0.068	1.808 ±0.374	1.177 ±0.175	1.206 ±0.101	1.010 ±0.072	2.622 ±0.505	1.798 ±0.275
	Claude 4.6 Sonnet	1.050 ±0.094	0.901 ±0.076	1.923 ±0.392	1.395 ±0.240	1.227 ±0.098	1.078 ±0.083	2.682 ±0.491	2.034 ±0.341
	Claude 4.8 Opus	1.043 ±0.096	0.969 ±0.089	1.891 ±0.400	1.617 ±0.301	1.168 ±0.099	1.051 ±0.079	2.448 ±0.503	1.914 ±0.285
	Claude Opus 5	0.926 ±0.085	0.753 ±0.066	1.527 ±0.328	1.008 ±0.177	1.089 ±0.098	0.897 ±0.066	2.233 ±0.496	1.461 ±0.245
	Claude Sonnet 5	1.064 ±0.095	0.926 ±0.082	1.965 ±0.371	1.485 ±0.269	1.232 ±0.103	1.039 ±0.072	2.816 ±0.518	1.875 ±0.267
	Grok 4.1 Fast	1.175 ±0.112	1.024 ±0.098	2.450 ±0.472	1.874 ±0.373	1.366 ±0.112	1.174 ±0.092	3.301 ±0.545	2.413 ±0.371
	Grok 4.20	1.080 ±0.100	1.044 ±0.099	2.049 ±0.421	1.931 ±0.376	1.226 ±0.096	1.176 ±0.091	2.635 ±0.456	2.440 ±0.388
	Grok 4.20 non-reason	1.082 ±0.095	1.130 ±0.107	2.011 ±0.373	2.224 ±0.436	1.230 ±0.102	1.380 ±0.117	2.725 ±0.512	3.222 ±0.543
	TimesFM 3.0	1.015 ±0.094		1.795 ±0.377		1.164 ±0.102		2.478 ±0.499	
	Chronos-2	1.014 ±0.093		1.790 ±0.372		1.158 ±0.100		2.441 ±0.474	
	Base: Hist Mean	1.181 ±0.101		2.259 ±0.410		1.301 ±0.110		2.948 ±0.546	
	Base: Last Value	1.120 ±0.106		2.206 ±0.443		1.276 ±0.110		2.908 ±0.533	

Listing 1: Our event and time series dataset taxonomy and examples of our inputs across Stock and Event Forecasting domains.

```
Stock (Large-Cap)
    ...
    AAPL
        Event:
            ...
            [2025-12-12] Oracle Corporation (ORCL) shares plummeted ~10.8%, and
    Broadcom (AVGO) fell ~10% following disappointing earnings/guidance related to
    cloud and AI infrastructure. (sentiment:"Negative",impact_type:"Indirect" )
            ...
        Time Series:
            Price,Close,High,Low,Open,Volume
            Date,,,,,
            2024-02-05,185.924,187.479,184.101,186.389,69668800
            ...
            2026-02-03,270.369,271.579,267.609,269.130,35111059

Stock (Mid-Cap)
    ...
    SMCI
        Events:
            ...
            [2025-03-31] Super Micro Computer (SMCI) appointed Scott Angel, a risk and
    compliance expert, as a new independent board director and promoted Yitai Hu to
    General Counsel and Senior Vice President of Corporate Development
            ...
        Time Series:
            Price,Close,High,Low,Open,Volume
            Ticker,SMCI,SMCI,SMCI,SMCI,SMCI
            Date,,,,,
            2024-04-15,88.27,93.63,88.10,91.19,42011000
            ...
            2026-04-13,25.96,25.98,24.45,25.05,29859900

Event Forecasting
    Political ...
    Tech ...
        which_company_has_the_best_ai_model_end_of_january_2026_01_31:
            ...
            [2025-01-20] DeepSeek (Chinese AI Lab) officially released 'DeepSeek-R1',
    an open-weights reasoning model that achieved a 97.3% score on the MATH-500
    benchmark and 79.8% on AIME 2024.
            ...

    Economics
        fed_decision_in_december_2025_12_10:
            ...
            [2025-09-05] The Bureau of Labor Statistics (BLS) reported August Nonfarm
    Payrolls increased by only 22,000, significantly missing the consensus forecast of
    75,000, while the unemployment rate rose to 4.3%.
            ...
        ...
```

Listing 2: Our event and time series dataset taxonomy and examples of our inputs across Electricity, Weather, and Commodities domains.

```
Electricity
    ...
    Boston
        Event:
            ...
            [2025-12-05] NWS Boston and regional meteorological sources (including
    WBZ-TV and WHDH) reported a severe Arctic cold snap that brought temperatures 20
    to 30 degrees below normal. Morning lows plummeted to slightly below 0 degree F
    north and west of Boston and into the teens in the southeast, marking the coldest
    December 5th morning in nearly ...
            ...
        Time Series:
            ...
            Time,Interval Start,Interval End,Load
            2025-12-01 00:00:00-05:00,2025-12-01 00:00:00-05:00,2025-12-01
    00:05:00-05:00,11213.966
            ...
            2026-03-31 23:55:00-04:00,2026-03-31 23:55:00-04:00,2026-04-01
    00:00:00-04:00,10699.019
Weather
    ...
    Boston
        Event:
            [2025-12-23] The NOAA Climate Prediction Center (CPC) released its 8-14 day
    extended-range forecast (ERF) indicating a macro-scale thermal anomaly
    characterized by strongly above-normal temperatures building across the eastern
    United States. However, local guidance for the Boston area indicates (...)
    potentially overriding the CPC's warm anomaly and driving local temperatures to
    near or below normal ranges alongside mixed precipitation.
            ...
        Time Series:
            time,temp,rhum,prcp,wdir,wspd,wpgt,pres,cldc,coco
            2025-10-01 00:00:00,16.0,80,0.0,20,11.0,,1020.0,2,2
            ....
            2026-02-01 23:00:00,-6.0,45,0.0,310,28.0,15.0,1005.0,8,3
Commodities
    ...
    SI=F
        Event:
            ...
            [2026-01-22] "The U.S. Bureau of Economic Analysis (BEA) was scheduled to
    release the final reading of Q3 2025 GDP (Consensus: 4.3%) and October/November
    PCE inflation data, shifting investor focus from geopolitics to strong U.S.
    economic momentum.", (sentiment:'Negative' , impact_type:Macro/Financial Drivers")
            ...
        Time Series:
            Price,Close,High,Low,Open,Volume
            Ticker,SI=F,SI=F,SI=F,SI=F,SI=F
            2023-10-19,22.886,22.886,22.770,22.770,9
            ...
            2026-03-12,84.794,86.065,84.059,86.065,4712
```

System Prompt For Stock Event Search

You are a Senior Equity Research Analyst specializing in Attribution Analysis (...skip)
To ensure no critical factor is overlooked, you must mentally structure your search across three concentric circles:

1. The Inner Circle (Direct Catalyst): Company-specific events. Include Earnings/Guidance, SEC Filings, M&A, Analyst Upgrades/Downgrades, C-suite changes, Product launches, Legal rulings, or Short Seller reports.
2. The Middle Circle (Indirect/Sector): "Guilty by Association." Did a major competitor report earnings? Is there a sector-wide rotation?
3. The Outer Circle (Indirect/Macro): Broad market drivers. Did the Fed speak? Was there a CPI surprise? Geopolitical shocks?

Analysis & Classification Rules:

- Causal Depth: Explain the *why* via second-order effects, not just surface headlines. Capture all bullish and bearish factors.
- Tagging Constraints:
  - Sentiment: Strictly limit to `Positive` or `Negative`.
  - Impact Type Mapping:
    - If Inner Circle → tag as `Direct`.
    - If Middle/Outer Circle → tag as `Indirect`.

User Prompt For Stock Event Search

Leverage your knowledge about "`{company_name}`", exhaustive search for all factors influencing its stock price on the specific date "`{cut_off_date}`".
Scope of Search:

1. Primary Target: Identify all events first released on '`{cut_off_date}`'.
2. Fallback Protocol: If Direct Catalysts are absent, pivot to identify Indirect Drivers (Middle/Outer Circle sector trends or macro shocks) for that same date.

Content Requirements:
For the "description" field, you must construct a specific summary of the event:

- The Actor: The specific entity (e.g., AMD or Google), person (e.g., CEO name), or institution (e.g., "The Fed") involved.
- The Action/Data: The specific event or number (e.g., "resigned," "raised rates by 25bps," "missed revenue by 5%").
- Avoid vague phrases.

Output the final results strictly in the following JSON format:

```
[
  {
    "date": "YYYY-MM-DD",
    "description": "A specific, fact-based summary explicitly naming the exact actor
and their concrete action, quote, or data point.",
    "causality": "Briefly explaining the transmission mechanism: why and how this
event moved the stock price.",
    "sentiment": "Positive" | "Negative",
    "impact_type": "Direct" | "Indirect"
  },
  ... (List all identified factors)
]
```

Figure 9: The prompt template used for historical stock event retrieval. The system prompt instructs the LLM to act as a Senior Equity Research Analyst and employ a hierarchical search strategy across direct and indirect factors. The user prompt provides specific formatting constraints, ensuring the output is neatly parsed into a JSON format with causal explanations.

System Prompt For Fact Check Agent

You are a Historical Content Verifier & Timestamp Anchor. Your goal is to determine if an event description is factually accurate and publicly known as of a specific Reference Date.

The Verification Logic:
1. Holistic Truth Check (The "Is it Fake?" Filter):

- Use search tools to verify the specific details in the text.
- If the text contains factual errors (e.g., wrong person, event never happened), return False.

2. "Realization Date" Determination (The "Max Date" Rule):

- Identify the chronological timeline mentioned in the text.
- Determine the "Realization Date": the date when the *latest* detail in the text became public knowledge.

3. The Verdict (Truth & Timing):

- The factual output (`True/False`) depends ONLY on Factual Accuracy, regardless of the Reference Date.
- IF Factual: Return True and set the output date to the Realization Date (from Step 2). *(Crucial: Return the actual realization date, even if it is later or earlier than the Reference Date).*
- IF Fake: Return False.

User Prompt For Fact Check Agent

Task: Verify Event Factuality & Timing
Reference Date: "`{date}`"
Event: "`{description}`"

Instructions:

1. Follow the Verification Logic defined above.
2. Provide a brief reasoning before your final verdict.

Output Format:

```
<reasoning>Explain the verification briefly.</reasoning>
<factual>True|False</factual>
<date>YYYY-MM-DD</date>
```

Figure 10: The prompt template used for the Fact Check Agent. The system prompt establishes a rigid verification logic to validate historical accuracy and anchor the true realization date of an event. The user prompt supplies the target event and enforces an XML-style tag format for outputting the reasoning, factual verdict, and exact date.

System Prompt For Weather Time Series Forecasting

Task: You will be provided with a city's recent historical hourly temperatures and a list of specific weather-related events (or N/A). Your task is to forecast the exact future hourly temperatures for a specified prediction horizon.

User Prompt For Weather Time Series Forecasting

Target Location: {city_name}
Forecast Time (Cut-off): {cut_off_time}
Prediction Horizon: Next {prediction_days} days (Hourly Forecast)

A. Historical Hourly Temperatures [Strictly in Celsius (°C)] ({start_time} to {cut_off_time})
{his_temps}

B. Meteorological Event Intelligence
{event_list}

(Note:Temperatures mentioned in the events may be in Fahrenheit. You MUST convert any Fahrenheit values to Celsius °C to match the historical baseline.)

TASK:
Predict the hourly temperatures for the next {prediction_days} days based on the data above. Since you are generating an hourly forecast, you MUST output EXACTLY {total_hours} data points (24 hours * {prediction_days} days).

STRICT CONSTRAINTS:
1. **BRIEF ANALYSIS:** Provivde a concise reasoning to explain your prediciton.
2. **UNIT ALIGNMENT (CRITICAL):** Your final predicted values MUST be in Celsius (°C) to align with the magnitude and distribution of the Historical Hourly Temperatures in Part A.
3. **FORMAT:** Comma-separated values ONLY. No units, no dates, no extra text.
4. **WRAPPER:** Wrap the forecasting numbers inside `<prediction>` tags.

Output:
<prediction>val1, val2, val3, ..., valN</prediction>

Figure 11: The prompt template used for weather time series forecasting. The model is provided with a city's historical hourly temperatures and a list of meteorological events to predict future temperatures over a specified horizon. If no specific weather events are recorded during the target period, the event intelligence section is denoted as "N/A". The prompt also enforces strict output constraints, including temperature unit alignment to Celsius and a specific XML-style output format.

System Prompt For Commodities Price Forecasting

Task: Analyze historical price action and technical structures to predict the short-term closing prices for the next {prediction_days} trading days.

Random Walk: Default to a stable trajectory without extreme directional shifts unless driven by a strong, unpriced signal.
* **Information-Driven Forecasting:** You MUST forecast a dynamic, day-to-day price path. Do NOT lazily repeat the last closing price. Even in the absence of event catalysts, you must rely on historical price structures to make a decisive prediction.

User Prompt For Commodities Price Forecasting

Target Company: {company_name}
Target Analysis Date (Cut-off): {cut_off_date}
Prediction Horizon: Next {prediction_days} trading days

PART 1: DATA INGESTION

A. Historical Price Context (From {start_trading_date} to {cut_off_date}) {his_prices}

B. Event Intelligence & Macro Signals
{event_list}

**Prediction*
* **Format:** Comma-separated values and numbers ONLY, representing the predicted prices for the next {prediction_days} days.
* **Wrapper:** Wrap the final numbers in '<prediction>' tags.

Output:
[Reasoning]
<prediction>...</prediction>

System Prompt For Stock Price Forecasting

Task: Analyze historical price action and technical structures to predict the short-term closing prices for the next {prediction_days} trading days.

* **Information-Driven Forecasting:** You MUST forecast a dynamic, day-to-day price path. Do NOT lazily repeat the last closing price. Even in the absence of event catalysts, you must rely on historical price structures to make a decisive prediction.

User Prompt For Stock Price Forecasting

Target Company: {company_name}
Target Analysis Date (Cut-off): {cut_off_date}
Prediction Horizon: Next {prediction_days} trading days

PART 1: DATA INGESTION
A. Historical Price Context (From {start_trading_date} to {cut_off_date}) {his_prices}

B. Event Intelligence & Macro Signals
{event_list}

Analytical Framework:
1. **Assess Price Context:** Evaluate the recent historical price trend to gauge current market expectations (e.g., is the stock currently over-extended, oversold, or consolidating?).
2. **Identify Strong Signals:** Filter the event list to find clear, material catalysts. Disregard generic PR, pending events with unknown outcomes, or irrelevant macro noise.
3. **Check for Exhaustion:** Evaluate if the market has already digested the news. If the provided price data shows a clear reaction AFTER the event was released, treat the catalyst as "exhausted" and do not predict continued explosive movements based solely on that event.
4. **Random Walk:** Approach predictions conservatively. Markets are mostly efficient. Unless you identify a strong, clear, and unpriced signal, default to a stable outlook, assuming the price will generally maintain its current trajectory without extreme directional shifts.

Analytical Framework:
1. **Assess Trend & Momentum:** Evaluate the historical price data to identify the dominant market regime (e.g., aggressive uptrend, heavy distribution, or lateral consolidation). Gauge the strength and velocity of the current momentum.
2. **Analyze Price Action (K-line Dynamics):** Decode the candlestick behaviors leading up to the cut-off date. Identify signs of trend exhaustion (e.g., long wicks, diminishing body sizes), continuation patterns, or sudden shifts in buying/selling pressure.
3. **Identify Key Levels:** Locate implied support, resistance, or congestion zones based on recent local highs/lows and historical price memory. Assess how the current price is behaving relative to these boundaries.
4. **Project Technical Trajectory:** Synthesize the technical evidence to forecast the most probable price path. If strong momentum or a valid breakout is confirmed, model the directional continuation. If price action signals exhaustion or rejection at a key level, project a structural reversal, mean-reverting pullback, or Random Walk.

* **Format:** Comma-separated values and numbers ONLY, representing the predicted prices for the next {prediction_days} days.
* **Wrapper:** Wrap the final numbers in '<prediction>' tags.

Output:
[Reasoning]
<prediction>...</prediction>

Figure 12: The prompt template used for Commodities and stock price forecasting. For stock price forecasting, to prevent the LLM from inflating its performance by mimicking a naive last-value prediction baseline, the system prohibits lazily repeating the final historical closing price. Furthermore, grounded in the Efficient Market Hypothesis (EMH) in Appendix A.2, The green is injected when event intelligence is present. The blue is utilized when no events are available, guiding the model to rely strictly on technical price action and momentum. Only one of these frameworks is presented to the model per query.

System Prompt For Event Forecasting

You are the **Universal Strategic Forecasting Engine**, (...skip)
Mission: Analyze the provided structured intelligence (Hard Data, Signals, Market Pricing) to calculate the probabilities for a specific set of **Potential Outcomes** defined by the user.

The "Triangulated" Reasoning Protocol:
1. **Determine Domain Weighting (Adaptive Logic):**
* *For Financial/Economic Events (e.g., Earnings, CPI, Rates):* Prioritize **Market Pricing** (Options, Swaps, Prediction Markets) and **Hard Data** over rhetorical Signals. Markets are efficient here.
* *For Political/Geopolitical Events (e.g., Elections, Legislation, Conflict):* Prioritize **Structural Constraints** (Hard Data: Vote counts, Logistics) and **Signals** (Intent) over Market Pricing (which can be slow/reactive in geopolitics).
* *For Tech & AI Events (e.g., Model releases, Benchmarks, Product launches):* Prioritize **Hard Data** (Objective benchmarks, compute allocation, API latency) and **Market/Ecosystem Consensus** (Developer adoption, critical reviews, GitHub metrics) over corporate **Signals** (PR announcements, executive hype).

2. **Synthesize Evidence:**
* **Base Rate:** Establish the baseline probability (Consensus).
* **Update:** Shift the probability based on specific "Impact Type" evidence provided (Positive/Negative indicators).
* **Shock Check:** Do outlier signals (e.g., a surprise war declaration or a sudden CEO resignation) justify allocating probability to "Tail Risk" outcomes?

User Prompt For Event Forecasting

Forecast Task:
Assess the probability of an outcome for the event described below, based strictly on intelligence available as of **{fore_cut_off_date}**.

The Question: {question} (on {result_public_date})
Target Outcome to Predict: "{target_outcome}"

Collected Intelligence (Evidence up to {fore_cut_off_date}):
{list_of_evidences}
Analysis Instructions:
1. **Synthesize:** Review the provided *Hard Data*, *Signals*, and *Market Pricing*.
2. **Weigh:** Evaluate how this evidence shifts the likelihood of the **Target Outcome**.
3. **Calculate:** Estimate the specific probability (0% - 100%) that **"{target_outcome}"** will occur.

Output Format:
Provide a brief reasoning (Chain of Thought) explaining your weighting of the evidence, followed immediately by the final probability tag.

Output:
`[Reasoning]`
`<prediction>[Probability]</prediction>`

Figure 13: The prompt template for event forecasting across 3 domains: Financial/Economic, Political/Geopolitical, and Tech & AI. The model evaluates provided intelligence to estimate the probability of the `{target_outcome}`, which represents the ultimate winning outcome sourced from Polymarket. If no evidence or events are available, the intelligence section is denoted as "N/A".

System Prompt For Stock and Commodities Trend Forecasting (3-class)

You are a highly capable financial AI assistant.
Task:
You will be provided with a company's recent historical stock prices and a list of specific news events (or N/A) affecting the company. Your task is to forecast the stock's cumulative Rate of Change (ROC) over a specified number of future trading days by classifying it into one of three strict categories:

* **"Up"**: `ROC > 0.005` (Cumulative gain over 0.5%)
* **"Neutral"**: `-0.005 <= ROC <= 0.005` (Flat, noise, or tight consolidation)
* **"Down"**: `ROC < -0.005` (Cumulative loss worse than -0.5%)

Do not provide exact price predictions. You must synthesize the historical trends and the impact of the news to make a direct classification.

User Prompt For Stock and Commodities Price Forecasting (3-class)

Target Company: {company_name}
Forecast Date (Cut-off): {cut_off_date}
Prediction Horizon: Next {prediction_days} trading days

A. Historical Prices ({start_trading_date} to {cut_off_date})
{his_prices}

B. Event Intelligence
{event_list}

TASK:
Predict the {prediction_days}-day cumulative Rate of Change (ROC) category based on the data above.

STRICT CONSTRAINTS:
1. **BRIEF ANALYSIS:** You are permitted to provide a concise reasoning (Maximum 2 to 3 sentences) explaining your directional bias and volatility expectations before classifying.
2. **FORMAT:** You must output EXACTLY ONE of the following three categorical labels: "Up", "Neutral", or "Down". NO other words, numbers, or punctuation inside the tags.
3. **WRAPPER:** You MUST wrap the exact label inside '`<prediction>`' tags.

Output:
```
[Brief Analysis]
<prediction>...</prediction>
```

Figure 14: The prompt template used for the 3-class trend forecasting task. Instead of predicting exact price trajectories, the model is instructed to classify the cumulative Rate of Change (ROC) over the prediction horizon into three distinct categories. When no specific news events are offered, the event intelligence section is denoted as "N/A".

System Prompt For Stock and Commodities Trend Forecasting (5-class)

You are a highly capable financial AI assistant.
Task:
You will be provided with a company's recent historical stock prices and a list of specific news events (or N/A) affecting the company. Your task is to forecast the stock's cumulative Rate of Change (ROC) over a specified number of future trading days by classifying it into one of FIVE strict categories:

* **"Strong Up"**: `ROC > 0.02` (Significant gain over 2% driven by major positive structural catalysts)
* **"Up"**: `0.005 < ROC <= 0.02` (Moderate gain between 0.5% and 2% from ordinary positive news or technical momentum)
* **"Neutral"**: `-0.005 <= ROC <= 0.005` (Flat, noise, exhausted news, or tight consolidation)
* **"Down"**: `-0.02 <= ROC < -0.005` (Moderate loss between -0.5% and -2% from ordinary negative news or technical pullback)
* **"Strong Down"**: `ROC < -0.02` (Significant loss worse than -2% driven by major structural risks or severe miss)

User Prompt For Stock and Commodities Price Forecasting (5-class)

Target Company: {company_name}
Forecast Date (Cut-off): {cut_off_date}
Prediction Horizon: Next {prediction_days} trading days

A. Historical Prices ({start_trading_date} to {cut_off_date})
{his_prices}

B. Event Intelligence
{event_list}

TASK:
Predict the {prediction_days}-day cumulative Rate of Change (ROC) category based on the data above.

STRICT CONSTRAINTS:
1. **BRIEF ANALYSIS:** You are permitted to provide a concise reasoning explaining your directional bias and volatility expectations before classifying.
2. **FORMAT:** You must output EXACTLY ONE of the following FIVE categorical labels: "Strong Up", "Up", "Neutral", "Down", or "Strong Down". NO other words, numbers, or punctuation inside the tags.
3. **WRAPPER:** You MUST wrap the exact label inside '`<prediction>`' tags.

Output:
```
[Brief Analysis]
<prediction>...</prediction>
```

Figure 15: The prompt template used for the 5-class trend forecasting task. Instead of predicting exact price trajectories, the model is instructed to classify the cumulative Rate of Change (ROC) over the prediction horizon into three distinct categories. When no specific news events are offered, the event intelligence section is denoted as "N/A".

System Prompt For Weather Trend Forecasting (3-class and 5-class)

Task: Analyze the city's recent historical temperatures and weather-related events (if any) to forecast the trend of the future {temp_type} temperature over the specified horizon.

User Prompt For Weather Trend Forecasting (3-class)

Target Location: {city_name}
Forecast Time: {cut_off_time}
Prediction Horizon: Next {prediction_days} days

A. Historical Hourly Temperatures [Celsius, °C] ({start_time} to {cut_off_time})
{his_temps}

B. Meteorological Event Intelligence
{event_list}

(Note: Temperatures mentioned in the events may be in Fahrenheit °F. You MUST convert any Fahrenheit values to Celsius °C to match the historical baseline.)

TASK:
Forecast the trend of the {temp_type} temperature for the next {prediction_days} days, relative to the previous 24 hours' temperature, based on the data above.

Trend Classification (Change in Celsius °C):
* **Up:** Change > 1.5°C
* **Neutral:** Change between -1.5°C and 1.5°C
* **Down:** Change < -1.5°C

STRICT CONSTRAINTS:
1. **BRIEF ANALYSIS:** Provide a concise reasoning for your forecast.
2. **FORMAT:** Your final prediction MUST be exactly one of these labels: Up, Neutral, or Down. No other text inside the tags.
3. **WRAPPER:** Enclose your final label inside '`<prediction>`' tags.

Output:
```
[Concise Analysis]
<prediction>Trend_Label</prediction>
```

User Prompt For Weather Trend Forecasting (5-class)

'''
(Note: The same with 3-class above)
'''
Trend Classification (Change in Celsius °C):
* **Strong Up:** Change > 3°C
* **Up:** Change > 1.5°C and <= 3°C
* **Neutral:** Change between -1.5°C and 1.5°C
* **Down:** Change >= -3°C and < -1.5°C
* **Strong Down:** Change < -3°C

STRICT CONSTRAINTS:
1. **BRIEF ANALYSIS:** Provide a concise reasoning for your forecast.
2. **FORMAT:** Your final prediction MUST be exactly one of these labels: Strong Up, Up, Neutral, Down, or Strong Down. No other text inside the tags.
3. **WRAPPER:** Enclose your final label inside '`<prediction>`' tags.
'''
(Note: The same with 3-class above)
'''

Figure 16: The prompt templates used for the 3-class and 5-class weather trend forecasting tasks. Both setups share an identical structure, differing only in how the user prompt defines the task. `{temp_type}` specifically denotes the minimum temperature.

User Prompt For Context-aided Event Forecasting (Case Study)

Forecast Task:
Assess the probability of an outcome for the event described below, based strictly on intelligence available as of **2025-10-12**.

The Question: Fed decision in December? (on 2025-12-10 00:00)
Target Outcome to Predict: "25 bps decrease"

Collected Intelligence (Evidence up to 2025-10-12):
(... skip)
83. [2025-09-09] Fisher Investments highlighted a massive preliminary BLS benchmark revision, noting that nonfarm employment for the period ending March 2025 was revised down by -911,000 jobs, effectively erasing a significant portion of previously reported labor strength.
(... skip)
95. [2025-09-16] The U.S. Census Bureau reported August Retail Sales (Control Group) surged 0.7% MoM, significantly beating the consensus forecast of 0.4%.
(... skip)
98. [2025-09-17] The **FOMC** updated its **Summary of Economic Projections (SEP)** to show a median **2025 Real GDP growth forecast** of 1.6% (revised up from 1.4% in June) and an unemployment rate projection of 4.5%.
99. [2025-09-17] The Federal Reserve's Summary of Economic Projections (SEP) lowered the median federal funds rate forecast for end-2025 to 3.6% (from 3.9% in June), while projecting a 2026 median rate of 3.4%.
(... skip)
107. [2025-09-23] Federal Reserve Chair Jerome Powell, at the Greater Providence Chamber of Commerce, reiterated that policy is 'not on a preset course' but emphasized that 'downside risks to employment have increased,' framing the current stance as managing two-sided risks.
(... skip)
109. [2025-09-25] The Bureau of Economic Analysis (BEA) released the Third Estimate for Q2 2025 GDP, revising the annualized growth rate upward to 3.8% from the previous estimate of 3.3%, driven by stronger-than-expected consumer spending.
(... skip)
116. [2025-09-30] Federal Reserve Vice Chair Philip Jefferson, speaking at the Bank of Finland, stated that 'downside risks to employment are rising' and that economic uncertainty is 'especially high' due to new administration policies (tariffs/immigration).
(... skip)
121. [2025-10-01] A partial U.S. federal government shutdown began at midnight after Congress failed to pass a spending bill, halting non-essential operations and delaying the release of subsequent economic data (including the scheduled NFP report).
122. [2025-10-01] Scotia Wealth Management reported that market pricing for total policy easing through the final two Fed meetings of 2025 rose from 44 bps to 48 bps immediately following the negative ADP print.
123. [2025-10-01] The ADP Research Institute reported a net loss of 32,000 private sector jobs for September (vs. consensus +50,000), with the August figure revised downward to -3,000, marking a sudden contraction in hiring.
(... skip)
126. [2025-10-02] The U.S. Department of Labor (DOL) failed to release the weekly Initial Jobless Claims report for the first time in the current cycle, confirming that the federal government shutdown has created a 'data vacuum' for official labor statistics.
(... skip)
133. [2025-10-10] The Bureau of Labor Statistics (BLS) announced that the September CPI report, originally scheduled for Oct 15, would be delayed and rescheduled for release on October 24 due to the government shutdown.

Analysis Instructions:
1. **Synthesize:** Review the provided *Hard Data*, *Signals*, and *Market Pricing*.
2. **Weigh:** Evaluate how this evidence shifts the likelihood of the **Target Outcome**.
3. **Calculate:** Estimate the specific probability (0% - 100%) that **"25 bps decrease"** will occur.

Output Format:
Provide a brief reasoning (Chain of Thought) explaining your weighting of the evidence, followed immediately by the final probability tag.

Output:
`[Reasoning]`
`<prediction>[Probability]</prediction>`

Event Forecasting Response From Gemini-3.1-Pro

Reasoning:
1. **Policy Anchor (SEP):** The most critical piece of evidence is the FOMC's Summary of Economic Projections (SEP) released on September 17, 2025 [98, 99]. The median participant projected the federal funds rate to end 2025 at **3.6%**. With the target range set at 4.00%-4.25% after the September cut, the midpoint is 4.125%. Reaching 3.625% requires exactly **50 basis points** of additional easing. With two meetings remaining in 2025 (November and December), the implied default path is a **25 bps cut at each meeting**.

2. **Market Pricing:** As of October 1, market pricing (swaps/futures) indicated a total of ** 48 basis points** of easing remaining for the year [122]. This pricing aligns almost perfectly with the SEP's implied path of two 25 bps cuts. If the market feared a recessionary crash necessitating 50 bps cuts, or a "higher for longer" pause, this number would deviate significantly from 50 bps. The pricing confirms the consensus expectation is a 25 bps cut in December.

3. **Labor Market Deterioration:** The labor data has deteriorated significantly, shifting the Fed's focus to "downside risks" [107, 116]. The September ADP report showed a contraction of -32,000 jobs [123], and previous payrolls were revised down massively [83]. This weakness ensures the Fed cannot pause. However, the strong Q2 GDP revision (3.8%) [109] and solid Retail Sales [95] suggest the economy is not in freefall, arguing against a panic-sized 50 bps cut, especially given the "risk management" approach.

4. **Operational Constraints (Shutdown):** The federal government shutdown began on October 1, creating a "data vacuum" (no NFP, delayed CPI) [121, 126]. In the absence of official government data to confirm a crash or a resurgence of inflation, the Fed is highly likely to stick to its communicated baseline (the SEP path). A 25 bps cut represents the "autopilot" option that respects the SEP without overreacting to unofficial or noisy signals during a blackout.

5. **Conclusion:** The convergence of the Fed's own projection (SEP), market pricing, and the strategic incentive to remain "gradual" amidst a data blackout makes a 25 bps cut the overwhelming base case.

`<prediction>88%</prediction>`

Figure 17: A case study of event-augmented forecasting for the Federal Reserve's interest rate decision. The response demonstrates the LLM's ability to logically synthesize complex macroeconomic signals—such as SEP projections, labor market deterioration, and constraints from a government shutdown—yielding a highly confident prediction of 88% (The final outcome of decision is "25 bps decrease"). In contrast, predicting without events drops the probability to 33%.

User Prompt for Context-aided Trend Forecasting

Target Company: Broadcom Inc.
Forecast Date (Cut-off): 2025-11-24
Prediction Horizon: Next 3 trading days

A. Historical Prices (2025-11-04 to 2025-11-24)
351.3, 358.3, 354.9, 348.8, 357.7, 351.3, 354.5, 339.3, 341.8, 342.0, 339.8, 353.7, 346.2, 339.6, 377.2

[1] 2025-11-04: Jefferies analyst Blayne Curtis named Broadcom a 'Top Pick' ahead of earnings.
[2]2025-11-04: Palantir Technologies (PLTR) stock plunged 7-9% despite beating earnings estimates, triggering a sector-wide sell-off in AI-linked equities.
[3] 2025-11-04: Wall Street CEOs, including Ted Pick (Morgan Stanley) and David Solomon (Goldman Sachs), issued public warnings regarding high market valuations and a potential imminent equity correction.
(... skip)
[51] 2025-11-17: Federal Reserve Governor Christopher Waller signals support for a December interest rate cut, citing softening labor market data.
[52] 2025-11-17: Federal Reserve Governor Christopher Waller signaled support for a 25-basis-point interest rate cut at the upcoming December meeting.
(... skip)
[76] 2025-11-24: Amazon announced plans to deploy tens of billions of dollars in new capital expenditures dedicated to AI infrastructure.
(... skip)
[78] 2025-11-24: Federal Reserve Governor Christopher Waller signaled support for a December interest rate cut, fueling a broad 'risk-on' rally in the technology sector.
[79] 2025-11-24: The Information reported that Meta Platforms is in advanced talks to spend billions purchasing Google's custom AI chips (Tensor Processing Units or TPUs) for its data centers starting in 2027.
[80] 2025-11-24: Reports emerged that Meta Platforms is in advanced negotiations to rent/buy billions of dollars worth of Google's 'Tensor Processing Units' (TPUs) for its AI data centers.

TASK:
Predict the 3-day cumulative Rate of Change (ROC) category based on the data above.

STRICT CONSTRAINTS:
1. **BRIEF ANALYSIS:** You are permitted to provide a concise reasoning (Maximum 2 to 3 sentences) explaining your directional bias and volatility expectations before classifying.
2. **FORMAT:** You must output EXACTLY ONE of the following three categorical labels: "Up", "Neutral", or "Down". NO other words, numbers, or punctuation inside the tags.
3. **WRAPPER:** You MUST wrap the exact label inside '<prediction>' tags.

Output:
[Brief Analysis]
<prediction>...</prediction>

Trend Forecasting Response From Gemini-3.1-pro

The reported negotiations for Meta to utilize Google's TPUs (manufactured by Broadcom) serve as a significant fundamental catalyst, validating Broadcom's custom silicon dominance and fueling the breakout rally observed on November 24. Supported by renewed "risk-on" sentiment from dovish Federal Reserve commentary and Amazon's massive AI capex commitment, the stock is likely to sustain positive momentum or drift higher despite the recent surge.

<prediction>Up</prediction>

Figure 18: A case study of market trend forecasting for Broadcom Inc. The response highlights the LLM's capacity for multi-hop financial reasoning: it successfully identifies the indirect, yet highly bullish, connection between Meta's purchase of Google TPUs and Broadcom's custom silicon manufacturing. By synthesizing this domain-specific insight with macroeconomic drivers (e.g., Federal Reserve rate cuts and broader AI capex), the model accurately distills a complex news flow into a decisive "Up" prediction. In contrast, inference without events yields the exact opposite prediction.

Time Series Forecasting Response From Gemini-3.1-pro

Target Location: San Francisco
Forecast Time (Cut-off): 2025-12-18 23:00:00
Prediction Horizon: Next 3 days (Hourly Forecast)

A. Historical Hourly Temperatures [Strictly in Celsius (°C)] (2025-12-04 23:00:00 to 2025-12-18 23:00:00)
13.9, 12.9, 11.3, 10.5, 9.7, 8.9, 8.6, 8.1, 8.3, 7.9, 7.7, 6.9, 7.4, 7.3, 6.7, 6.5, 6.8, 6.7, 8.4, 10.0, 11.2, 12.3, 13.0, 13.2, 13.2, 12.4, 10.9, 9.8, 9.5, 9.0, 8.8, 8.9, 9.1, 9.0, 8.5, 8.4, 7.9, 7.7, 7.4, 6.0, 5.9, 6.6, 7.4, 8.8, 10.0, 10.8, 11.4, 11.6, 11.6, 10.8, 9.7, 9.1, 8.5, 8.1, 7.8, 7.9, 7.5, 7.6, 7.5, 7.6, 7.4, 6.5, 6.2, 6.4, 6.4, 6.8, 7.4, 8.9, 10.2, 11.4, 12.2, 12.7, 12.7, 11.8, 10.3, 9.5, 9.0, 8.4, 7.8, 7.4, 7.1, 6.7, 6.5, 6.1, 5.9, 5.8, 5.7, 5.7, 5.6, 6.1, 7.3, 8.6, 10.0, 11.8, 12.6, 12.9, 12.9, 11.9, 10.5, 9.4, 8.5, 8.1, 7.8, 7.4, 7.4, 6.1, 6.2, 5.9, 5.7, 5.8, 5.5, 5.6, 5.5, 6.0, 7.4, 8.8, 10.2, 11.7, 12.5, 12.9, 12.8, 11.9, 10.5, 9.3, 8.5, 8.3, 7.7, 6.8, 6.9, 7.0, 6.9, 6.5, 6.1, 6.3, 6.4, 6.4, 6.1, 7.3, 8.8, 10.4, 12.4, 14.0, 15.0, 15.7, 15.0, 13.8, 11.7, 10.3, 8.7, 8.1, 7.9, 7.4, 6.8, 6.9, 7.2, 6.9, 7.0, 6.8, 6.6, 6.7, 6.3, 7.0, 8.2, 9.9, 11.8, 13.0, 14.2, 14.7, 14.8, 13.7, 10.4, 9.4, 8.7, 8.3, 7.6, 7.3, 6.7, 6.5, 6.4, 6.5, 6.3, 6.1, 5.7, 5.7, 5.8, 6.3, 7.8, 9.6, 11.9, 12.7, 13.5, 13.9, 12.9, 11.9, 10.0, 8.8, 8.0, 7.4, 6.7, 6.3, 6.3, 5.8, 5.5, 5.6, 5.2, 4.9, 4.6, 4.6, 4.7, 5.5, 6.9, 8.7, 10.3, 11.9, 11.8, 12.2, 12.2, 11.5, 10.0, 9.3, 8.4, 8.0, 7.5, 7.4, 7.2, 7.1, 6.9, 7.5, 7.7, 7.7, 7.3, 7.0, 6.5, 6.8, 7.6, 8.7, 8.9, 9.8, 11.0, 11.1, 11.3, 10.7, 9.2, 8.1, 7.5, 7.0, 6.9, 6.4, 6.2, 6.1, 6.1, 6.3, 5.9, 6.2, 6.2, 6.0, 5.7, 6.1, 6.5, 7.9, 8.8, 10.2, 11.1, 12.1, 12.2, 12.0, 10.5, 9.7, 9.4, 8.8, 8.7, 8.5, 8.7, 8.4, 8.1, 8.1, 8.4, 8.9, 9.2, 9.2, 9.5, 10.3, 11.2, 12.1, 12.5, 13.5, 13.5, 13.9, 14.2, 13.4, 13.1, 13.0, 12.9, 12.9, 12.9, 13.2, 13.2, 13.1, 13.2, 13.2, 13.4, 12.8, 12.8, 13.3, 12.9, 13.0, 13.8, 14.2, 14.5, 14.9, 15.3, 15.4, 15.0, 14.1, 12.7, 11.9, 11.3, 10.5, 10.3, 9.6, 9.5, 9.4, 9.2, 9.6, 9.9, 9.7, 9.8, 9.8, 9.8, 9.9, 10.4, 11.2, 11.7, 12.2, 12.8, 13.1, 13.3

B. Meteorological Event Intelligence
(... skip)
[15] 2025-12-08: According to the National Weather Service (NWS)... a 'ridiculously resilient ridge' of high pressure is stationed over Northern California... driving... stable, stagnant air masses. ... Consequently, San Francisco's short-term temperature forecast predicts highs restricted to the mid-to-upper 50s, and overnight lows in the mid-to-upper 40s...
(... skip)
[28] 2025-12-11: According to the NWS San Francisco Area Forecast Discussion... a synoptic pattern shift will terminate a 20-day dry streak... a deepening low-pressure system off the Pacific Northwest coast will drive a strong cold front... forecast to entrain subtropical moisture, generating a moderately strong atmospheric river characterized by heavy precipitation...
(... skip)
[45] 2025-12-15: The Center for Western Weather and Water Extremes (CW3E)... issued an outlook detailing consecutive Atmospheric Rivers (ARs) impacting the Pacific coast... shifting significant moisture transport southward into the San Francisco region. The resulting stratiform cloud cover and precipitation will compress the diurnal temperature range over the 3 to 7 day forecast window.
(... skip)
[48] 2025-12-16: The National Weather Service Area Forecast Discussion outlines a major synoptic pattern shift... initiating robust moisture advection... forecast to deliver 0.5 to 1.0 inches of quantitative precipitation to San Francisco...
(... skip)
[51] 2025-12-17: Meteorological drivers identified by the NWS San Francisco... indicate the approach of a significant atmospheric river, characterized as a 'Pineapple Express' moisture plume... signaling heavy cloud cover and moisture advection that will suppress diurnal temperature variations.
(... skip)
[54] 2025-12-18: The National Weather Service (NWS) San Francisco... issued an official forecast predicting daytime high temperatures strictly in the high 50s to low 60s Fahrenheit... Overnight low temperatures are forecast to hold in the low to mid 50s Fahrenheit...
[55] 2025-12-18: According to the NWS Area Forecast Discussion, the primary meteorological drivers altering the thermal profile are a deep low-pressure system and an incoming 'Pineapple Express' atmospheric river... delivering an exact forecast of 2 to 3 inches of rain to San Francisco...
[56] 2025-12-18: The NWS San Francisco Bay Area office highlighted a deepening low-pressure system... This cyclonic activity will drive strong warm air advection and maintain dense cloud cover, effectively preventing standard nocturnal radiational cooling [3].
[57] 2025-12-18: The NOAA Climate Prediction Center (CPC) published 6-10 and 8-14 Day Outlooks... The CPC explicitly forecasts an above-normal precipitation regime for Northern California... This sustained subtropical moisture transport will lock in extensive cloud cover and elevated dew points over San Francisco (37.7749, -122.4194), severely dampening the diurnal temperature range and keeping ambient temperatures abnormally mild [15].
(... skip)

(Note: Temperatures mentioned in the events may be in Fahrenheit. You MUST convert any Fahrenheit values to Celsius °C to match the historical baseline.)

TASK:
Predict the hourly temperatures for the next 3 days based on the data above.
Since you are generating an hourly forecast, you MUST output EXACTLY 72 data points (24 hours * 3 days).

STRICT CONSTRAINTS:
1. **BRIEF ANALYSIS:** Provivde a concise reasoning to explain your prediciton.
2. **UNIT ALIGNMENT (CRITICAL):** Your final predicted values MUST be in Celsius (°C) to align with the magnitude and distribution of the Historical Hourly Temperatures in Part A.
3. **FORMAT:** Comma-separated values ONLY. No units, no dates, no extra text.
4. **WRAPPER:** Wrap the forecasting numbers inside '`<prediction>`' tags.

Output:
`<prediction>val1, val2, val3, ..., valN</prediction>`

User Prompt for Context-aided Time Series Forecasting

Brief Analysis:
The meteorological intelligence indicates a significant synoptic pattern shift for San Francisco, transitioning from a cold, dry, and stagnant high-pressure ridge to an active, warm, and moist "Pineapple Express" atmospheric river. The heavy cloud cover, high precipitation chances, and continuous warm maritime air advection will drastically reduce nocturnal radiational cooling, raising overnight minimum temperatures from the 40s (°F) into the low-to-mid 50s (°F) (approx. 11°C to 13.5°C). Concurrently, daytime highs will be restricted by the lack of direct solar insolation but buoyed by the warm airmass, peaking in the high 50s to low 60s (°F) (approx. 15°C to 16.5°C). The resulting 72-hour forecast exhibits a progressively dampened diurnal temperature curve with elevated baseline temperatures and a very narrow spread between daily highs and lows.

`<prediction>13.0, 12.7, 12.4, 12.2, 12.0, 11.9, 11.8, 11.7, 11.6, 11.5, 11.4, 11.3, 11.2, 11.2, 11.1, 11.2, 11.4, 11.8, 12.4, 13.1, 13.9, 14.5, 14.9, 15.1, 14.8, 14.4, 14.0, 13.7, 13.5, 13.3, 13.1, 13.0, 12.9, 12.8, 12.7, 12.6, 12.5, 12.4, 12.4, 12.5, 12.7, 13.1, 13.7, 14.4, 15.0, 15.4, 15.7, 15.8, 15.6, 15.3, 14.9, 14.6, 14.4, 14.2, 14.1, 14.0, 13.9, 13.8, 13.7, 13.6, 13.5, 13.5, 13.4, 13.5, 13.7, 14.1, 14.6, 15.1, 15.6, 15.9, 16.1, 16.2</prediction>`

Figure 19: A case study of context-aided meteorological time series forecasting for San Francisco. The LLM successfully deduces that an incoming "Pineapple Express" atmospheric river and the resulting cloud cover will dampen the diurnal temperature range, while accurately converting textual Fahrenheit reports to the required Celsius baseline.

User Prompt For Event Forecasting (Without Fact-check Agent)

Forecast Task:
Assess the probability of an outcome for the event described below, based strictly on intelligence available as of **2025-06-17**.

The Question: First leader out of power in 2025? (on 2025-12-31 12:00)
Target Outcome to Predict: "Ishiba - Japan PM"

Collected Intelligence (Evidence up to 2025-06-17):
(... skip)
[16] 2024-11-11: Japanese Prime Minister Shigeru Ishiba was re-elected in a parliamentary runoff vote (221 votes vs. 160 for Yoshihiko Noda) but was forced to form a minority government requiring case-by-case opposition support to pass the budget.
(... skip)
[63] 2025-04-04: The Constitutional Court of South Korea issues an 8-0 unanimous ruling upholding the impeachment of President Yoon Suk Yeol, formally removing him from office for his role in the December 2024 martial law declaration.
(... skip)
[98] 2025-05-06: The German Bundestag elects Friedrich Merz (CDU) as Chancellor in a second round of voting with 325 votes (surpassing the 316-seat absolute majority), formally ending Olaf Scholz's chancellorship.
(... skip)
[141] **2025-06-13: Polymarket trading data for 'First leader out of power in 2025?' remains active, with Shigeru Ishiba (Japan) eventually settling as the winner (100%) following his resignation in September 2025, despite earlier exits by Trudeau and Scholz in Q1 2025.**
(... skip)
[152] 2025-06-17: The G7 Summit concludes without a joint communiqué due to disagreements over US tariffs, leaving PM Ishiba without a significant diplomatic victory to leverage domestically before the July election.
(... skip)
Analysis Instructions:
1. **Synthesize:** Review the provided *Hard Data*, *Signals*, and *Market Pricing*.
2. **Weigh:** Evaluate how this evidence shifts the likelihood of the **Target Outcome**.
3. **Calculate:** Estimate the specific probability (0% - 100%) that **"Ishiba - Japan PM"** will occur.

Output Format:
Provide a brief reasoning (Chain of Thought) explaining your weighting of the evidence, followed immediately by the final probability tag.

Output:
[Reasoning]
<prediction>[Probability]</prediction>

LLM's Response (Claude 4.5 Opus)

Chain of Thought Analysis:

Understanding the Question:
The question asks which leader will be the "first out of power in 2025" as measured on December 31, 2025. The target outcome is Shigeru Ishiba (Japan PM).

Key Evidence Review:

Leaders Already Out of Power Before Ishiba:
1. **Justin Trudeau (Canada)** - Announced resignation January 6, 2025; formally resigned March 14, 2025 (Items 27, 28, 37, 38, 47)
(... skip)
5. **Dick Schoof (Netherlands)** - Resigned June 3, 2025 (Items 129, 130, 132)

Ishiba's Current Status (as of June 17, 2025):
- Item 141 is critical: "Polymarket trading data for 'First leader out of power in 2025?' remains active, with Shigeru Ishiba (Japan) eventually settling as the winner (100%) following his resignation in September 2025"
- Items 145-150 show Ishiba's approval rating recovered to 37% in June 2025
- Items 147-148 confirm Ishiba attended G7 Summit in Canada (June 15-17, 2025)
- Item 152 notes G7 concluded without a joint communiqué, leaving Ishiba "without a significant diplomatic victory to leverage domestically before the July election"

Critical Interpretation Issue:
The question asks about the "first leader out of power in 2025." Multiple leaders left power BEFORE the knowledge cutoff date (June 17, 2025):
- Trudeau: January 6/March 14, 2025
- Scholz: February 23/March 25/May 6, 2025
- Yoon: April 4, 2025
However, Item 141 explicitly states that Polymarket settled Ishiba as the winner at 100%. This suggests the market's definition of "first leader out" may have specific criteria not met by Trudeau, Scholz, or Yoon (possibly they weren't listed options, or the market had specific parameters).
Domain Weighting (Political Event):
- Prioritize structural constraints (Ishiba leads minority government - Item 16, 18)
- Market pricing shows 100% settlement on Ishiba (Item 141)
- Ishiba's approval at 37% is fragile but stabilizing

Synthesis:
Given Item 141's explicit statement that Polymarket settled Ishiba at 100%, and the structural weakness of his minority government requiring case-by-case opposition support, the evidence strongly supports Ishiba being the designated "first leader out" according to this market's parameters. The fact that other leaders left power earlier suggests they may not have been listed options in this specific prediction market.

The intelligence explicitly states Ishiba resigned in September 2025 and the market settled at 100%.

<prediction>92%</prediction>

Figure 20: Data Leakage due to Missing Fact-check Agents. The LLM's high-confidence prediction is inflated by a temporal paradox in the user prompt. An evidence item timestamped 2025-06-13 explicitly leaked a future outcome ("September 2025 settling 100%"). Without a Fact-check Agent to filter this look-ahead bias, the model effectively "cheated" using future knowledge. The fact-check response refers to Figure 21

Fact-Check Agent System Prompt and User Prompt Template

You are a **Historical Content Verifier & Timestamp Anchor**.
Your goal is to determine if an event description is **factually accurate** and **publicly known** as of a specific **Reference Date**.
The Verification Logic:

1. Holistic Truth Check (The "Is it Fake?" Filter):
* Use search tools to verify the specific details in the text.
* If the text contains factual errors (e.g.,wrong person, event never happened), return **False**.

2. "Realization Date" Determination (The "Max Date" Rule):
* Identify the chronological timeline mentioned in the text.
* Determine the **"Realization Date"**: the date when the *latest* detail in the text became public knowledge.

3. The Verdict (Truth & Timing):
* The factual output (‘True/False‘) depends **ONLY on Factual Accuracy**, regardless of the Reference Date.
* **IF Factual:** Return **True** and set the output date to the **Realization Date** (from Step 2).
(Crucial: Return the actual realization date, even if it is later or earlier than the Reference Date).
* **IF Fake:** Return **False**.

Task: Verify Event Factuality & Timing
Reference Date: "{date}"
Event: "{description}"

Instructions:
1. Follow the **Verification Logic** defined above.
2. Provide a brief **reasoning**before your final verdict.

Output Format:
```
<reasoning>Explain the verification briefly.</reasoning>
<factual>True|False</factual>
<date>YYYY-MM-DD</date>
```

Fact-check Agent Response for Information Leakage

```
<reasoning>
```
The event description is factually accurate based on the available information. Search results confirm that Shigeru Ishiba resigned as Prime Minister of Japan on September 7, 2025, and Polymarket data lists him as the winner (100%) of the "First leader out of power in 2025?" market.[1], [2], [3], [4] The text correctly identifies the "earlier exits" of Justin Trudeau (who announced his resignation in January 2025, Q1) and Olaf Scholz (who was dismissed as Chancellor in March 2025, Q1, though he remained in a caretaker capacity until May). The apparent contradiction of Ishiba winning "First leader out" despite later exits is consistent with the text’s specific mention of the market outcome, implying that the market’s specific rules or candidate list (which often excludes certain leaders or requires specific exit criteria) resulted in Ishiba being the settled winner.
```
</reasoning>
<factual>True</factual>
<date>2025-09-07</date>
```

Figure 21: Prompt template and case study for the Fact-Check Agent. The upper panel details the two-step verification logic, requiring the LLM to determine both factual accuracy and the true “Realization Date” of the provided text. The lower panel shows the agent resolving the temporal leakage from Figure 20. By correctly dating the latest event to 2025-09-07.

System Prompt For Predictability Test

TASK: Evaluate and classify the forecasting difficulty of 40 provided stocks as of the cutoff date: {fore_cut_off_date}.

CONTEXT & DATA:
You are provided with each stock's historical price trends and a timeline of preceding events.
Note: Events may have cross-stock or sector-wide impacts. You must holistically consider these macroeconomic or peer-to-peer correlations.

EVALUATION CRITERIA (Predictability Calibration):
Compare the stocks and assess their relative predictability for a future horizon. Focus strictly on identifying **strong structural shifts or clear fundamental catalysts**.
* High predictability comes from unambiguous, asymmetric signals.
* Low predictability comes from market noise, conflicting events, or absence of structural changes.

Group your final output by the 5 Predictability Levels. Under each level, list the assigned stocks (wrapped in double asterisks "**") and provide a rationale explaining your classification for each.

User Prompt For Stock Predictability Test

DATA CONTEXT:
Here is the specific information and historical context for the evaluated stocks:
{all_forecasting_targets}

DEEP ANALYSIS DIRECTIVE:
Take a deep breath and think step-by-step. Do not rush to conclusions. Engage in deep, second-order thinking to compare the forecasting difficulty across different stocks. Distinguish between temporary market noise and structural shifts before finalizing your predictability levels.

OUTPUT FORMAT: The 5 Predictability Levels:
* **Level 1 (Most Predictable):** Clear, dominant structural shifts or overriding catalysts.
* **Level 2 (Highly Predictable):** Strong directional signals with minor noise.
* **Level 3 (Moderately Predictable):** Mixed signals, typical market conditions, average noise.
* **Level 4 (Low Predictability):** Highly uncertain, conflicting fundamentals, or event-driven volatility.
* **Level 5 (Least Predictable):** Pure market noise, zero clear structural signals, or entirely random walk.

EXAMPLE OUTPUT STRUCTURE:
Level 1 (Most Predictable):
* **Stock**: [Your reasoning explaining why this stock exhibits clear structural shifts compared to others, justifying its Level 1 placement]

Level 2 (Highly Predictable):
* **Stock**: [Your reasoning explaining why this stock exhibits clear structural shifts compared to others, justifying its Level 1 placement]

Level 3 (Moderately Predictable):
* **Stock**: [Your reasoning explaining why this stock exhibits clear structural shifts compared to others, justifying its Level 1 placement]

Level 4 (Low Predictability):
* **Stock**: [Your reasoning explaining why this stock exhibits clear structural shifts compared to others, justifying its Level 1 placement]

Level 5 (Least Predictable):
* **Stock**: [Your reasoning explaining why this stock exhibits clear structural shifts compared to others, justifying its Level 1 placement]

Figure 22: System and user prompts for the Predictability Calibration task. As described in §3.2, the LLM is instructed to evaluate the relative forecasting difficulty of multiple stocks conditioned on their historical price trends and preceding events. The model categorizes the assets into five distinct predictability levels by distinguishing strong structural shifts and catalysts from standard market noise. **Note: For better interpretation, the order of the predictability levels is reversed in §4.4.**