Can Large Language Models Find Connections between Social Beliefs?

Anonymous ACL submission

Abstract

Understanding how people's perspectives on different issues change in correspondence with one another is essential for modeling collective reasoning and social dynamics. However, this problem remains underexplored due to the absence of standardized benchmarks and evaluation protocols. In this work, we introduce BELIEFBENCH, a new benchmark for evaluating whether large language models (LLMs) can detect when shifts in beliefs about one real-world event are accompanied by corresponding shifts in beliefs about another. The benchmark is constructed from Polymarket, a prediction market platform where event probabilities are updated daily, reflecting crowd belief over time. We formulate a classification task in which event pairs are labeled based on a combination of time-series co-movement, semantic similarity, and other metadata. Label quality is validated by human annotators. Our evaluation reveals two key findings: (1) LLMs consistently outperform heuristic and neural baselines in identifying meaningful belief correlations across diverse domains; (2) Chain-of-Thought prompting improves performance in settings that require multi-step reasoning, such as politics and elections, but can hurt performance in domains where surface-level signals are more predictive. BELIEFBENCH thus provides a challenging testbed for evaluating how well LLMs capture the co-evolution of perspectives and the underlying temporal and causal reasoning processes.

1 Introduction

Social beliefs represent an individual's perspective about uncertain future events—for example, presidential election outcomes or financial market trends. Each person holds many such beliefs shaped by their prior education and life experiences, and these beliefs are rarely independent. When concrete new evidence becomes available, multiple beliefs may shift simultaneously rather than in isolation. For instance, the release of a major poll showing a surge in support for Donald Trump could not only update one's belief about "Will Trump win the 2024 election?" but also shift expectations about "Will Republicans control the Senate?", since the two outcomes are closely tied. Likewise, an unexpected announcement from the Federal Reserve about tightening monetary policy may change beliefs about "Will the Fed raise interest rates?" and simultaneously alter expectations on "Will Bitcoin fall below $30,000?", capturing broader macroeconomic sentiment (Li et al., 2024a; Lee et al., 2025; Radivojevic et al., 2024).

At the collective level, such updates can trigger a "butterfly effect," where a single event reshapes

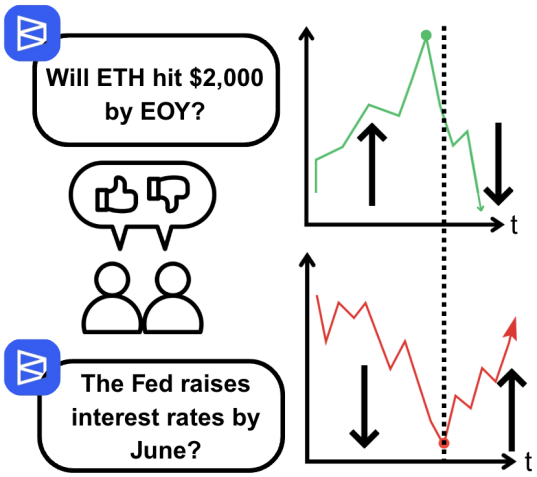

Figure 1: **Using betting platform data to study connections between social beliefs.** On platforms like Polymarket, users bet on the outcomes of real-world events. We show two markets above: *Will ETH hit $2000?* and *Will the Fed raise interest rates?*. Each market provides a time series of probabilities, reflecting collective beliefs about future events. We observe that time series for each market are often correlated—for example, the two above show contradictory trends. Such co-movements reveal connections between beliefs, motivating our study of whether LLMs can provide the reason behind such correlation.

public beliefs across society, leading to broad consequences. A financial crisis, for example, can alter expectations about interest rates, currency stability, and political outcomes, amplifying systemic risks. Understanding the correlations among social beliefs—such as co-occurrence (Kolajo et al., 2022; Minnema et al., 2023; Glandt et al., 2021), semantic similarity (Wang et al., 2024a; Cann et al., 2023; Lu et al., 2025), or implicit causal alignment (Deng et al., 2021; Wang et al., 2024b; Peng et al., 2021)—is therefore crucial for interpreting societal trends and forecasting collective behavior. Moreover, these correlations often extend beyond direct pairs: if belief A is linked to B, and B to C, people may implicitly associate A with C as well. Detecting such transitive or multi-hop relationships requires models to reason beyond surface cues and integrate temporal and institutional knowledge.

Despite its importance, automatically uncovering these latent structures remains a significant challenge. Traditional approaches (Guan et al., 2021; Ren et al., 2022; Li et al., 2024b) rely on structured datasets with fixed ontologies, which generalize poorly to the dynamic, unstructured belief expressions found in real-world forecasting or market settings. The task is further complicated by scarce and noisy belief-linked time-series data, limited leakage-free benchmarks, the difficulty of defining reliable ground truth, and the need to reason beyond lexical overlap to capture temporal patterns and latent semantics. Large language models (LLMs) are particularly promising here: pretrained on vast, diverse corpora, they can flexibly interpret unstructured language, infer implicit causal and semantic links, and reason over multi-hop dependencies without requiring rigid ontologies. This naturally raises the question: *Can LLMs identify correlations between real-world social beliefs?*

To address this question, we first build a new benchmark with high-quality labels, constructed from SocialWM (Anonymous, 2025), a dataset collected from Polymarket[1], a decentralized forecasting platform. Each event pair is labeled based on a combination of time-series co-movement and semantic similarity, with labels generated automatically and validated against human annotations. We then evaluate a diverse set of LLM families on this benchmark to assess their ability to identify meaningful belief correlations. Our results reveal several insights: (1) LLMs consistently outperform heuristic and neural baselines, with GPT-4o achieving Spearman correlations up to 0.53 and QWK scores above 0.52, significantly surpassing all heuristic baselines; (2) While Chain-of-Thought prompting (Wei et al., 2023) enhances interpretability by surfacing intermediate reasoning steps that support multi-hop associations, its impact on performance is mixed, with improvements observed in some settings but degradations in others; (3) performance drops on post-cutoff events, underscoring limits in temporal generalization due to pretraining.

[1] https://polymarket.com/

2 Related Work

Social belief correlation detection. Identifying correlations between beliefs associated with real-world events, such as co-occurrence, semantic relevance, or implicit causal links, is fundamental to understanding social dynamics. Cataldi et al. (2010) propose a co-occurrence graph to detect tweet topics. The Whatsup framework (Hettiarachchi et al., 2023) resolves co-occurring events using self-learned word embeddings. TimeBank (Gast et al., 2016) and MATRES (Ning et al., 2018) provide structured datasets for temporal and causal relation extraction. Zhou et al. (2021) introduce a BERT-based model for reasoning over event correlations. In the financial setting, MARKETGPT (Wheeler and Varner, 2024) and PLUTUS (Xu et al., 2024) develop pretrained models for market belief understanding. However, many of these studies rely on synthetic setups or structured event representations, limiting their applicability to noisy, ambiguous real-world beliefs. Our work differs by introducing a realistic evaluation task constructed from real-world market data, enabling systematic measurement of LLMs' ability to identify belief correlations under temporal uncertainty and semantic sparsity.

Social reasoning Prior work uses the term "social reasoning" to refer to tasks like understanding social norms, commonsense interactions, or modeling human mental states. For example, SocKET benchmarks LLMs on social-concept understanding and moral expectations (Choi et al., 2023), while Gandhi & colleagues study mental-state reasoning for theory-of-mind modeling (Gandhi et al., 2024). Other work evaluates LLMs' understanding of social norms in large-scale benchmark settings, such as the Social Norm dataset (Yuan et al., 2024) and the NormAd cultural adaptability framework (Rao et al., 2025). Prior work often defines social reasoning through individual or small-group

cognition, focusing on human-centric scripts or moral norms. In contrast, we define it as identifying meaningful connections between real-world social beliefs, capturing co-occurrence, semantic relevance, and implicit causality in different domains. Our task centers on reasoning over collective dynamics using noisy, unstructured signals (*e.g.*, prediction markets), shifting focus from interpersonal commonsense to event-level inference relevant for social science and forecasting.

3 Preliminary

Social belief We define a *social belief* as a collective expectation about the outcome of a real-world event. For example, people may hold beliefs such as "Candidate X will win the 2024 election," "Bitcoin will surpass $100,000 this year," or "Team Y will win the championship." These beliefs are inherently dynamic: they evolve as new information emerges, such as polling updates, economic indicators, or breaking news. Unlike static opinions, social beliefs can be quantified and tracked over time, providing a lens into how collective expectations shift in response to external events.

Data source for social belief To capture such evolving beliefs at scale, we use Polymarket[2], a decentralized prediction market where users trade on outcomes of real-world events with clear resolution criteria (*e.g.*, "Will Candidate X win the 2024 election?"). Each market typically has a binary "Yes"/"No" structure, and daily trading prices provide time-stamped, financially incentivized probability estimates that reflect the crowd's collective belief. Polymarket spans diverse domains, including politics, economics, cryptocurrency, and sports, and markets often remain active for weeks or months, allowing us to observe the temporal dynamics of belief formation and revision. Compared to text sources such as news or social media, prediction markets aggregate dispersed signals into quantitative, externally verifiable probabilities, making them a robust proxy for real-time social belief.

Connections between social beliefs While individual beliefs can be measured, identifying meaningful relationships between them remains challenging. Such connections are often implicit, multi-hop, and cross-domain—for instance, an election outcome may shift financial markets, or geopolitical tensions may influence energy prices via global supply chains. Detecting these links requires reasoning beyond surface semantics, such as identifying shared actors, institutional dependencies, or indirect causal chains. Our task examines whether LLMs can uncover these latent connections between real-world social beliefs.

[2] https://polymarket.com/

4 Problem Definition

Notation Let $\mathcal{E}$ be the set of market events from Polymarket. Each event $e \in \mathcal{E}$ is defined by a natural language question q_e and a belief trajectory $\{p_t^e\}_{t=1}^{T_e}$, where each $p_t^e \in (0,1)$ denotes the market-implied probability of event e on day t, derived from trading activity. Thus, $\{p_t^e\}_{t=1}^{T_e}$ forms a time series capturing the temporal dynamics of the evolving social belief, while q_e provides the semantic content of the event.

Social belief correlation task Formally, given a pair of events $(A, B) \in \mathcal{E} \times \mathcal{E}$, each represented by their natural language descriptions (q_A, q_B) and belief trajectories $\{p_t^A\}_{t=1}^{T_A}$ and $\{p_t^B\}_{t=1}^{T_B}$, where each $\{p_t^e\}$ is a time series of daily market-implied probabilities, the goal is to estimate the strength of correlation between their associated social beliefs. We define a mapping

$$f : \mathcal{E} \times \mathcal{E} \longrightarrow \mathcal{Y}, \tag{1}$$

where the label space $\mathcal{Y} = \{1, 2, 3, 4, 5\}$ encodes ordinal levels of correlation: 1 (very weak), 2 (weak), 3 (medium), 4 (strong), and 5 (very strong). The task is formulated as ordinal classification, where the model predicts a discrete label $\hat{y} = f(A, B) \in \mathcal{Y}$. In our setting, f can be a LLM, which jointly reasons over the semantic content (q_A, q_B) and the temporal patterns $(\{p_t^A\}, \{p_t^B\})$ to infer correlation strength.

5 BELIEFBENCH: A Benchmark for Tracking Co-evolution of Social Beliefs

To benchmark the social belief correlation task, we construct a dataset based on SocialWM (Anonymous, 2025), derived from Polymarket. In this section, we first explain how belief pairs are selected, then describe our procedure for collecting ground-truth correlation labels, and how we verify them. Finally, we provide details on the evaluation methodology.

5.1 Belief Pair Selection

Not all markets offer informative or reliable signals for belief reasoning. To ensure that the included

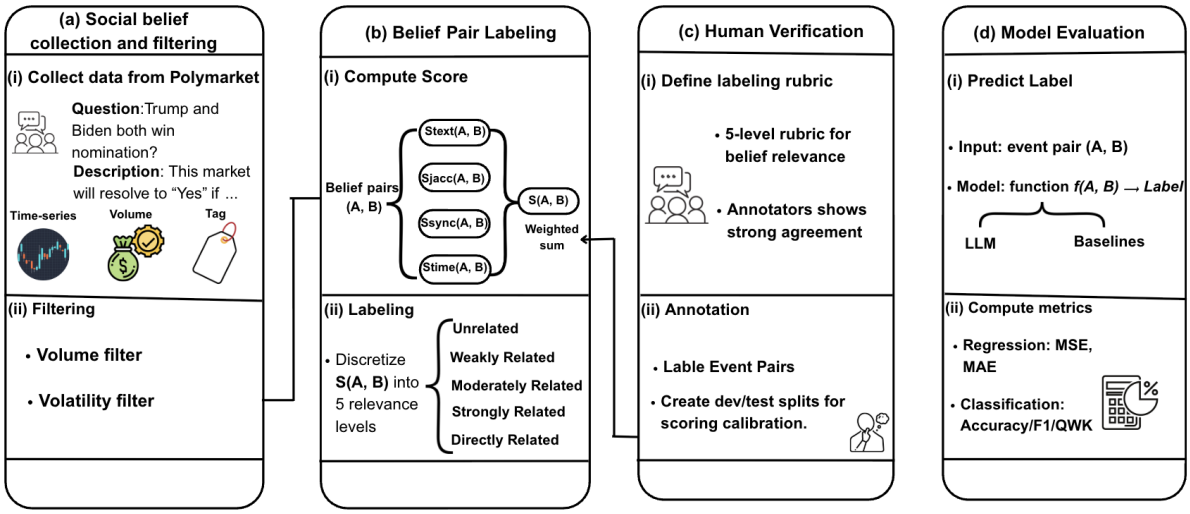

Figure 2: **Overview of the BELIEFBENCH construction and evaluation pipeline.** (a) Social belief data is collected and filtered from Polymarket. (b) Event pairs are constructed and scored based on semantic and temporal signals. The resulting correlation score $S(A, B)$ is discretized into five relevance levels. (c) Human verification is conducted to validate labeling quality, with annotators reaching strong agreement using a shared rubric. (d) Models predict relevance labels for event pairs. Performance is evaluated using both regression and classification metrics.

events reflect collective crowd beliefs rather than noise, we apply two filters: one based on trading volume, the other on volatility of belief movement.

Volume filter Markets with very low trading volume are often driven by isolated trades and do not reflect meaningful aggregation of public belief. We remove the bottom 25% of events by trading volume within each domain. This helps exclude illiquid or inactive markets where probability shifts are unreliable.

Volatility filter We require the event to have sufficient probability movement. A flat probability series provides little statistical signal. By imposing a minimum volatility threshold, we ensure that the probability series contain enough variation to make the correlation test meaningful. Let $r_t = \text{logit}(p_t) - \text{logit}(p_{t-1})$ be logit return, *i.e.*, day-to-day changes in log-odds. Denote by $\sigma_t^{(w)}$ the rolling standard deviation of $\{r_\tau\}_{\tau=t-w+1}^{t}$ over a window of w days. Let γ be the volatility threshold and α be the required proportion. We retain a base market only if

$$\frac{1}{T-w+1}\sum_{t=w}^{T}\mathbf{1}\left[\sigma_t^{(w)} \geq \gamma\right] \geq \alpha, \qquad (2)$$

i.e., at least α fraction of the windows have the standard deviation of the daily logit returns above γ.

5.2 Belief Pair Labeling

To establish ground-truth labels for belief correlation, we adopt a hybrid scoring framework that integrates both semantic and temporal information. This enables us to measure not only surface-level similarity but also co-movement in belief dynamics across event pairs. Each pair (A, B) is assigned a composite score $S(A, B)$ based on four interpretable features.

Feature1: Change-point synchrony We identify time points where an event's belief trajectory exhibits statistically significant shifts by applying z-score thresholding to the price deltas in its time series. For each event, we extract a set of such change points. The synchrony score then measures the fraction of change points in event A that align within a short temporal window δ of any change point in event B. This captures the intuition that jointly fluctuating beliefs are likely to be correlated:

$$s_1(A, B) = \frac{1}{|T_A|}\sum_{t\in T_A} 1\left[\exists\, t' \in T_B,\ |t - t'| < \delta\right]. \qquad (3)$$

Feature2: Tag Jaccard similarity To estimate topical overlap, we use the Jaccard index over tag sets from Polymarket metadata. Each event includes tags that describe its domain or subject matter. A high Jaccard score indicates that two events are framed under similar categories or themes,

which may reflect a shared discourse context:

$$s_2(A, B) = \frac{|\mathcal{K}_A \cap \mathcal{K}_B|}{|\mathcal{K}_A \cup \mathcal{K}_B|}. \quad (4)$$

Feature3: Minimum time gap We compute the minimum absolute time difference between any change point in event A and any in event B. This measures how closely belief shifts in the two events occur in time. We convert this to a soft similarity score using a monotonic inverse transformation:

$$s_3(A, B) = \frac{1}{1 + \min\limits_{t \in T_A, t' \in T_B} \frac{|t - t'|}{\tau}}. \quad (5)$$

Feature4: Textual similarity We embed the text descriptions of events using sentence-transformer models and compute cosine similarity between the resulting embeddings. This feature captures semantic proximity at the lexical and conceptual level, and complements the tag-based feature with more nuanced language modeling:

$$s_4(A, B) = 1 - \cos\left(\mathbf{e}_A, \mathbf{e}_B\right). \quad (6)$$

Overall The four feature scores are linearly combined into a single heuristic correlation score. The weights are optimized on a development set to best match human relevance judgments. We discretize $S(A, B)$ into five relevance classes: *very weak* (0.0–0.2), *weak* (0.2–0.4), *medium* (0.4–0.6), *strong* (0.6–0.8), and *very strong* (0.8–1.0). These bucketed labels serve as groundtruth in evaluation:

$$S(A, B) = \sum_{i=1}^{4} w_i \cdot s_i(A, B) \quad (7)$$

where w_i is tuned to make the prediction highly aligned with a small set of human annotation results.

5.3 Human Verification

To validate the quality of the heuristic labels used in evaluation, we conducted a human annotation study on a representative subset of 50 event pairs, stratified across five correlation levels. Three annotators rated each pair based solely on textual semantics and related news, without access to belief trajectories or model predictions. Inter-annotator agreement was strong, with pairwise Pearson correlations ranging from 0.75 to 0.81 and an intra-class correlation (ICC) of 0.77. Furthermore, the aggregated human judgments exhibit high alignment with the heuristic scores used in our benchmark, achieving a Pearson correlation of 0.689. These results confirm that the scoring function reflects intuitive assessments of belief correlation. Full protocol details and annotation examples are provided in Appendix §G.

5.4 Model Evaluation

We evaluate model predictions against labels using both regression and classification metrics. For regression, we first map the five discrete relevance classes to their midpoint values, and then compute Mean Squared Error (MSE), Mean Absolute Error (MAE) between predicted scores and these bin centers. For classification, we discretize model outputs into five bins and report accuracy, macro-averaged F1 score, and quadratic-weighted kappa (QWK) to assess ordinal agreement. These metrics collectively measure both score precision and the ability to capture ordinal relevance patterns.

6 Experiments

LLMs have powerful social reasoning capability. Here we used the data in SocialWM to test whether LLMs can judge the degree of relevance between real-world events, using a five-level scale from very weak to very strong.

6.1 Experimental Settings

For experiments, we consider two categories of models: (1) baselines using heuristic rules or pretrained neural encoders, and (2) prompting-based LLMs that generate predictions.

Baselines We implement several heuristic baselines that rely on simple similarity or overlap metrics computed from event metadata. We also include a neural baseline using a cross-encoder model (`nli-deberta-v3-base`), which computes a scalar relevance score from the concatenated text of the two event descriptions. These continuous scores are then discretized into the same five relevance bins for evaluation.

Prompting-based LLMs We evaluate a diverse set of language models, including GPT-4o, Qwen2-72B-Instruct, DeepSeek-R1, and others, using a prompting-based classification setup. Each model receives as input the titles and descriptions of two events and is asked to assess how strongly the events are related by selecting one of five predefined relevance intervals. For each model, we com-

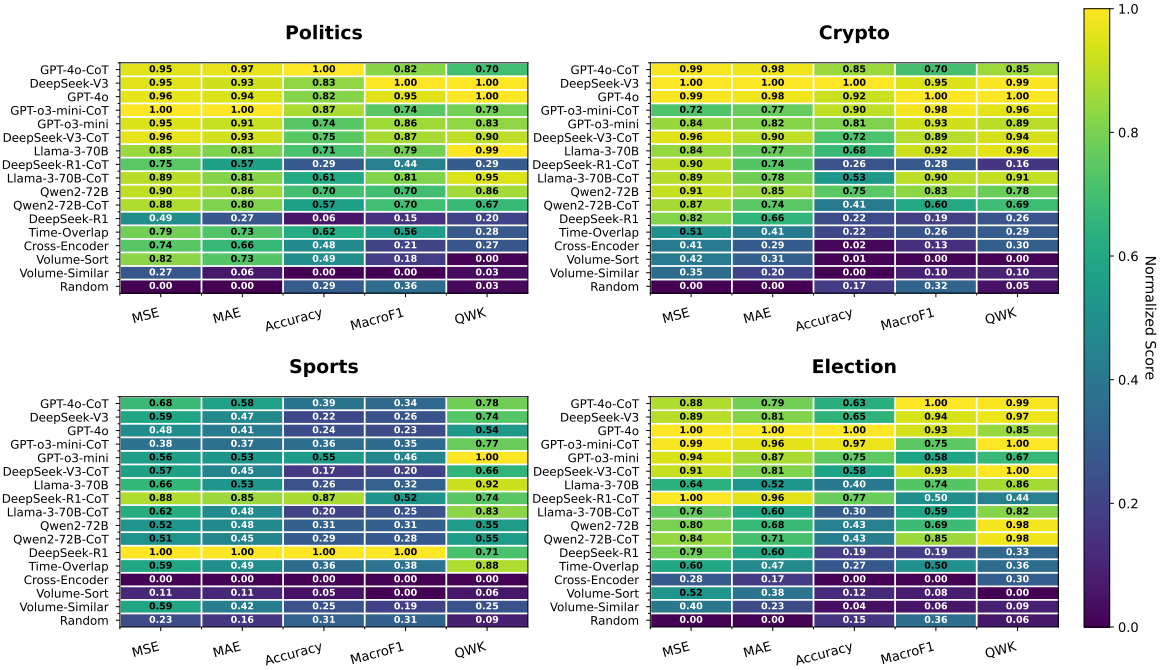

Figure 3: **Overall performance heatmaps across four domains: Politics, Crypto, Sports, and Election.** Each heatmap shows model performance across five metrics: MSE, MAE (lower is better), and Accuracy, Macro-F1, QWK (higher is better). To make metrics comparable, error metrics are negated and all values are min-max normalized within each metric and dataset. Models are sorted by their average normalized score across all datasets. Higher values indicate better normalized performance.

pare two prompting strategies: a direct classification setting and a Chain-of-Thought (CoT) variant. In the CoT setting, the model first generates a brief explanation describing any semantic, causal, or temporal connections it identifies between the events. It then outputs a structured JSON object indicating the predicted relevance label. This design allows us to evaluate both the classification performance and the interpretability of the model's reasoning process.

6.2 Experimental Results

GPT-4o+CoT demonstrates strong and consistent performance across domains. In Figure 3, we evaluate model performance on four domains: Politics, Election, Crypto, and Sports. GPT-4o+CoT consistently ranks among the top models in both classification metrics (Accuracy, F1, QWK) and regression metrics (MSE, MAE). Compared to smaller closed-source models such as GPT-o3-mini and open-source models including Meta-Llama-3-70B, Qwen2-72B, GPT-4o+CoT achieves high overall scores with stable results across domains. While heuristic baselines (e.g., based on time overlap or volume) are included for reference, they typically lag behind LLM-based models. This suggests that even without access to metadata, large language models can effectively infer event relationships from text alone.

LLMs perform well in semantically dense domains but struggle in sparse ones. As shown in Figure 3, model performance varies by domain structure. In Crypto and Election, where events share entities, timelines, or institutions, models achieve stronger results. Even simple heuristics perform well due to the rich semantic context. In contrast, Sports events are often isolated and actor-specific, leading to the weakest performance. Political events fall in between, requiring both structural and contextual reasoning. These patterns suggest that LLMs are most effective in domains with coherent and recurring semantics.

The effectiveness of CoT prompting varies by domain and reasoning complexity. Chain-of-Thought (CoT) prompting is most effective in domains requiring multi-hop reasoning, such as Politics and Election. For example, GPT-4o-CoT and GPT-o3-mini-CoT achieve strong regression performance (MSE, MAE) in Politics, while GPT-o3-mini-CoT and DeepSeek-V3-CoT improve ranking consistency in Election. However,

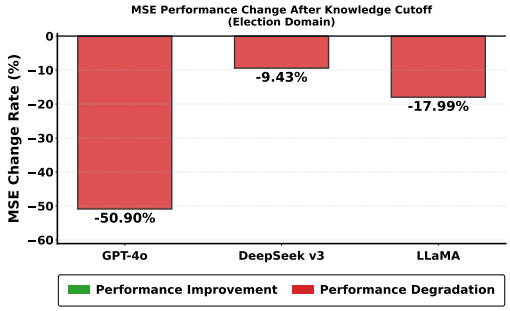

Figure 4: **Performance degradation on post-cutoff events across models (Election domain).** All models exhibit worse MSE performance on post-cutoff event pairs, highlighting challenges in temporal generalization. GPT-4o shows the smallest increase in error. (MSE values are sign-inverted for visualization clarity.)

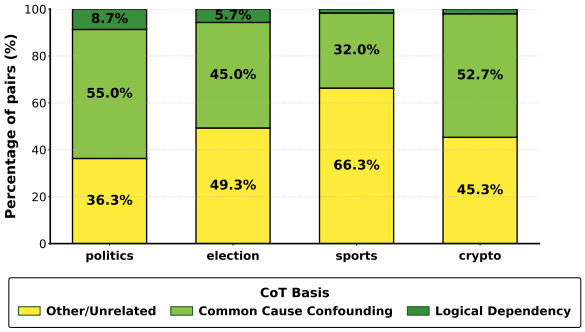

Figure 5: **Distribution of reasoning types from GPT-4o's Chain-of-Thought outputs across domains**. Each pair was labeled based on the explanation produced by the model. A majority of predictions are based on shared context (*confounding*) or loose narrative links (*CoT basis*), while only a small portion exhibit explicit logical or causal reasoning. This suggests the model is primarily identifying correlations rather than inferring direct causal links.

gains are uneven: GPT-4o shows better calibration in Politics but lower Macro-F1 and QWK, and Election improvements remain mostly in regression metrics. In contrast, CoT consistently reduces performance in Sports and Crypto, where relevance depends on surface-level patterns; here, CoT variants of GPT-o3-mini and DeepSeek underperform across Accuracy and QWK. Overall, CoT helps when latent dependencies must be inferred but introduces noise in domains driven by direct signals.

7 Discussion

We investigate how LLMs reason about social event relevance through two research questions. RQ1 shows performance degrades after the knowledge cutoff (Figure 4), highlighting LLMs' dependence on up-to-date factual knowledge. RQ2 analyzes Chain-of-Thought outputs, revealing that LLMs rely mainly on confounding signals and narrative connections rather than explicit logic.

RQ1: Does the knowledge cutoff influence performance? We investigate whether LLMs rely on factual knowledge from pretraining or can generalize to unseen events. To this end, we compare model performance on event pairs occurring before and after the model's knowledge cutoff. As shown in Figure 4, all evaluated models exhibit clear performance degradation on post-cutoff examples in the *election* domain, measured by percentage change in MSE. For instance, GPT-4o shows a substantial drop of over 50%, while DeepSeek-v3 and LLaMA-3 also experience notable declines. These results suggest that while LLMs may generalize to unseen patterns to some extent, their ability to capture belief correlation often depends on up-to-date world knowledge learned during pretraining.

RQ2: How do LLMs judge belief correlation? To better understand what types of relationships LLMs rely on when judging belief correlation, we analyze their Chain-of-Thought (CoT) outputs and categorize the reasoning basis. As shown in Figure 5, only a small fraction of cases reflect explicit logical connections: approximately 8.7% in politics and less than 5.7% in sports. In contrast, a large proportion of predictions fall under *confounding* relationships (e.g., shared context or common background factors), accounting for 55% in politics and 32% in sports. These results suggest that LLMs do not primarily rely on formal logic or direct causality. Instead, they often identify perceived connections through narrative, intuition, or shared framing. This supports our interpretation that the LLM captures *relatedness* rather than strict *causal inference*.

8 Case Study

To better understand how LLMs assess event relevance, we analyze predictions across representative event pairs. We group cases into two types: (1) 0-hop pairs, which exhibit surface-level thematic overlap, and (2) 1-hop or multi-hop pairs, which require reasoning over latent causal or institutional structures. For each category, we examine the behavior of GPT-o3-mini with and without Chain-of-Thought (CoT) prompting. Further qualitative

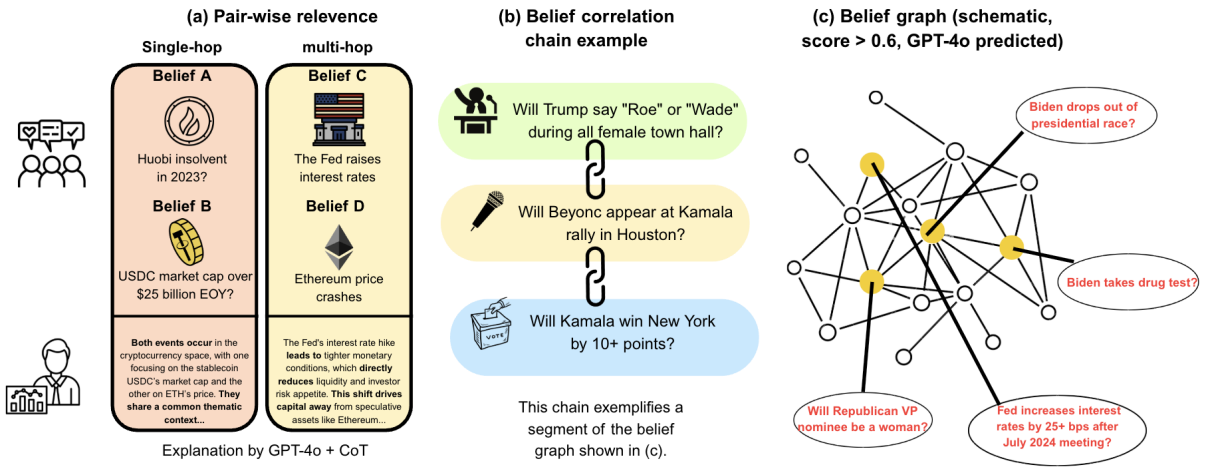

Figure 6: **Case study of social belief reasoning with LLMs**. **(a)** Pairwise relevance cases: GPT-4o+CoT explains both single-hop (topical) and multi-hop (causal) belief pairs. **(b)** An example belief chain constructed from high-relevance event pairs (score > 0.6), with links inferred by GPT-4o using Chain-of-Thought reasoning. **(c)** A schematic belief correlation graph, where edges denote LLM-predicted high-confidence belief correlations. The chain in (b) exemplifies a segment within this network.

examples and detailed outputs are included in Appendix §C.

Single-hop: Topical overlap These pairs often belong to the same domain (*e.g.*, cryptocurrency) and share broad semantic context, but lack direct interaction or dependency. In this setting, CoT prompting frequently introduces spurious logic, weakening relevance predictions. For example, the model underestimates the relationship between ETH price and USDC market cap by overemphasizing their economic independence. In contrast, the non+CoT variant better preserves local topical proximity, yielding predictions that align more closely with human annotations. This suggests that when cues are shallow but sufficient, simple inference is preferable to added reasoning steps.

Multi-hop: Causal or institutional reasoning We define multi-hop reasoning as cases where the relevance between events depends on indirect or institutional links, such as shared political actors, procedural dependencies, or regulatory chains, rather than simple topical overlap. In such settings, particularly in the politics and election domains where multi-hop structures are more common, CoT prompting tends to improve overall prediction metrics. For instance, when evaluating how nomination outcomes relate to vice-presidential picks, or how financial regulations affect related assets, CoT helps the model trace relevant dependencies. However, despite the overall gains, a non-negligible number of cases still suffer from flawed reasoning, such as hallucinated links or incoherent logic.

Butterfly-effect: Chainable social beliefs In addition to reasoning over isolated event pairs, LLMs can identify extended chains of belief correlations. As illustrated in Figure 6, GPT-4o with CoT generates coherent reasoning paths between events that are not obviously related on the surface, such as political discourse, celebrity actions, and electoral outcomes. The center panel shows a plausible belief chain inferred from pairwise high-scoring links, which forms a subgraph of the larger belief network shown on the right. This reflects what we refer to as the *social butterfly effect*, where local signals propagate through institutional or topical structures to shape broader expectations. These high-confidence structures are observed not only in the Election domain but also in other domains, suggesting that LLMs are capable of reconstructing latent belief networks from unstructured input.

9 Conclusion

We present BELIEFBENCH, a new benchmark for evaluating LLMs' ability to reason over real-world social belief correlations derived from prediction market data. LLMs consistently outperform heuristic and embedding-based methods across domains, revealing their capacity to identify semantic and temporal relationships in belief dynamics. While Chain-of-Thought prompting helps in complex reasoning cases, it may reduce accuracy in simpler contexts. Our findings highlight both the promise and current limits of LLMs in modeling evolving social beliefs, and point to future directions in adaptive prompting and temporal modeling.

10 Limitations

Heuristic-score-based ground truth Our ground-truth labels are derived from a weighted heuristic score $S(A, B)$ that combines temporal synchrony, textual similarity, tag overlap, and time alignment (see Section §5.2). Although this method improves over pure correlation-based approaches (e.g. Kendall's τ), it can still assign high scores to spurious pairs, for example events with spikes in coincident volatility or shared metadata but without substantive connection. Such false positives can penalize models that correctly reject these superficial links, limiting the fidelity of the supervision signal.

Platform and domain bias Polymarket does not list every real-world event - in many domains, the coverage is patchy.

Black-box prompt-based design Our study intentionally focuses on black-box, prompt-based approach without task-specific fine-tuning or custom model architectures. While this choice aligns with our goal of evaluating LLMs in realistic usage scenarios, it limits our ability to optimize performance on this task. Future work could explore fine-tuning, retrieval-augmented methods, or specialized architectures to better capture subtle belief relationships.

Temporal overlap assumption Our approach focuses on social belief pairs with overlapping active periods to ensure that the measured time-series correlations capture dynamic co-movement as traders respond to new information. While this design helps reduce noise in estimating relevance, it also limits the benchmark's ability to evaluate delayed or indirect causal links that might manifest outside of these overlapping windows. Future work could explore more advanced temporal modeling strategies, such as lag-aware correlation measures or causal inference techniques to better capture these complex, cross-temporal relationships.

11 Ethical Statement

This work analyzes public event data from Polymarket, a prediction market platform that provides open-access market-level data without any user-identifiable information. We do not collect or process individual-level data, and all analysis is conducted at the event level. Thus, privacy concerns are minimal.

Our evaluation framework involves using large language models (LLMs) to assess the relevance between social events. These models, while powerful, may exhibit unintended biases, particularly in politically sensitive or socially charged domains. We caution against using these models as authoritative predictors or decision-making tools in high-stakes environments.

Additionally, while our work aims to understand event relationships, it does not attempt to forecast outcomes or provide trading recommendations. The models are evaluated solely on their reasoning and ranking capability and should not be interpreted as reliable financial or political forecasting instruments.

Finally, while our method is training-free, the evaluation dataset itself may reflect biases from Polymarket's coverage, which is shaped by community interest and market dynamics. As a result, certain domains, such as Sports or Politics, may be overrepresented, potentially influencing model predictions or evaluation trends. We encourage future work to broaden coverage to include a more balanced set of social domains.

References

2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

2024. Qwen2 technical report. *Preprint*, arXiv:2407.10671.

2025a. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

2025b. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

Anonymous. 2025. Title omitted for double-blind review. Under review.

Tristan J. B. Cann, Ben Dennes, Travis Coan, Saffron O'Neill, and Hywel T. P. Williams. 2023. Using semantic similarity and text embedding to measure the social media echo of strategic communications. *Preprint*, arXiv:2303.16694.

Mario Cataldi, Luigi Di Caro, and Claudio Schifanella. 2010. Emerging topic detection on twitter based on temporal and social terms evaluation. In *Proceedings of the Tenth International Workshop on Multimedia Data Mining*, MDMKDD '10, New York, NY, USA. Association for Computing Machinery.

Minje Choi, Jiaxin Pei, Sagar Kumar, Chang Shu, and David Jurgens. 2023. Do LLMs understand social

knowledge? evaluating the sociability of large language models with SocKET benchmark. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11370–11403, Singapore. Association for Computational Linguistics.

Songgaojun Deng, Huzefa Rangwala, and Yue Ning. 2021. Causal knowledge guided societal event forecasting. *Preprint*, arXiv:2112.05695.

Saumya Gandhi, Ritu Gala, Vijay Viswanathan, Tongshuang Wu, and Graham Neubig. 2024. Better synthetic data by retrieving and transforming existing datasets. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6453–6466, Bangkok, Thailand. Association for Computational Linguistics.

Volker Gast, Lennart Bierkandt, Stephan Druskat, and Christoph Rzymski. 2016. Enriching TimeBank: Towards a more precise annotation of temporal relations in a text. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3844–3850, Portorož, Slovenia. European Language Resources Association (ELRA).

Kyle Glandt, Sarthak Khanal, Yingjie Li, Doina Caragea, and Cornelia Caragea. 2021. Stance detection in COVID-19 tweets. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1596–1611, Online. Association for Computational Linguistics.

Saiping Guan, Xueqi Cheng, Long Bai, Fujun Zhang, Zixuan Li, Yutao Zeng, Xiaolong Jin, and Jiafeng Guo. 2021. What is event knowledge graph: A survey. *CoRR*, abs/2112.15280.

Hansi Hettiarachchi, Mariam Adedoyin-Olowe, Jagdev Bhogal, and Mohamed Medhat Gaber. 2023. Whatsup: An event resolution approach for co-occurring events in social media. *Information Sciences*, 625:553–577.

Taiwo Kolajo, Olawande Daramola, and Ayodele A Adebiyi. 2022. Real-time event detection in social media streams through semantic analysis of noisy terms. *Journal of Big Data*, 9(1):90.

Jean Lee, Nicholas Stevens, and Soyeon Caren Han. 2025. Large language models in finance (finllms). *Neural Computing and Applications*.

Lincan Li, Jiaqi Li, Catherine Chen, Fred Gui, Hongjia Yang, Chenxiao Yu, Zhengguang Wang, Jianing Cai, Junlong Aaron Zhou, Bolin Shen, Alex Qian, Weixin Chen, Zhongkai Xue, Lichao Sun, Lifang He, Hanjie Chen, Kaize Ding, Zijian Du, Fangzhou Mu, and 28 others. 2024a. Political-llm: Large language models in political science. *Preprint*, arXiv:2412.06864.

Pu Li, Xiaoyan Yu, Hao Peng, Yantuan Xian, Linqin Wang, Li Sun, Jingyun Zhang, and Philip S. Yu. 2024b. Relational prompt-based pre-trained language models for social event detection. *Preprint*, arXiv:2404.08263.

Yiwei Lu, Zhengtao Yu, Yantuan Xian, Yuxin Huang, and Yan Xiang. 2025. Unsupervised social media event detection method combining semantic edge pruning and community discovery. In *Proceedings of the 4th International Conference on Computer, Artificial Intelligence and Control Engineering*, CAICE '25, page 1045–1052, New York, NY, USA. Association for Computing Machinery.

Gosse Minnema, Huiyuan Lai, Benedetta Muscato, and Malvina Nissim. 2023. Responsibility perspective transfer for Italian femicide news. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7907–7918, Toronto, Canada. Association for Computational Linguistics.

Qiang Ning, Hao Wu, and Dan Roth. 2018. A multi-axis annotation scheme for event temporal relations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1318–1328, Melbourne, Australia. Association for Computational Linguistics.

Hao Peng, Jianxin Li, Yangqiu Song, Renyu Yang, Rajiv Ranjan, Philip S. Yu, and Lifang He. 2021. Streaming social event detection and evolution discovery in heterogeneous information networks. *Preprint*, arXiv:2104.00853.

Kristina Radivojevic, Nicholas Clark, and Paul Brenner. 2024. Llms among us: Generative ai participating in digital discourse. In *Proceedings of the AAAI Symposium Series*, volume 3, pages 209–218.

Abhinav Sukumar Rao, Akhila Yerukola, Vishwa Shah, Katharina Reinecke, and Maarten Sap. 2025. NormAd: A framework for measuring the cultural adaptability of large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2373–2403, Albuquerque, New Mexico. Association for Computational Linguistics.

Jiaqian Ren, Lei Jiang, Hao Peng, Zhiwei Liu, Jia Wu, and Philip S. Yu. 2022. Evidential temporal-aware graph-based social event detection via dempster-shafer theory. *Preprint*, arXiv:2205.12179.

Haoyu Wang, Hongming Zhang, Kaiqiang Song, Dong Yu, and Dan Roth. 2024a. Event semantic classification in context. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1395–1407, St. Julian's, Malta. Association for Computational Linguistics.

Zimu Wang, Lei Xia, Wei Wang, and Xinya Du. 2024b. Document-level causal relation extraction with knowledge-guided binary question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16944–16955, Miami, Florida, USA. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

Aaron Wheeler and Jeffrey D. Varner. 2024. Marketgpt: Developing a pre-trained transformer (gpt) for modeling financial time series. *Preprint*, arXiv:2411.16585.

Yuanjian Xu, Anxian Liu, Jianing Hao, Zhenzhuo Li, Shichang Meng, and Guang Zhang. 2024. Plutus: A well pre-trained large unified transformer can unveil financial time series regularities. *Preprint*, arXiv:2408.10111.

Ye Yuan, Kexin Tang, Jianhao Shen, Ming Zhang, and Chenguang Wang. 2024. Measuring social norms of large language models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 650–699, Mexico City, Mexico. Association for Computational Linguistics.

Yucheng Zhou, Xiubo Geng, Tao Shen, Guodong Long, and Daxin Jiang. 2021. Eventbert: A pre-trained model for event correlation reasoning. *Preprint*, arXiv:2110.06533.

A Artifact Details

A.1 Artifact Information

This artifact contains all components required to reproduce the results in our study of belief correlation reasoning in large language models (LLMs). It includes:

- **Code:** A complete implementation of the pairwise belief correlation scoring pipeline, including preprocessing, model inference (with and without Chain-of-Thought prompting), and evaluation metrics.
- **Data:**
  - Manually annotated development and test sets across four domains: Politics, Election, Cryptocurrency, and Sports.
  - Rubric definitions used to guide annotation.
  - Annotation metadata and inter-annotator agreement statistics.
- **Models:** Inference scripts for querying multiple foundation models via standard APIs. Specifically, GPT-4o and GPT-o3-mini were accessed through the official OpenAI API, while Meta-Llama-3, DeepSeek-V3, and Qwen2 series were accessed via the Together.ai inference platform. All calls are wrapped with reproducible configurations, and API versions are specified to ensure consistent results across runs. For models supporting Chain-of-Thought (CoT) prompting, the corresponding CoT-enabled variants are also included.
- **Evaluation:** Scripts to compute both regression and classification metrics, including MSE, MAE, Accuracy, Macro-F1, QWK. Also included are scripts to produce the figures and tables in the main paper and appendix.
- **Case Study Tools:** Utilities for constructing belief chains, visualizing belief graphs, and analyzing CoT rationales.

The artifact is designed for easy replication and modification. Each script is documented with usage instructions, input formats, and expected outputs. Running the default configuration will reproduce all key results from the paper. At the time of submission, these materials are under preparation for release. We will make the code and data available upon publication.

A.2 Artifact License

All components of our artifact are intended for research use and will be released under open-source or permissive licenses upon publication.

- **Codebase:** The full codebase, including preprocessing, inference, and evaluation scripts, will be released under the MIT License.
- **Annotated Data:** The manually labeled development and test sets, along with rubric definitions and annotation metadata, are original contributions of this work. These datasets will be released under the CC BY 4.0 License, permitting reuse with attribution for research and non-commercial purposes.
  - **Codebase:** The full codebase, including preprocessing, inference, and evaluation scripts, will be released under the MIT License.
  - **Annotated Data:** The manually labeled development and test sets, along with rubric definitions and annotation metadata, are original contributions of this work. These datasets will be released under the CC BY 4.0 License, permitting reuse with attribution for research and non-commercial purposes.
  - **Model Usage:** Our study relies on querying several pretrained language models. We use **GPT-4o** and **GPT-o3-mini** via the OpenAI API,[3] which are proprietary models licensed by OpenAI. We also evaluate open-weight models including **Meta-Llama-3 70B** (gra, 2024), **DeepSeek-V3** (dee, 2025b), **DeepSeek-R1** (dee, 2025a), and **Qwen2** (yan, 2024), accessed through the Together.ai inference platform, all released under Apache 2.0 or similar permissive licenses. For comparison, we include a **cross-encoder baseline** using `nli-deberta-v3-base`[4] from Hugging Face, licensed under the MIT License.

We respect all license terms associated with the use of these third-party models and APIs. No model weights are redistributed. All data and code will be clearly marked with their respective licenses in the released repository.

A.3 Data Usage

Our dataset includes events across four domains: Politics, Election, Cryptocurrency, and Sports. We use a subset of Polymarket data curated by prior work currently under review (Anonymous, 2025). The final dataset will be released under the MIT License for academic use.

- **Source and Licensing:**

[3] https://platform.openai.com/docs/models/gpt-4o

[4] https://huggingface.co/cross-encoder/nli-deberta-v3-base

- **Use Consistency:** Our data usage is consistent with the intended purpose of the source materials, which were either licensed for research or created explicitly for this project. No repurposing beyond research evaluation has been conducted.
- **Human Annotation:** Each belief correlation pair in the development and test sets was labeled by multiple annotators using a rubric-based scale. Inter-annotator agreement scores are included in the Appendix §G to reflect labeling quality.
- **Privacy and Safety:** The dataset does not contain any personally identifiable information (PII), user metadata, or social media handles. All text has been reviewed to exclude offensive content, and no inference was made regarding demographic or protected attributes.
- **Intended Use:** The dataset is intended exclusively for research on social reasoning, belief dynamics, and LLM evaluation. It is not suitable for deployment in user-facing applications or downstream tasks involving sensitive decision-making.

A.4 Data Statistics

Our benchmark covers four domains: Politics, Election, Cryptocurrency, and Sports.

The final benchmark includes:

- **Total event pairs:** 8,839
- **Label format:** Each pair is assigned a continuous belief correlation score in the range $[0, 1]$, reflecting graded relatedness. For classification-based analyses, scores are mapped to a 5-point ordinal scale (from strongly unrelated to strongly related) using predefined thresholds.
- **Label source:** The majority of labels were derived programmatically via rubric-based scoring; a small subset was verified by human annotators for calibration and quality assurance.
- **Agreement check:** For the verified subset, each pair was annotated by 3 annotators. The average inter-annotator correlation exceeds 0.78, indicating strong agreement on the ordinal scale used for verification.

B LLM Prompt

The prompt of the zero shot and CoT version are in Table 1

C Case Study Examples

C.1 0-hop: Thematic Overlap without Direct Link

These pairs are topically similar but lack direct causal or temporal connection.

Example 1 **Event 1:** USDC market cap over $25 billion EOY?
Event 2: Will ETH hit $2,000 by EOY?
Ground Truth: `0.4-0.6`
GPT-o3-mini (CoT): `0.2-0.4`
GPT-o3-mini (No CoT): `0.4-0.6`
CoT: "Both events occur in the cryptocurrency space, with one focusing on the stablecoin USDC's market cap and the other on ETH's price. They share a common thematic context but are measuring distinct characteristics of the crypto market with minimal direct causal influence between them."

Example 2 **Event 1:** Will ETH hit $2,000 by EOY?
Event 2: Huobi insolvent in 2023?
Ground Truth: `0.4-0.6`
GPT-o3-mini (CoT): `0.2-0.4`
GPT-o3-mini (No CoT): `0.4-0.6`
CoT: "Both events relate to the cryptocurrency industry. However, while ETH reaching a certain price is a market performance event, Huobi's insolvency is a corporate financial event. Their connection is thematic but not directly causal or temporally linked."

C.2 1-hop or Multi-hop: Causal or Institutional Links

These examples involve shared actors, institutional decisions, or policy-level connections.

Example 3 **Event 1:** Will SEC delay Bitcoin ETF decision?
Event 2: Ethereum spot ETF approved by Mar 31?
Ground Truth: `0.6-0.8`
GPT-o3-mini (CoT): `0.6-0.8`
GPT-o3-mini (No CoT): `0.4-0.6`
CoT: "Both events involve SEC decisions regarding cryptocurrency ETFs. They are thematically linked, as they both pertain to regulatory actions on crypto investments and have overlapping timelines, despite focusing on different digital assets (Bitcoin and Ethereum)."

Example 4 **Event 1:** Will Haley beat DeSantis in Iowa?

Original Prompt	CoT Prompt

You are an expert in interpreting and reasoning about real-world events, such as those found on prediction markets.
Your task is to read two events and **classify** how related they are using *exactly one* of the label keys below. Base your judgement on any causal, temporal, or thematic relationship that may exist.
 	
Return **only** a valid JSON object in the form: `{"label": "0.4-0.6"}`. Do *not* include explanations, formatting, or any additional text.
 	
Labels (choose one key only):
`"0.0-0.2"` → Unrelated — No real connection in topic, time, or influence.
`"0.2-0.4"` → Weakly Related — Small thematic overlap, but no causal or temporal influence.
`"0.4-0.6"` → Moderately Related — Events share context or actors but remain largely independent.
`"0.6-0.8"` → Strongly Related — One event influences or is likely affected by the other.
`"0.8-1.0"` → Highly Related — One event is clearly a consequence, cause, or restatement of the other.
 	
Event 1
`{e1_title}`
`{e1_desc}`
 	
Event 2
`{e2_title}`
`{e2_desc}`
 	
Strictly respond with a JSON object like: `{"label": "0.6-0.8"}`
	You are an expert in interpreting and reasoning about real-world events, such as those found on prediction markets.
	Your task is to read two events and **classify** how related they are using *exactly one* of the label keys below. Base your judgement on any causal, temporal, or thematic relationship that may exist.
	
	First, in *a few concise sentences*, explain any causal, temporal, or thematic links you see.
	Then, on a new line, output *ONLY* one valid JSON object like: `{"label": "0.4-0.6"}`. Do *not* include any additional text.
	
	Labels (choose one key only):
	`"0.0-0.2"` → Unrelated — No real connection in topic, time, or influence.
	`"0.2-0.4"` → Weakly Related — Small thematic overlap, but no causal or temporal influence.
	`"0.4-0.6"` → Moderately Related — Events share context or actors but remain largely independent.
	`"0.6-0.8"` → Strongly Related — One event influences or is likely affected by the other.
	`"0.8-1.0"` → Highly Related — One event is clearly a consequence, cause, or restatement of the other.
	
	Event 1
	`{e1_title}`
	`{e1_desc}`
	
	Event 2
	`{e2_title}`
	`{e2_desc}`
	
	Give your concise reasoning, *then* output the JSON object on the next line.

Table 1: Comparison of Original vs. CoT prompt templates

Event 2: Will Kristi Noem win the U.S. 2024 Republican VP nomination?
Ground Truth: `0.2-0.4`
GPT-o3-mini (CoT): `0.2-0.4`
GPT-o3-mini (No CoT): `0.4-0.6`
CoT: "Both events occur within the context of U.S. Republican political dynamics in the 2024 election cycle, sharing a general theme of intra-party competition. However, they concern distinct contests with different candidates and positions, and there is no direct causal or temporal link between them."
We include the full benchmark results across all domains and metrics in Tables 2–5.

D Full Benchmark Results

We include the full benchmark results across all domains and metrics in Tables 2–5.

E Use Of AI Assistants

We did use ChatGPT as the writing assistant to help us write part of the paper. Additionally, we utilize the power of CodePilot to help us code faster. However, all the AI-generated writing and coding components assisted by AI are manually checked and modified. There is no full AI-generated content in the paper.

Method	MSE	MAE	Accuracy	Macro-F1	QWK
random	0.1459	0.3130	0.1977	0.1377	0.0071
heuristic (vol. max→min)	0.0411	0.1674	0.2860	0.0910	-0.0040
heuristic (vol. sim.)	0.1113	0.3003	0.0691	0.0414	0.0089
heuristic (time overlap)	0.0459	0.1687	0.3437	0.1913	0.1121
GPT-4o	0.0234	0.1258	0.4317	0.2978	0.4094
GPT-4o + CoT	0.0250	0.1214	0.5116	0.2621	0.2843
GPT-o3-mini	0.0253	0.1322	0.3973	0.2722	0.3415
GPT-o3-mini + CoT	0.0188	0.1147	0.4561	0.2411	0.3238
Meta-Llama3-70B	0.0377	0.1532	0.3847	0.2543	0.4084
Meta-Llama3-70B + CoT	0.0327	0.1518	0.3400	0.2593	0.3887
DeepSeek-V3	0.0250	0.1291	0.4383	0.3100	0.4105
DeepSeek-V3 + CoT	0.0236	0.1286	0.4006	0.2752	0.3697
DeepSeek-R1	0.0830	0.2600	0.0940	0.0807	0.0786
DeepSeek-R1 + CoT	0.0512	0.1996	0.1959	0.1585	0.1162
Qwen2-72B	0.0309	0.1426	0.3800	0.2292	0.3526
Qwen2-72B + CoT	0.0338	0.1534	0.3200	0.2285	0.2749
cross-encoder (nli-deberta-v3-base)	0.0519	0.1812	0.2797	0.0969	0.1076

Table 2: **Performance on Politics domain.** Evaluation across selected metrics.

Method	MSE	MAE	Accuracy	Macro-F1	QWK
random	0.1330	0.3000	0.2019	0.1541	0.0125
heuristic (vol. max→min)	0.0878	0.2447	0.1430	0.0560	-0.0140
heuristic (vol. sim.)	0.0945	0.2645	0.1403	0.0864	0.0398
heuristic (time overlap)	0.0779	0.2274	0.2179	0.1344	0.1427
heuristic (tag overlap)	0.0152	0.1014	0.5471	0.5691	0.6999
GPT-4o	0.0252	0.1265	0.4683	0.3584	0.5227
GPT-4o + CoT	0.0256	0.1274	0.4433	0.2687	0.4447
GPT-o3-mini	0.0412	0.1549	0.4284	0.3360	0.4645
GPT-o3-mini + CoT	0.0543	0.1638	0.4632	0.3531	0.5011
Meta-Llama3-70B	0.0416	0.1640	0.3828	0.3349	0.5004
Meta-Llama3-70B + CoT	0.0364	0.1612	0.3303	0.3278	0.4733
DeepSeek-V3	0.0242	0.1230	0.4974	0.3428	0.5187
DeepSeek-V3 + CoT	0.0289	0.1402	0.3963	0.3244	0.4886
DeepSeek-R1	0.0441	0.1833	0.2206	0.1137	0.1267
DeepSeek-R1 + CoT	0.0352	0.1698	0.2320	0.1420	0.0713
Qwen2-72B	0.0336	0.1488	0.4067	0.3069	0.4024
Qwen2-72B + CoT	0.0387	0.1683	0.2867	0.2385	0.3550
cross-encoder (nli-deberta-v3-base)	0.0888	0.2483	0.1466	0.0967	0.1491

Table 3: **Performance on Cryptocurrency domain.** Evaluation across selected metrics.

Method	MSE	MAE	Accuracy	Macro-F1	QWK
random	0.1423	0.3093	0.2016	0.1759	0.0099
heuristic (vol. max→min)	0.1612	0.3197	0.1090	0.0490	-0.0030
heuristic (vol. sim.)	0.0885	0.2531	0.1780	0.1289	0.0941
heuristic (time overlap)	0.0877	0.2383	0.2157	0.2058	0.4190
heuristic (tag overlap)	0.0229	0.1298	0.4407	0.4982	0.7932
GPT-4o	0.1042	0.2558	0.1746	0.1431	0.2418
GPT-4o + CoT	0.0744	0.2209	0.2267	0.1890	0.3678
GPT-o3-mini	0.0931	0.2305	0.2840	0.2383	0.4805
GPT-o3-mini + CoT	0.1199	0.2654	0.2182	0.1922	0.3620
Meta-Llama3-70B	0.0772	0.2312	0.1813	0.1790	0.4399
Meta-Llama3-70B + CoT	0.0838	0.2404	0.1629	0.1523	0.3916
DeepSeek-V3	0.0884	0.2432	0.1678	0.1558	0.3438
DeepSeek-V3 + CoT	0.0909	0.2480	0.1500	0.1305	0.3026
DeepSeek-R1	0.0258	0.1307	0.4409	0.4599	0.3327
DeepSeek-R1 + CoT	0.0442	0.1625	0.3972	0.2620	0.3454
Qwen2-72B	0.0983	0.2422	0.2000	0.1751	0.2478
Qwen2-72B + CoT	0.1006	0.2476	0.1933	0.1625	0.2499
cross-encoder (nli-deberta-v3-base)	0.1779	0.3432	0.0916	0.0496	-0.0360

Table 4: **Performance on Sports domain.** Evaluation across selected metrics.

Method	MSE	MAE	Accuracy	Macro-F1	QWK
random	0.1268	0.2914	0.2058	0.1558	0.0077
heuristic (vol. max→min)	0.0719	0.2227	0.1940	0.0870	-0.0200
heuristic (vol. sim.)	0.0850	0.2504	0.1610	0.0835	0.0181
heuristic (time overlap)	0.0639	0.2063	0.2570	0.1906	0.1380
heuristic (tag overlap)	0.0175	0.1121	0.4721	0.5775	0.6283
GPT-4o	0.0219	0.1112	0.5575	0.2940	0.3522
GPT-4o + CoT	0.0346	0.1489	0.4033	0.3100	0.4149
GPT-o3-mini	0.0278	0.1344	0.4548	0.2088	0.2752
GPT-o3-mini + CoT	0.0231	0.1187	0.5451	0.2496	0.4183
Meta-Llama3-70B	0.0596	0.1970	0.3103	0.2468	0.3598
Meta-Llama3-70B + CoT	0.0470	0.1834	0.2660	0.2118	0.3397
DeepSeek-V3	0.0330	0.1456	0.4132	0.2953	0.4087
DeepSeek-V3 + CoT	0.0312	0.1450	0.3836	0.2930	0.4197
DeepSeek-R1	0.0441	0.1833	0.2206	0.1137	0.1267
DeepSeek-R1 + CoT	0.0220	0.1192	0.4636	0.1893	0.1715
Qwen2-72B	0.0430	0.1696	0.3233	0.2345	0.4127
Qwen2-72B + CoT	0.0383	0.1639	0.3200	0.2737	0.4104
cross-encoder (nli-deberta-v3-base)	0.0972	0.2604	0.1436	0.0688	0.1117

Table 5: **Performance on Election domain.** Evaluation across selected metrics.

Table 6: **Annotation scale with definitions and representative examples.** Each bin corresponds to a level of relevance used in rating event pairs.

Label Range	Definition	Example Event Pair
0.0–0.2	Unrelated; events concern different topics, entities, or timelines.	Will China invade Taiwan in 2024? vs. Karine Jean-Pierre out as Press Secretary by July 31?
0.2–0.4	Weakly related; minimal topical overlap, but no structural link.	U.S. military action against Iran in 2024? vs. Democrats win popular vote by 4–5%?
0.4–0.6	Moderately related; shared actors, parties, or contexts.	Will another candidate win NY-16 Democratic Primary? vs. Will a candidate from another party win NY Senate?
0.6–0.8	Strongly related; possible causal or strategic link.	Will Trump tweet 90+ times Oct 25–Nov 1? vs. Will Trump win 30% of Black men?
0.8–1.0	Highly related; one event entails the other.	Biden resign during his speech today? vs. Biden removed via 25th Amendment?

F Heuristic Selection Methods

To provide interpretable baselines for belief correlation reasoning, we introduce a set of heuristic scoring methods for ranking candidate event pairs. Unlike learned models, these heuristics use domain knowledge and surface-level attributes to estimate correlation scores without language understanding or reasoning. They serve as simple, zero-shot approximations to relevance or co-movement between beliefs.

Random We assign a uniform random score to each candidate event. This provides a lower-bound reference for performance and reflects the difficulty of the task in the absence of any meaningful signal.

Volume-Based Sorting We hypothesize that highly traded events are more likely to be central or influential in public discourse. For each candidate, we compute its total market trading volume (over the active time window) and use this as a relevance score. We experiment with two variants:

- **Volume Max-to-Min:** Assigns the candidate's normalized trading volume as its correlation score. Events with higher volume are assumed to be more generally relevant, independent of the base event.
- **Volume Similarity:** Computes the absolute difference in trading volume between the base and candidate events. Event pairs with more similar volumes receive higher scores, under the assumption that similarly salient events may co-occur in public discourse or exhibit belief co-activation.

Temporal Overlap We compute the degree of overlap in time between the base and candidate event windows. Events that occur in similar timeframes may be causally or contextually linked. The score is computed as the ratio of overlapping duration to union duration.

Cross-Encoder Baseline We include a strong neural retrieval baseline using the `nli-deberta-v3-base` cross-encoder. It jointly encodes event pairs and outputs a real-valued relevance score. Although trained on general-purpose sentence similarity or natural language inference tasks, it often captures surface-level lexical or semantic overlap, making it a competitive 0-hop semantic baseline.

G Human Evaluation of Heuristic Scoring

G.1 Setup

Objective and Sampling. To assess whether our heuristic scoring function aligns with human intuition, we conducted an annotation study over 50 event pairs. These pairs were drawn evenly across five correlation levels (very weak to very strong) according to the algorithmic relevance scores described in Section §5.2. This stratified sampling ensured that the full range of belief correlation strengths was represented, enabling consistent evaluation across relevance levels.

Annotators and Conditions. Three annotators, who were NLP researchers involved in the project, participated in the study. While familiar with the

modeling setup, they lacked domain-specific expertise in forecasting or geopolitical reasoning. Annotations were conducted non-blind: annotators shared the same rubric and examples to guide their judgments

G.2 Annotation Protocol

Rubric Development and Scoring Process. Prior to annotation, the three annotators collaboratively developed a shared rubric to define five levels of belief correlation, ranging from unrelated to highly related. This rubric was iteratively refined through internal calibration rounds, ensuring that all annotators applied consistent semantic and causal reasoning. During annotation, each annotator independently rated all 50 event pairs on a continuous scale from 0.0 to 1.0 using the agreed rubric. Table 6 summarizes the scoring bins and includes representative examples for each level.

Label Aggregation and Annotation Conditions. Although annotators shared a rubric, the annotation process itself was conducted independently without real-time coordination. Final labels were aggregated by majority vote; in cases of complete disagreement, we averaged the three scores. To prevent bias, annotators were shown only the event texts, without access to belief trajectories, model predictions, or algorithmic scores. This ensured that all judgments reflected semantic reasoning alone.

Annotator Agreement. We evaluate inter-annotator reliability using both pairwise Pearson correlations and intra-class correlation (ICC). As shown in Table 7, pairwise Pearson scores range from 0.752 to 0.814, indicating strong linear consistency among annotators. The highest alignment is observed between Annotators A and B (0.814), while A and C show slightly lower but still robust agreement (0.752). To complement this, we compute ICC(2,1) under a two-way random effects model, yielding a value of 0.771. This reflects substantial agreement across annotators and confirms the reliability of the human labels as a benchmark for model alignment.

G.3 Alignment with Heuristic Model

To measure how well the heuristic score $S(A, B)$ matches human judgment, we compute the Pearson correlation between model predictions and the aggregated human labels. The resulting correlation of $\rho = 0.689$ (Table 8) indicates strong alignment between the scoring function and human reasoning.

Table 7: **Inter-annotator agreement.** Pearson correlation coefficients between annotators.

	Annotator A	Annotator B	Annotator C
Annotator A	1.000	0.814	0.752
Annotator B	0.814	1.000	0.798
Annotator C	0.752	0.798	1.000

Table 8: **Model-human alignment.** Pearson correlation between the heuristic score and human annotations.

Method	Pearson Correlation
Heuristic score $S(A, B)$	0.689

H Detailed Performance Degradation After Cutoff

I Demo Interface Overview

We build a web-based demo to showcase how our system connects real-time news and prediction market data. The interface allows users to explore forecastable events, understand model-generated reasoning, and vote on likely outcomes. Below, we walk through its key components.

Main Event Grid. Upon entering the demo (Figure 8), users see a grid of active prediction questions. Each card displays an event (*e.g.*, "Will X and Truth Social merger be announced before August?") along with real-time probability estimates for each outcome (Yes/No), sourced from Polymarket. Users can filter events by domain (*e.g.*, politics, crypto) via the dropdown menu. Clicking on the "News" tab navigates to a dedicated news feed page. Selecting an individual event card leads to a detailed view for reasoning and voting.

News Integration. The "News" section (Figure 9) presents a chronological list of recent headlines. Clicking on any headline redirects users to the original article. Users can also expand or collapse a card by clicking the dropdown triangle on the right. When expanded, the card reveals any prediction events automatically identified as semantically or causally related to the article, bridging news and belief markets.

Detailed Event View. When clicking on a grid cell, users are taken to a dedicated page for that

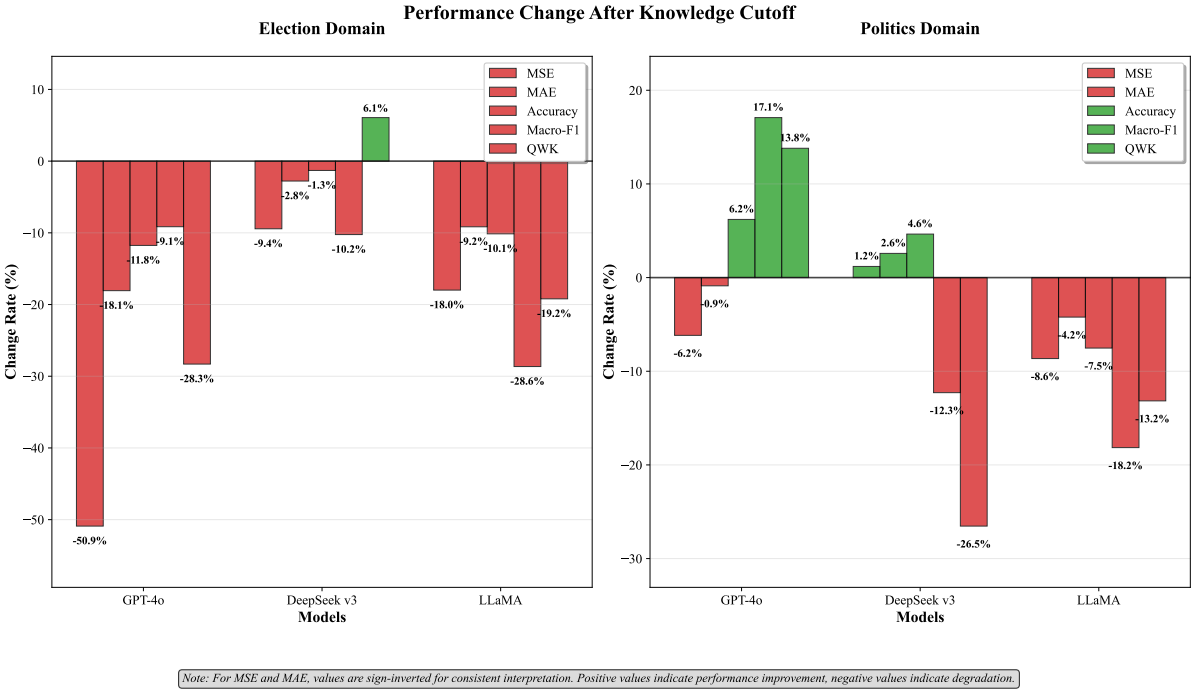

Figure 7: **Performance change after knowledge cutoff across domains and models.** Bars show the relative change in evaluation metrics on post-cutoff event pairs, compared to pre-cutoff ones. For metrics like MSE and MAE, values are sign-inverted to ensure a consistent interpretation, where negative values indicate degraded performance. GPT-4o shows a substantial decline across most metrics in the election domain, while performance remains more stable in the politics domain.

prediction question (Figure 10). Here, they can select an outcome and choose from a list of candidate reasons generated by an LLM. These explanations help users interpret possible causal mechanisms. The right panel shows a time-series chart visualizing real-time market probabilities for each option. After selecting both an outcome and a reason, users can vote to register their belief.

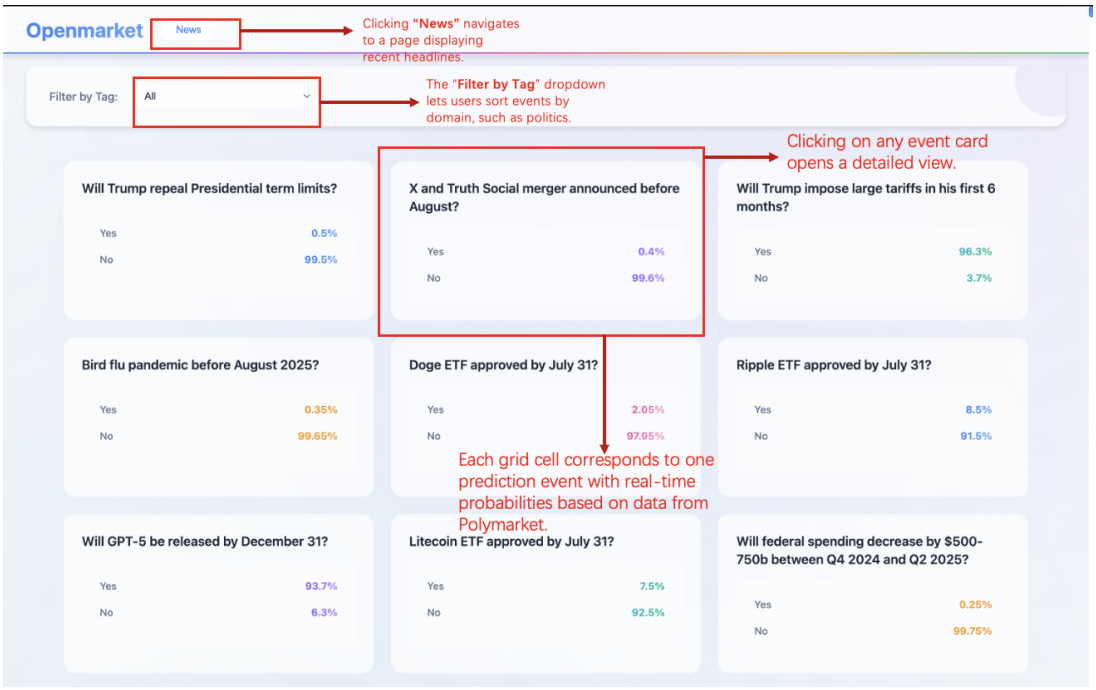

Figure 8: Main interface with real-time prediction events. Cards show current market probabilities and are filterable by topic.

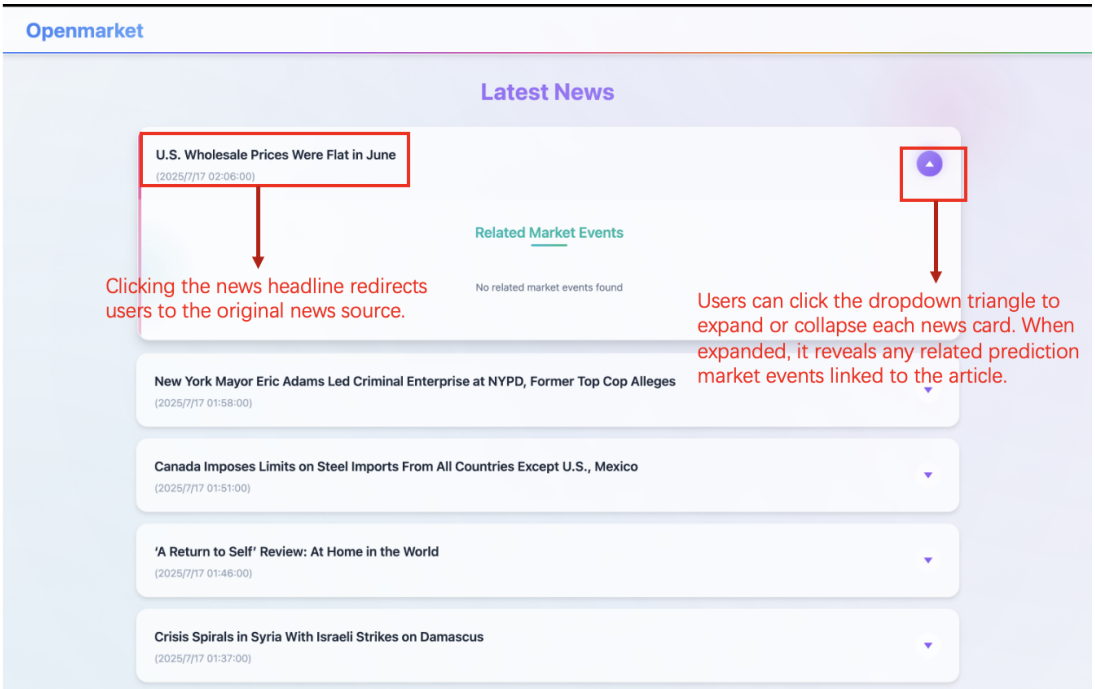

Figure 9: News page interface. Each news item links to the source and may surface relevant market events.

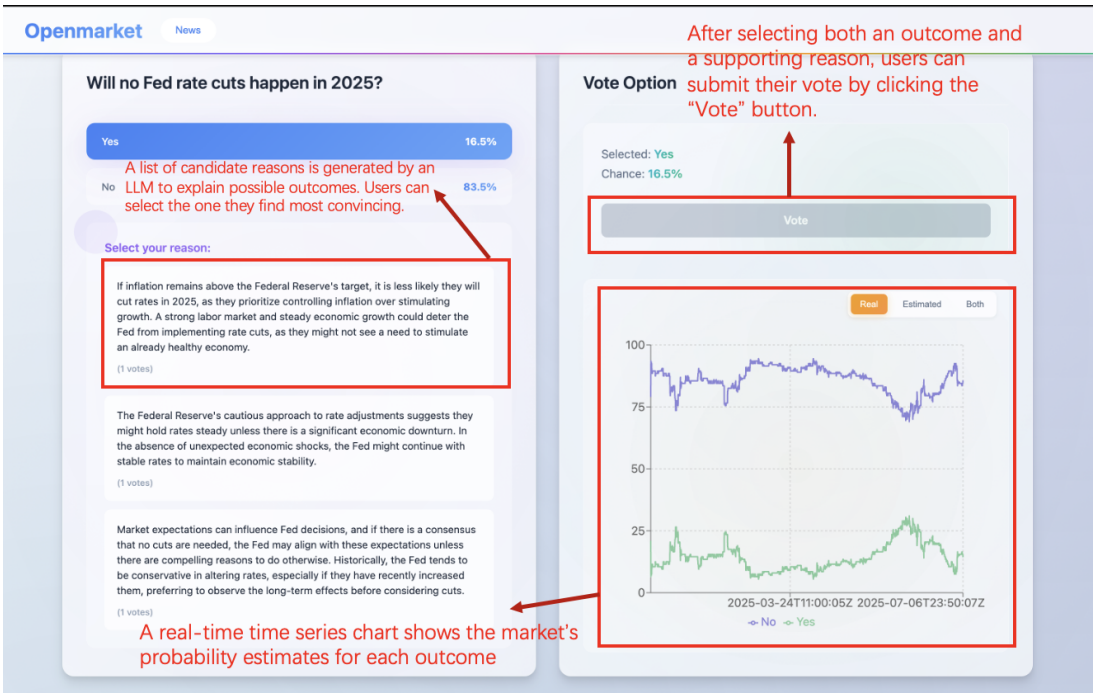

Figure 10: Detailed view of a prediction event. Users select an outcome and reason, then submit their vote.