LLM as a Risk Manager: LLM Semantic Filtering for Lead–Lag Trading in Prediction Markets

Sumin Kim[1,*], **Minjae Kim**[1,*], **Jihoon Kwon**[1,*],
Yoon Kim[2], **Oscar Levy**[3], **Alejandro Lopez-Lira**[4], **Yongjae Lee**[5,**], **Chanyeol Choi**[1,**]

[1]LinqAlpha [2]Massachusetts Institute of Technology [3]University of California, Berkeley
[4]University of Florida [5]UNIST

[*]Equal Contribution. [**]Corresponding authors.

Abstract

Prediction markets provide a unique setting where event-level time series are directly tied to natural-language descriptions, yet discovering robust lead–lag relationships remains challenging due to spurious statistical correlations. We propose a hybrid two-stage causal screener to address this challenge: (i) a statistical stage that uses Granger causality to identify candidate leader–follower pairs from market-implied probability time series, and (ii) an LLM-based semantic stage that re-ranks these candidates by assessing whether the proposed direction admits a plausible economic transmission mechanism based on event descriptions. Because causal ground truth is unobserved, we evaluate the ranked pairs using a fixed, signal-triggered trading protocol that maps relationship quality into realized profit and loss (PnL). On Kalshi Economics markets, our hybrid approach consistently outperforms the statistical baseline. Across rolling evaluations, the win rate increases from 51.4% to 54.5%. Crucially, the average magnitude of losing trades decreases substantially from \$649 to \$347. This reduction is driven by the LLM's ability to filter out statistically fragile links that are prone to large losses, rather than relying on rare gains. These improvements remain stable across different trading configurations, indicating that the gains are not driven by specific parameter choices. Overall, the results suggest that LLMs function as *semantic risk managers* on top of statistical discovery, prioritizing lead–lag relationships that generalize under changing market conditions.

1 Introduction

Real-world events rarely occur in isolation; information about one event often changes expectations about what happens next. For example, an inflation surprise can raise expectations of tighter monetary policy. These dependencies often manifest as *lead–lag* structure, where movement in one event systematically precedes movement in another (Bennett et al., 2022).

Identifying such structure from observational time series remains challenging in practice. Traditional statistical methods such as Granger causality can detect directional predictability, but the resulting relationships are often unstable over time and may not reflect meaningful underlying mechanisms (Rossi and Wang, 2019; Phillips, 1986). In large-scale settings, statistically significant lead–lag links frequently fail to generalize out of sample and can induce substantial losses when used in downstream decision-making.

Meanwhile, large language models (LLMs) have shown strong performance in time-series forecasting tasks, including predicting future trajectories from noisy observations (Gruver et al., 2023; Jin et al., 2024). However, these approaches typically focus on individual time series and do not address whether relationships *between* events are meaningful or robust. This raises a distinct question: *can LLMs help distinguish mechanistically plausible inter-event relationships from brittle statistical correlations?*

Our key idea is to use LLMs as *semantic filters* that evaluate the plausibility of statistically identified lead–lag relationships. We study this question in a prediction-market setting, where event-level time series continue to update over time and relationships can be evaluated on future, unseen periods under changing market conditions.

We implement a two-stage framework. First, we apply Granger causality to identify candidate leader–follower event pairs from market-implied probability time series (Statistical Approach). Second, we use an LLM to re-rank these candidates based solely on the events' natural-language descriptions, prioritizing relationships that admit a plausible real-world mechanism (Hybrid Approach).

Across 18 rolling evaluations on Kalshi Economics markets, LLM-based semantic filtering consistently improves trading performance relative to Granger screening alone. At a 7-day holding horizon, total PnL more than doubles, driven mainly by a pronounced reduction in downside risk. These results suggest that LLMs act as robustness filters on top of statistical discovery, helping to exclude fragile lead–lag relationships that fail under changing market conditions.

2 Related Work

Limits of statistical significance in lead–lag discovery. Lead–lag patterns, in which movements in one time series appear to precede those in another, are widely documented in economics and finance (Lo and MacKinlay, 1990; Hasbrouck, 1995). Motivated by these empirical regularities, a large body of work seeks to formalize directional relationships using statistical discovery tools such as Granger causality and related predictive tests (Granger, 1969; Kinnear and Mazumdar, 2023; Chaudhry et al., 2017).

However, even when statistically significant, many discovered lead–lag relationships remain fundamentally ambiguous and limited. In large-scale settings, pairwise screening inevitably produces links that survive selection purely by chance, a concern long emphasized in the data-snooping and multiple-testing literature (Sullivan et al., 1999). These issues are further exacerbated when relationships evolve over time or exhibit structural breaks, leading to instability in Granger-causal links across samples (Ventosa-Santaulària and Vera-Valdés, 2008; Rossi and Wang, 2019; Phillips, 1986). As a result, statistical significance alone provides limited guidance for distinguishing mechanistically meaningful inter-event relationships from brittle correlations.

To assess robustness, discovered relationships are therefore often evaluated using out-of-sample predictive performance or rolling evaluation schemes (Diebold and Mariano, 2002; Rossi, 2021). Yet even when validated out of sample, such evaluations provide limited guidance about whether an identified lead–lag direction corresponds to a coherent real-world transmission mechanism.

Semantic evaluation with LLMs and event-based settings. Recent work has explored the use of large language models (LLMs) beyond traditional language tasks, including evaluating explanations, relationships, and counterfactuals expressed in natural language (Zheng et al., 2023; Yang et al., 2024; Wang, 2024). In these settings, LLMs are not used to infer causal structure from data, but rather to assess the coherence or plausibility of candidate relationships. This perspective is particularly relevant for lead–lag analysis, where temporal ordering alone is insufficient to establish whether a relationship admits a meaningful real-world transmission mechanism.

Prediction markets provide a natural platform for this type of evaluation. Each contract corresponds to a clearly defined event described in natural language and yields a time series of market-implied probabilities reflecting collective expectations (Wolfers and Zitzewitz, 2004; Berg et al., 2008). This structure enables the study of inter-event relationships while allowing discovered dependencies to be evaluated on future, unseen periods under changing market conditions.

3 Background

This section introduces the key concepts and notation underlying our analysis of directional relationships between event-level time series. Our objective is to study directional lead–lag relationships between event-level time series and establish a framework for evaluating their predictive relevance.

We formally define directional lead–lag relationships in Section. 3.1. Section 3.2 describes preprocessing procedure of time-series data, and Section 3.3 introduces Granger causality as a statistical tool for operationalizing directional predictability. Finally, Section 3.4 defines how to annotate discovered lead–lag relationships.

3.1 Directional lead–lag relationships

Let $\{x_t\}_{t=1}^T$ and $\{y_t\}_{t=1}^T$ be two real-valued time series indexed by time t. We say that x *leads* y if past values of x_t contain predictive information about future values of y_t beyond what is contained in the past of y_t itself.

Formally, x leads y if there exists a forecast horizon $h > 0$ such that

$$\mathbb{E}[y_{t+h} \mid x_{1:t}, y_{1:t}] \neq \mathbb{E}[y_{t+h} \mid y_{1:t}] \,.$$

This definition is directional and asymmetric: x leading y does not imply that y leads x. Throughout this paper, we study ordered event pairs exhibiting directional lead–lag structure—also referred to as leader–follower pairs (L, F).

3.2 Time-series representation of prediction-market prices

Prediction markets trade contracts whose payoff depends on the realization of a future event. For each event i, the daily YES price forms an event-level time series that reflects the market's collective belief about the probability of that event occurring.

Let $p_{i,t} \in [0, 100]$ denote the daily YES price (in percentage points) of event i at time t. To obtain an unbounded real-valued signal suitable for time-series analysis, we apply the log-odds transformation

$$\ell_{i,t} = \log\left(\frac{p_{i,t}}{100 - p_{i,t}}\right), \quad (1)$$

where $\ell_{i,t} \in \mathbb{R}$.

This transformation mitigates boundary effects near 0 and 100 and renders price changes approximately additive, which facilitates statistical modeling.

3.3 Directional predictability via Granger causality

To test lead–lag relationships, we require a statistical notion of directional predictability. We use Granger causality, which assesses whether incorporating the lagged history of one series improves prediction of another beyond what is explained by the target series' own past.

Let $\{x_t\}_{t=1}^{T}$ and $\{y_t\}_{t=1}^{T}$ be two time series, and let p denote the chosen lag length in the autoregressive model. Granger causality tests are typically formulated under the assumption that the input series are approximately stationary. In practice, stationarity is commonly assessed using unit-root tests such as the Augmented Dickey–Fuller (ADF) test, which evaluates whether a time series exhibits a unit root (i.e., behaves like a random walk). When non-stationarity is detected, differencing is a common remedy (Dickey and Fuller, 1979; Said and Dickey, 1984; Hamilton, 2020).

A standard approach to testing Granger causality is to use vector autoregressive (VAR) models, which capture linear dependencies between multiple time series through their lagged values (Hamilton, 2020). In the bivariate case, y_t can be modeled with optional lags of x_t as

$$y_t = \alpha_0 + \sum_{k=1}^{p} \alpha_k \, y_{t-k} + \sum_{k=1}^{p} \beta_k \, x_{t-k} + \varepsilon_t, \quad (2)$$

where ε_t is a mean-zero error term.

The Granger test evaluates the null hypothesis $H_0 : \beta_1 = \cdots = \beta_p = 0$, corresponding to the absence of directional predictability from x to y under lag length p. Rejection of H_0 indicates a lead–lag relationship from x to y. Thus, Granger causality provides a testable proxy for the theoretical lead–lag definition in Section 3.1.

3.4 Sign of co-movement

Beyond temporal direction, lead–lag relationships can be further characterized by whether the two series tend to move in the same or opposite directions. For an ordered lead–lag (leader–follower) pair $(L \rightarrow F)$ with corresponding time series $(x_{L,t}, x_{F,t})$, we define the sign of co-movement as

$$s = \operatorname{sgn}(\operatorname{corr}(x_{L,t}, x_{F,t})) \in \{-1, +1\}, \quad (3)$$

where $\operatorname{corr}(\cdot,\cdot)$ denotes the Pearson correlation coefficient. Here, $s = +1$ indicates aligned movement and $s = -1$ indicates opposing movement, conditional on the identified lead–lag ordering. This quantity is purely descriptive and does not imply any causal or structural interpretation. It indicates whether increases in the leading series tend to be associated with increases or decreases in the lagging series.

4 Methods

Recall that our study aims to assess whether LLMs can distinguish causal relationships that are mechanistically meaningful. To assess this question, we compare two methods for ranking candidate causal relationships: a purely *statistical* approach, and a *hybrid* approach that applies LLM-based reranking on top of the statistical approach. We evaluate each ranking using the same trading protocol, where investment performance reflects the reliability of the identified causal relationships.

In this section, we introduce the two ranking methods: *statistical* and *hybrid* (Section. 4.1), and describe the trading protocol (Section. 4.2).

4.1 Ranking Methods for Candidate Causal Relationships

Figure 1 summarizes the two-stage causal filtering framework used in this study. Starting from event-level prediction market price time series, the pipeline consists of (i) a statistical filtering stage

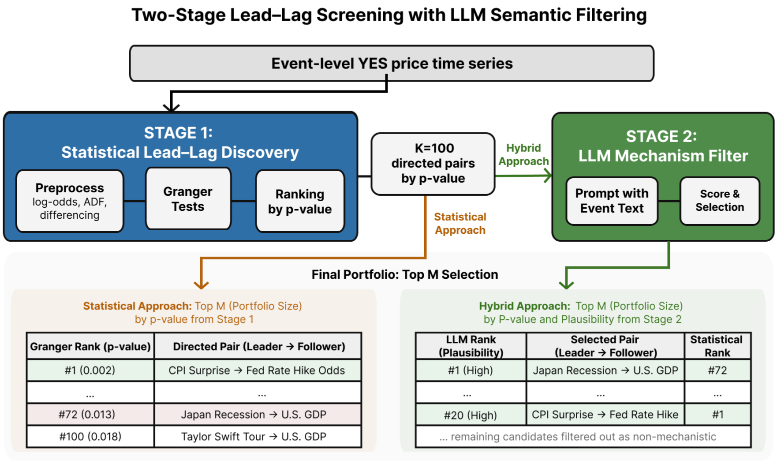

Figure 1: Two-stage framework for leader–follower pair discovery in prediction markets. Stage 1 produces a candidate set of Top K directed pairs (K=100) ranked by Granger significance, and Stage 2 applies LLM-based semantic re-ranking to select the final Top M portfolio (M=20).

that identifies candidate leader–follower relationships based on time-series evidence, and (ii) a semantic filtering stage that re-ranks these candidates using an LLM. The two stages correspond to the *Statistical Approach* and *Hybrid Approach* described below, and both produce ranked lists of directed event pairs that are evaluated using the same trading protocol.

For each event i, we obtain a daily price time series $\mathcal{P}_i = \{p_{i,t}\}_{t=1}^{T}$, where $p_{i,t}$ denotes the market-implied price at time t and T is the length of the observation window. From $\mathcal{P}_i$, we construct a stationary market signal $\mathcal{X}_i = \{x_{i,t}\}_{t=1}^{T}$ by transforming prices into log-odds. Standard stationarity preprocessing is applied prior to Granger testing.

Statistical Approach. Given a pair of events (i, j), the statistical approach aims to determine whether there exists a directional lead–lag relationship between the two events based *purely on time-series evidence*. To this end, we apply Granger causality tests to the stationary market series constructed from their daily price data.

For each unordered event pair $\{i, j\}$, we evaluate both possible causal directions, $(i \to j)$ and $(j \to i)$. In each direction, we estimate vector autoregressive (VAR) models over multiple lag lengths and assess whether the lagged history of the candidate leader provides incremental predictive power for the follower beyond its own past.

We retain the direction exhibiting stronger statistical evidence of directional predictability as the candidate causal relationship for the event pair. Finally, all candidate relationships are ranked by their statistical strength, yielding an ordered list of directed leader–follower pairs from strongest to weakest evidence of Granger-based predictability.

Hybrid Approach. The hybrid approach takes as its starting point the set of candidate causal relationships identified by the statistical approach. The role of the LLM is not to discover new relationships, but to re-rank these statistically validated candidates based on semantic and mechanism-level plausibility. In doing so, the approach preserves statistical validity while adjusting priorities to favor relationships that have coherent economic interpretations.

Specifically, for each directed event pair retained by the statistical approach, we prompt a LLM with the event titles and descriptions and ask whether a plausible economic mechanism exists by which movement in the leader could precede movement in the follower. The model evaluates the semantic coherence of the proposed causal direction and assigns a plausibility score, with higher scores indicating stronger mechanism-level support.

We then use these plausibility scores to re-rank the statistically screened candidate relationships. The final hybrid ranking thus preserves the statistical validity ensured by Granger causality while prioritizing relationships that are also supported by coherent economic reasoning.

4.2 Trading-Based Evaluation of Ranking Methods

Since the ground truth of causal relationships is not directly observable, evaluating the quality of

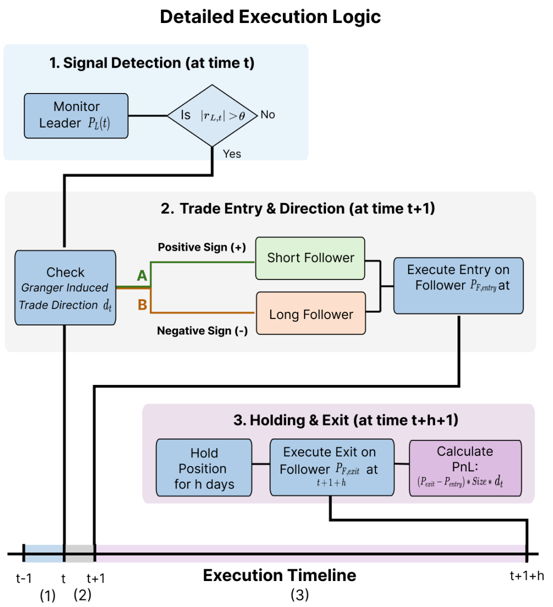

Figure 2: Signal-triggered trading protocol used to evaluate ranked lead-lag relationships from Figure 1: leader price moves trigger follower trades, with direction determined by the Granger-induced trade sign and out-of-sample PnL used to evaluate the ranked pair list.

different relationship rankings is inherently challenging. To address this issue, we evaluate the ranking approaches based on trading performance. The trading rule is intentionally designed to be simple and signal-driven: when a sufficiently large price movement occurs in a leader event, the strategy mechanically follows the corresponding follower. Therefore, differences in realized trading performance can be attributed to differences in the quality of the underlying causal rankings. Figure 2 illustrates the resulting signal-based trade lifecycle, which we detail step by step below.

Step 1: leader-based trigger Consider a directed event pair (L, F), where L denotes the leader event and F the corresponding follower as identified by a causal ranking. At each time t, we monitor the daily price movement of the leader event and trigger a trading signal when the magnitude of the price change exceeds a predefined threshold. Formally, let $p_{L,t}$ denote the market-implied price (in percent) of the leader event at time t. We define the one-day relative price change as $r_{L,t} \triangleq (p_{L,t} - p_{L,t-1})/p_{L,t-1}$. A trading signal is generated if $|r_{L,t}| > \theta$, where θ is a fixed threshold in decimal form.

Step 2: Follower Entry. When a trading signal is generated for a leader–follower pair (L, F) at time t, we enter a position in the follower event at time $t + 1$. We define the *Granger-induced trade direction* as

$$d_t = \mathrm{sign}(r_{L,t}) \times s(L \to F),$$

where $s(L \to F)$ is the sign of the identified Granger relationship. If $d_t = +1$, we take exposure to an increase in the follower event probability (buy YES shares), and if $d_t = -1$, we take exposure to a decrease in the follower event probability (sell YES shares). This construction follows directly from the lead–lag interpretation of the Granger sign described in Section 3.1.

Step 3: Holding Period and Exit. Once entered, the follower position is held for a fixed horizon of h days and is exited mechanically at time $t + h + 1$. The realized profit or loss is computed based on the change in the follower's market-implied price over the holding period.

5 Experiments

5.1 Experimental Setup

Data and windows. We use Kalshi prediction-market data in the Economics category spanning October 2021 to November 2025. After filtering markets with insufficient activity or negligible price variation, we retain 554 event markets. Evaluation follows a rolling-window protocol with 60-day training windows and 30-day testing windows, yielding 18 non-overlapping test periods.

Pipeline and hyperparameters. Within each training window, we perform ADF unit-root testing and apply first differencing when necessary prior to Granger analysis. We then run pairwise Granger causality tests across all available markets. For each event pair, we evaluate both causal directions and sweep lag lengths $L \in \{1, 2, 3, 4, 5\}$, retaining the strongest statistical evidence of directional predictability. Because these optimal lag orders are short in market-implied probability dynamics, we treat the trading holding horizon h as an independent evaluation parameter and report robustness across multiple horizons. From these tests, we retain the top $K = 100$ directed pairs ranked by statistical significance. From this candidate set, we form a portfolio by selecting the top $M = 20$ directed pairs under two ranking conditions: the *statistical* approach and the *hybrid* approach (see Section 4.1). For the semantic re-ranking stage, we use `GPT-5-nano`; robustness to an alternative model

Table 1: Trading-based evaluation summary. We compare portfolios constructed under the *statistical approach* and the *hybrid approach* across rolling test windows. Metrics include trade count, win rate, average win/loss per trade, and total PnL; performance gains are driven primarily by reductions in average loss magnitude.

Metric	Statistical Approach	Hybrid Approach	Change
Win Rate	51.4%	54.5%	+3.1pp
Avg Win	$724	$636	-12%
Avg Loss	-$649	-$347	+46.5%
Total PnL	**$4,100**	**$12,500**	**+205%**

Table 2: Comparison of average loss per trade under same-event and different-event settings. Same-event pairs share the same underlying event with different formulations, whereas different-event pairs correspond to distinct events. Lower average loss indicates better downside control.

Method	Same-Event	Different-Event
Avg. Loss of Statistical Approach	-$700	-$642
Avg. Loss of Hybrid Approach	-$400	-$333
Loss Reduction	42.9%	48.1%

variant (`GPT-5-mini`) is reported in Appendix A.4. Each trade is executed with position size of 100 per contract. Unless otherwise specified, we report results under a default configuration with holding horizon $h = 7$ days, trigger threshold $\theta = 0$ (i.e., $|r_{L,t}| > 0$), and portfolio size $M = 20$. To assess robustness, we additionally vary the holding horizon over $h \in \{1, 3, 5, 7, 10, 14, 21\}$ (Appendix A.1).

5.2 Results

LLM filtering yields substantial reductions in downside risk. Table 1 reports the primary comparison between the statistical and hybrid approaches under the default evaluation setting. LLM-based re-ranking leads to a substantial improvement in total PnL (+205%) accompanied by a modest increase in win rate. This pattern indicates that the gains are not driven by a small number of outlier wins or aggressive upside bets. Instead, the dominant driver of the performance improvement is a large reduction in average loss magnitude (46.5%), highlighting enhanced downside control as the primary mechanism.

Loss reduction persists for distinct event pairs. To examine whether the observed loss reduction is driven by trivial semantic overlap, we analyze performance composition under the default setting.

Table 3: Comparison between proposed Hybrid method and a random-direction baseline. The random baseline trades the same Granger-surviving pairs as the Statistical and Hybrid approaches but randomizes the predicted sign, isolating the contribution of directional information from pair selection. Default setting: K=100, M=20, h=7d, θ=0.

Metric	Statistical	Random-Dir.	Hybrid
Avg Win	$724	$610	$636
Avg Loss	-$649	-$495	-$347
Total PnL	**$4,100**	**$6,500**	**$12,500**

Table 4: Comparison between proposed Hybrid method and a semantic-similarity baseline. The semantic-similarity baseline selects the same Granger-surviving pairs but re-ranks them by cosine similarity of event-title embeddings instead of LLM plausibility, retaining the Granger-estimated sign as the trading direction. *Change* is the relative difference of the Semantic column from the Hybrid column. Default setting: K=100, M=20, h=7d, θ=0.

Metric	Hybrid	Semantic	Change
Avg Win	$636	$512	-19.5%
Avg Loss	-$347	-$628	-81.0%
Total PnL	**$12,500**	**-$8,100**	-165%

As shown in Table 2, LLM filtering reduces average loss magnitude by 48.1% for different-event pairs (from $642 to $333), comparable to the reduction observed for same-event pairs (42.9%). This indicates that the gains are not attributable to trivial event overlap, but persist for genuinely distinct event relationships.

Gains exceed a direction-randomized baseline. To verify that the hybrid approach's advantage is not attributable to pair selection alone, we compare against a random-direction baseline that trades the same Granger-surviving pairs but randomizes the predicted sign. As shown in Table 3, the hybrid approach outperforms random-direction trading in 26 of 42 configurations (61.9%), with total PnL of $12,500 versus $6,500. This indicates that the LLM contributes directional information beyond filtering which pairs to trade.

LLM gains are not explained by surface semantic similarity. To examine whether the hybrid approach's advantage stems from textual similarity between event descriptions rather than mechanism-level reasoning, we compare against a semantic-similarity baseline that re-ranks the same Granger-surviving pairs by cosine similar-

Table 5: Win rate (WR) by the magnitude of leader movement, indicating which side outperforms in investment outcomes. Here, the magnitude of leader movement is defined as the amount of change in the leader event's price p at a given time. Win rates (WR) are reported for Statistical Apporach (Statistical WR) and Hybrid Approach (Hybrid WR) portfolios; Δ denotes the win-rate difference in percentage points.

Magnitude of leader movement (pt)	Statistical WR	Hybrid WR	Winner	Δ
5–10 pt	57.1%	66.7%	LLM	**+9.5pp**
10+ pt	53.8%	71.4%	LLM	**+17.6pp**

ity of event-title embeddings computed by the `text-embedding-3-small` model from OpenAI. As shown in Table 4, semantic ranking yields total PnL of $-\$8{,}100$, far below the hybrid approach (\$12,500), with average loss magnitude ($-\$628$) close to the statistical baseline and well above the hybrid's ($-\$347$). This indicates that the LLM's downside-risk reduction cannot be reduced to description-level similarity, but reflects genuine mechanism-level filtering.

Semantic filtering is most valuable during large leader moves. We next examine performance by the magnitude of leader price changes. As shown in Table 5, semantic re-ranking delivers larger gains when leader markets experience larger discrete repricings. For 5–10pt moves, win rate increases from 57.1% to 66.7% (+9.5pp), and for moves exceeding 10pt, it rises from 53.8% to 71.4% (+17.6pp). This pattern suggests that LLM filtering is particularly effective at mitigating regime-sensitive failures that arise during large market moves. Beyond these aggregate patterns, the hybrid ranking also elevates economically coherent relationships that rank low under Granger screening; representative examples are provided in Appendix A.3.

6 Conclusion

We introduced a dual-stage causal filtering framework for prediction markets, combining statistical lead–lag screening via Granger causality with LLM-based semantic verification. Across rolling-window evaluations, LLM re-ranking consistently improves trading performance, with gains driven primarily by substantial reductions in average loss magnitude (downside risk) and accompanied by modest improvements in win rate. Taken together, these results suggest that LLMs can act as semantic risk managers on top of statistical discovery, prioritizing relationships that generalize more reliably under changing market conditions.

Limitations

Our experiments focus on Kalshi Economics markets, where each contract is paired with a natural-language description and a daily probability series; thus, the benefits of LLM-based semantic re-ranking may not generalize to other market types or platforms. Because causal ground truth is unobserved, we evaluate rankings indirectly via an offline rolling-window backtest under a fixed rule-based trading protocol. Although we prevent lookahead bias by restricting decisions to information available at each time step, the strategy has not been tested in a live trading environment. In addition, the semantic filtering stage relies on an LLM whose plausibility judgments may be sensitive to prompting and model versions.

References

Stefanos Bennett, Mihai Cucuringu, and Gesine Reinert. 2022. Lead–lag detection and network clustering for multivariate time series with an application to the us equity market. *Machine Learning*, 111:4497–4538.

Joyce E Berg, Forrest D Nelson, and Thomas A Rietz. 2008. Prediction market accuracy in the long run. *International Journal of Forecasting*, 24(2):285–300.

Aditya Chaudhry, Pan Xu, and Quanquan Gu. 2017. Uncertainty assessment and false discovery rate control in high-dimensional granger causal inference. In *International Conference on Machine Learning*, pages 684–693. PMLR.

David A. Dickey and Wayne A. Fuller. 1979. Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association*, 74(366):427–431.

Francis X Diebold and Robert S Mariano. 2002. Comparing predictive accuracy. *Journal of Business & economic statistics*, 20(1):134–144.

Clive WJ Granger. 1969. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: journal of the Econometric Society*, pages 424–438.

Nate Gruver, Marc Finzi, Shikai Qiu, and Andrew G Wilson. 2023. Large language models are zero-shot time series forecasters. *Advances in Neural Information Processing Systems*, 36:19622–19635.

James D Hamilton. 2020. *Time series analysis*. Princeton university press.

Joel Hasbrouck. 1995. One security, many markets: Determining the contributions to price discovery. *The journal of Finance*, 50(4):1175–1199.

Ming Jin, Shiyu Wang, Lintao Ma, Zhixuan Chu, James Zhang, Xiaoming Shi, Pin-Yu Chen, Yuxuan Liang, Yuan-Fang Li, Shirui Pan, and Qingsong Wen. 2024. Time-llm: Time series forecasting by reprogramming large language models. In *International Conference on Learning Representations*, volume 2024, pages 23857–23880.

Ryan J Kinnear and Ravi R Mazumdar. 2023. Exact recovery of granger causality graphs with unconditional pairwise tests. *Network Science*, 11(3):431–457.

Andrew W Lo and A Craig MacKinlay. 1990. An econometric analysis of nonsynchronous trading. *Journal of Econometrics*, 45(1-2):181–211.

OpenAI. 2026. Compare models. `https://platform.openai.com/docs/models/compare`. Last accessed February 2, 2026.

Peter CB Phillips. 1986. Understanding spurious regressions in econometrics. *Journal of econometrics*, 33(3):311–340.

Barbara Rossi. 2021. Forecasting in the presence of instabilities: How we know whether models predict well and how to improve them. *Journal of Economic Literature*, 59(4):1135–1190.

Barbara Rossi and Yu Wang. 2019. Vector autoregressive-based granger causality tests in the presence of instabilities. *The Stata Journal*, 19(4):883–899.

Said E Said and David A Dickey. 1984. Testing for unit roots in autoregressive-moving average models of unknown order. *Biometrika*, 71(3):599–607.

Ryan Sullivan, Allan Timmermann, and Halbert White. 1999. Data-snooping, technical trading rule performance, and the bootstrap. *The journal of Finance*, 54(5):1647–1691.

Daniel Ventosa-Santaulària and J Eduardo Vera-Valdés. 2008. Granger-causality in the presence of structural breaks. *Economics Bulletin*, 3(61).

Zeyu Wang. 2024. Causalbench: A comprehensive benchmark for evaluating causal reasoning capabilities of large language models. In *Proceedings of the 10th SIGHAN Workshop on Chinese Language Processing (SIGHAN-10)*, pages 143–151.

Justin Wolfers and Eric Zitzewitz. 2004. Prediction markets. *Journal of Economic Perspectives*, 18(2):107–126.

Linying Yang, Vik Shirvaikar, Oscar Clivio, and Fabian Falck. 2024. A critical review of causal reasoning benchmarks for large language models. arXiv. *Preprint*, arXiv:2407.08029.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

A Appendix

A.1 Holding-period ablation

We examine whether the loss-reduction effect persists across alternative holding horizons $h \in \{1, 3, 5, 7, 10, 14, 21\}$ under otherwise identical settings.

Table 6: Hold period ablation (trades with $|\Delta p_{L,t}| > 0$ only). We report win rate and average loss for Statistical Approach versus Hybrid Approach portfolios as the holding horizon varies; Loss Reduction is the relative decrease in average loss magnitude under Hybrid filtering.

Hold	Stat. WR	Hybrid WR	Stat. Avg Loss	Hybrid Avg Loss	Loss Reduction
1d	56.0%	66.7%	-$536	-$283	47.2%
3d	44.7%	62.8%	-$773	-$469	39.4%
5d	47.5%	66.0%	-$745	-$500	32.9%
7d	**51.4%**	**54.5%**	**-$649**	**-$347**	**46.5%**
10d	51.4%	56.9%	-$497	-$371	25.3%
14d	49.5%	56.1%	-$661	-$400	39.5%
21d	47.3%	53.1%	-$973	-$753	22.6%
Mean	∼50%	∼59%	-$689	-$446	36.2%

A.2 Threshold sensitivity

We examine whether the loss-reduction effect persists when the trigger threshold θ is tightened so that trades are only entered after larger leader moves.

Table 7: Performance across trigger thresholds θ at h=7d, M=20. The statistical baseline degrades sharply as θ tightens, while the hybrid approach remains stable.

θ	Statistical	Hybrid	Advantage
0.01	$2,400	$3,900	+$1,500
0.05	$4,100	$12,500	+$8,400
0.10	-$2,600	$10,000	+$12,600

As small movements are filtered out, the statistical baseline degrades from $4,100 at θ=0.05 to $-$2,600 at θ=0.10, consistent with spurious statistical correlations failing under stricter conditions. The hybrid approach remains in the $10,000–$12,500 range across the same sweep, indicating that semantically grounded pairs generalize more reliably when triggers are restricted to economically meaningful moves. The LLM advantage therefore *increases* rather than decreases as microstructure noise is suppressed.

A.3 Representative pairs elevated by LLM re-ranking

We provide illustrative examples of LLM-selected pairs that fall outside the Granger Top-M ranking yet yield positive realized outcomes.

Table 8: Representative pairs ranked low by the Statistical Approach but elevated by the Hybrid Approach. SR = Statistical Rank (Granger p-value); HR = Hybrid Rank (after LLM re-ranking).

Leader	Follower	SR	HR	PnL
China 2022 GDP Growth >5%	World 2022 GDP Growth >3%	#23	#17	+$1,100
Japan 2026 Recession	US Q1 2025 GDP Growth >2%	#71	#5	+$700
US 2025 Oil Prod. >14.5M bbl/day	Brazil 2025 Inflation >5.5%	#24	#15	+$600
India 2026 Recession	CRE Delinq. Q4 2024 >3%	#51	#3	+$200

Beyond aggregate performance gains, the semantic filtering stage surfaces economically coherent lead–lag relationships that would be excluded under purely statistical screening. Table 8 reports representative *LLM-only* trades that fall outside the Granger Top-M portfolio by p-value ranking, yet yield positive realized outcomes after LLM-based re-ranking. Notably, the Japan Recession → U.S. GDP pair has a

Granger rank of #71—far outside the Top-M cutoff—yet is selected by the LLM based on a cross-border macroeconomic transmission mechanism. Despite weak Granger statistics and low surface similarity, the model assigns a negative causal sign consistent with economic reasoning linking recessions to spillovers through trade and financial channels. These examples illustrate that the LLM stage does not merely refine statistical ordering, but can recover structurally meaningful relationships that are brittle under purely statistical ranking.

A.4 Additional Experiments

In this appendix, we provide robustness analyses to examine whether the semantic re-ranking behavior we observe is stable under a post-cutoff evaluation setting and across different LLM model variants.

Table 9: Trading performance after training cutoff (entry date $>$ May 31, 2024).

Metric	Statistical Approach	Hybrid Approach	Change
Win Rate	61.8%	62.1%	+0.3pp
Avg Win	$614	$656	+6.8%
Avg Loss	-$700	-$418	**+40.3%**
Total PnL	**$3,800**	**$7,200**	**+89%**

Post-cutoff evaluation to mitigate lookahead bias. To mitigate potential lookahead from an LLM's pretraining knowledge, we evaluate our method on a post-cutoff dataset consisting only of test periods after May 31, 2024, based on the publicly documented training data cutoff of the LLM (OpenAI, 2026). All other components of the pipeline are kept unchanged.

As shown in Table 9, the hybrid approach continues to outperform the statistical baseline under post-cutoff evaluation. As in the main results, the average loss magnitude is reduced by 40.3% (from $700 to $418), yielding an 89% improvement in total PnL. These results reinforce the interpretation of the LLM as a semantic risk manager that systematically suppresses large downside losses by deprioritizing statistically fragile relationships.

Robustness across LLM model variants. We further repeat the entire semantic re-ranking using an alternative model variant, `GPT-5-mini`, and observe the same loss-reduction behavior under otherwise identical settings. Together with the deterministic inference setup (temperature$=0$) and the single-question prompt design used in our main experiments, this consistency suggests that the risk-management effect is robust across model choices rather than an artifact of a specific LLM or stochastic sampling variance.

Risk-adjusted performance across rolling windows. To verify that the improvement is not driven by a few favorable windows, we report an annualized Sharpe ratio computed as the per-window mean PnL divided by the per-window standard deviation, scaled by $\sqrt{N}$ with N=16 active windows. The hybrid approach attains a Sharpe of 1.92 versus 0.61 for the statistical baseline, a 3.1$\times$ improvement, with comparable volatility ($1,630 vs. $1,675) but roughly 3$\times$ higher mean window PnL ($781 vs. $256), indicating that the advantage is risk-adjusted rather than concentrated in a few outlier periods.

A.5 Prompt Template for LLM-Based Semantic Filtering

We use a fixed prompt format across all candidate pairs to ensure consistent semantic scoring and to avoid pair-specific tuning.

```
Analyze this prediction market event pair for causal relationship:

  Leader: {Leader Event Title}
  Follower: {Follower Event Title}

  Assess:
  1. Is there a real economic/causal mechanism (not just correlation)?
  2. Strength: high/medium/low/none
  3. Sign: positive (same direction) or negative (opposite)

  Respond in JSON:
  {
     "has_mechanism": true/false,
     "strength": "high/medium/low/none",
     "sign": "positive/negative",
     "reason": "brief explanation"
  }

  Be skeptical - many statistical correlations are spurious.
```

Figure 3: Prompt template used for LLM-based semantic filtering. Given a directed leader–follower event pair, the model assesses whether a plausible economic transmission mechanism exists (beyond correlation), assigns a strength level, and predicts the expected sign of co-movement, returning a structured JSON output.