Can Large Language Models Uncover the Structure of Social Opinions?

Anonymous ACL submission

Abstract

Understanding how opinions on different issues evolve together is essential for modeling collective intelligence, yet this remains underexplored due to the absence of standardized benchmarks. We introduce the concept of an opinion graph, where nodes represent social opinions on real-world events (e.g., presidential elections, stock predictions) and edges capture pairwise relationships between them. Building on this, we present OPINIONBENCH, a new benchmark designed to evaluate whether large language models (LLMs) can uncover the hidden structure within evolving social opinions. Constructed from Polymarket prediction markets, OPINIONBENCH labels event pairs using time-series co-movement, semantic similarity, and metadata, followed by human validation. Experiments show that (1) LLMs consistently outperform baselines in identifying opinion correlations across domains, and (2) LLMs can infer the underlying graph structure through edge prediction. OPINIONBENCH provides a challenging testbed for assessing LLMs' ability to capture complex patterns of social opinion co-evolution.

1 Introduction

Social opinions represent an individual's subjective perspective about uncertain future events—for example, presidential election outcomes, economic trends, or technological breakthroughs. Each person holds a wide range of such opinions shaped by their education, experiences, and social context. These opinions encode people's internal beliefs and expectations about how the world will unfold and serve as the building blocks of collective reasoning and societal decision-making. Social opinions exhibit two key structural features: (1) **Correlation** – Social opinions are not formed or updated in isolation. Multiple opinions are often highly correlated and can shift together when new information emerges. For example, a major poll showing a surge in support for Donald Trump may simultaneously update beliefs about "Will Trump win the 2024 election?" and "Will Republicans control the Senate?", as the two events are tightly connected. Likewise, an unexpected Federal Reserve policy announcement can jointly influence opinions about interest rates and Bitcoin prices, reflecting shared macroeconomic sentiment (Li et al., 2024; Lee et al., 2025; Radivojevic et al., 2024). (2) **Transductivity** – Beyond pairwise edge relationships, if an opinion A is related to B, and B to C, individuals may implicitly associate A with C. Such multi-hop dependencies allow information or belief changes to propagate indirectly across cross-domain opinions (Zhu et al., 2003).

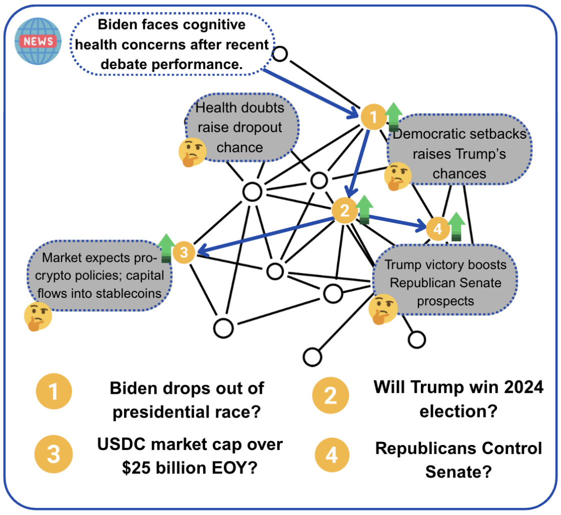

NEWS
Biden faces cognitive health concerns after recent debate performance.
Health doubts raise dropout chance
Democratic setbacks raises Trump's chances
Market expects pro-crypto policies; capital flows into stablecoins
Trump victory boosts Republican Senate prospects
1
2
3
4
1 Biden drops out of presidential race?
2 Will Trump win 2024 election?
3 USDC market cap over $25 billion EOY?
4 Republicans Control Senate?

Figure 1: **An example of the opinion graph**. In the opinion graph, each node represents a social opinion (data collected from PolyMarket), and edges denote highly possible reasoning rationale between two social opinions (data constructed with feature design and verified by humans). Similar social opinions are closely correlated with each other and form a graph structure. When the news appears, multiple social opinions in the graph would co-evolve accordingly.

These properties make it natural to represent social opinions using an **opinion graph**. In this graph, each node corresponds to a specific social opinion toward a real-world event, and each edge represents a relationship between two opinions, such as temporal co-movement, semantic similarity, or causal alignment(Kazemi et al., 2020). Because opinions are often correlated and exhibit transductive dependencies, the resulting graph naturally contains clusters and higher-order structures(Fortunato, 2010), where groups of related opinions cohere into tightly connected subgraphs. This graph-based formulation for social opinions provides a principled way to model how opinions interact locally while also capturing how information can propagate globally through complex multi-hop connections(Hegselmann and Krause, 2002).

Identifying and predicting these latent structures within social opinions is crucial for understanding collective dynamics and forecasting societal change. Structural patterns reveal how shocks can propagate through interconnected opinions, potentially amplifying into large-scale shifts—a "butterfly effect" at the level of public belief. Mapping these structures enables applications such as detecting emerging narratives, anticipating coordinated opinion shifts during crises, modeling systemic financial or political risks, and improving the interpretability of large language models analyzing social behavior (Kolajo et al., 2022; Minnema et al., 2023; Glandt et al., 2021; Wang et al., 2024; Cann et al., 2023; Deng et al., 2021; Peng et al., 2021). This raises a central research question: *Can we leverage LLMs to scalably discover the hidden structure inside a large number of social opinions?* Addressing this question would open up new opportunities for analyzing collective reasoning at unprecedented scale and granularity.

To address this question, we first collect high-quality opinion graphs from SWM (Anonymous, 2025), a dataset compiled from Polymarket[1] , a decentralized forecasting platform. Each node corresponds to a specific social opinion toward a real-world event, and each edge is labeled based on a combination of time-series co-movement and semantic similarity, with labels generated automatically and validated against human annotations to ensure reliability. We then apply LLMs to perform pairwise edge prediction, enabling us to reconstruct the entire opinion graph from local edge-level inferences and assess whether models can capture both local correlations and global structural patterns.

Our findings reveal two conclusions: (1) LLMs can accurately predict edges in the opinion graph, consistently outperforming heuristic and neural baselines. For example, GPT-4o achieves QWK scores above 0.52 on Cryptocurrency domain, significantly surpassing all heuristic methods. (2) Beyond local edges, LLMs can also recover the global structure of the opinion graph: the predicted graphs exhibit clustering patterns and structural alignments that closely match the ground truth, indicating that LLMs implicitly capture transductive and higher-order regularities in social opinions.

2 Related Work

Social opinion dynamics. Understanding the dynamics of social opinions 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) introduces 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 social opinion understanding. However, many of these studies rely on synthetic setups or structured event representations, limiting their applicability to noisy, ambiguous real-world social opinions. Our work differs by introducing a realistic evaluation task constructed from real-world market data, enabling systematic measurement of LLMs' ability to identify social opinion 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 frame-

[1] https://polymarket.com/

work (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 opinions, capturing co-occurrence, semantic relevance, or 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

Definition of opinion graphs. We represent the relationships among social opinions using an opinion graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is the set of nodes corresponding to individual social opinions, and $\mathcal{E}$ is the set of edges capturing meaningful rationales between pairs of opinions. This graph provides a unified representation of how collective expectations about different real-world events are connected through semantic, temporal, or causal dependencies.

Social opinions as nodes. Each node $v \in \mathcal{V}$ represents a *social opinion* toward a real-world event $e \in \mathcal{E}$. Formally, a node is defined as $v_e := \left(q_e, \{p_t^e\}_{t=1}^{T_e}\right)$, where q_e is the natural language description of the event, and p_t^e is the time series of daily market-implied probabilities extracted from Polymarket. This time series reflects the evolving collective belief about the event, while q_e provides semantic context. In our framework, only the textual descriptions are used for model inference, whereas temporal signals are utilized for constructing and validating the ground-truth correlations.

Reasoning rationales as edges. Each edge $(v_i, v_j) \in \mathcal{E}$ represents a *reasoning rationale* that links two social opinions. A rationale provides a textual explanation describing why the two opinions are related, such as shared semantic content or correlated temporal dynamics. In other words, edges encode interpretable relational explanations that justify why two opinions should be connected. Because social opinions are often correlated and exhibit transductive dependencies, the resulting opinion graph naturally forms clusters and higher-order structures.

Opinion structure prediction task. The task we defined on the opinion graph aims to reconstruct the structure of the underlying opinion graph $\mathcal{G}$ from observed social opinions $\mathcal{V}$ by evaluating the plausibility of edges between node pairs. Formally, given a set of candidate node pairs $\mathcal{P} \subseteq \mathcal{V} \times \mathcal{V}$, the goal is to learn a scoring function $s : \mathcal{P} \to \mathbb{R}$, where $s(v_i, v_j)$ quantifies the predicted rationale strength between two social opinions v_i and v_j. High scores correspond to pairs that are strongly connected through semantic, temporal, or causal reasoning, while low scores indicate weak or spurious associations. By applying a threshold τ to the predicted scores,

$$\hat{\mathcal{E}} = \{(v_i, v_j) \in \mathcal{P} \mid s(v_i, v_j) \geq \tau\}, \qquad (1)$$

We aim to obtain a filtered set of edges that define the predicted opinion graph $\hat{\mathcal{G}} = (\mathcal{V}, \hat{\mathcal{E}})$ that should be as similar as possible with the ground-truth one.

4 Constructing the Opinion Graph

To benchmark the social opinion correlation task, we construct a dataset based on SWM (Anonymous, 2025), derived from Polymarket. In this section, we first explain how social opinion pairs are selected, then describe our procedure for collecting ground-truth relationship labels.

4.1 Social Opinion Node Collection

Not all markets offer informative or reliable signals for social opinion reasoning. To ensure that the included events reflect collective crowd social opinions rather than noise, we apply two filters: one based on trading volume, the other on volatility of social opinion movement.

Volume filter. Markets with very low trading volume are often driven by isolated trades and do not reflect meaningful aggregation of public social opinion. 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 a sufficient probability of movement. A flat probability series provides little statistical signal. By imposing a minimum volatility threshold, we ensure that the probability series contains enough variation to make the correlation test meaningful. Details are available in Appendix §C.1.

4.2 Reasoning Rationale Edge Collection

To construct ground-truth edges for the opinion graph, we adopt a hybrid scoring framework that

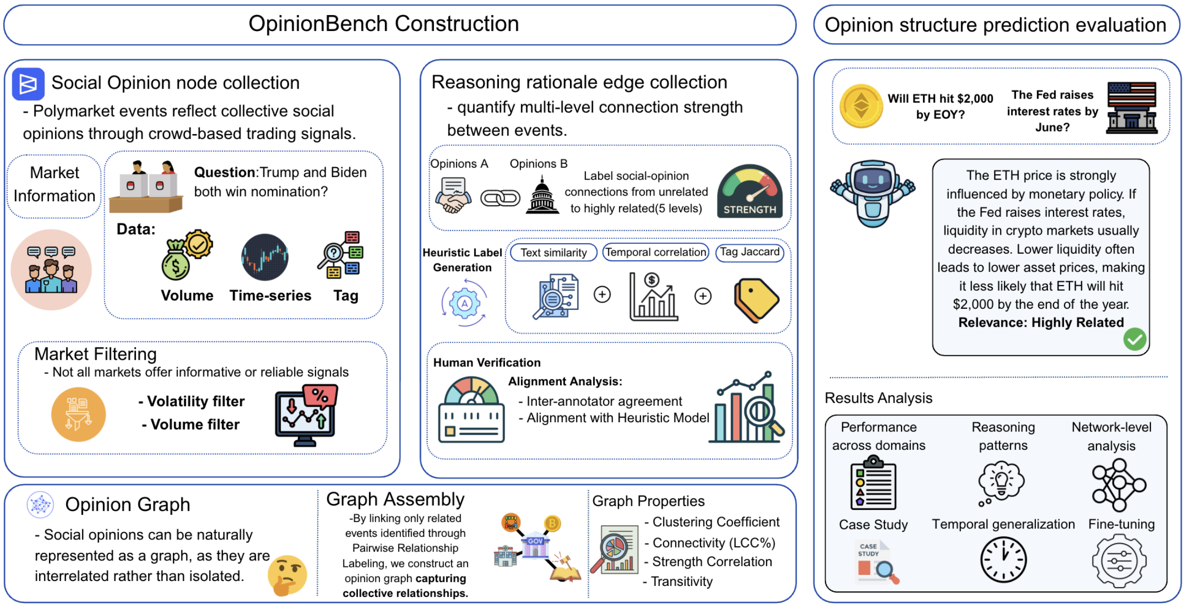

OpinionBench Construction
Opinion structure prediction evaluation
Social Opinion node collection
- Polymarket events reflect collective social opinions through crowd-based trading signals.
Market Information
Question:Trump and Biden both win nomination?
Data:
Volume
Time-series
Tag
Market Filtering
- Not all markets offer informative or reliable signals
- Volatility filter
- Volume filter
Reasoning rationale edge collection
- quantify multi-level connection strength between events.
Opinions A
Opinions B
Label social-opinion connections from unrelated to highly related(5 levels)
STRENGTH
Heuristic Label Generation
Text similarity
Temporal correlation
Tag Jaccard
Human Verification
Alignment Analysis:
- Inter-annotator agreement
- Alignment with Heuristic Model
Opinion Graph
- Social opinions can be naturally represented as a graph, as they are interrelated rather than isolated.
Graph Assembly
-By linking only related events identified through Pairwise Relationship Labeling, we construct an opinion graph capturing collective relationships.
GOV
Graph Properties
- Clustering Coefficient
- Connectivity (LCC%)
- Strength Correlation
- Transitivity
Will ETH hit $2,000 by EOY?
The Fed raises interest rates by June?
The ETH price is strongly influenced by monetary policy. If the Fed raises interest rates, liquidity in crypto markets usually decreases. Lower liquidity often leads to lower asset prices, making it less likely that ETH will hit $2,000 by the end of the year.
Relevance: Highly Related
Results Analysis
Performance across domains
Reasoning patterns
Network-level analysis
Case Study
CASE STUDY
Temporal generalization
Fine-tuning

Figure 2: **Overview of the OpinionBench pipeline.** (a) Social opinion node collection from *Polymarket* with market filtering for reliability. (b) Reasoning rationale edge collection quantifies multi-level connections (5 levels) via semantic, temporal, and tag similarities, verified by human annotation. (c) Opinion structure prediction evaluation measures models' ability to capture opinion relevance across domains, reasoning types, temporal generalization, and network structure.

integrates multiple complementary signals. This approach captures both semantic and temporal aspects of social opinion relationships, allowing us to go beyond surface similarity and identify deeper correlations between event pairs. For each pair (A, B), we compute four interpretable feature scores, s_1 through s_4, and combine them into a single composite relevance score.

Feature design. The four features capture distinct yet complementary aspects of social opinion relationships. (1) *Change-point synchrony* s_1 detects statistically significant shifts in each event's social opinion trajectory and measures how frequently these shifts occur close together in time, capturing coordinated changes in collective expectations. (2) *Tag Jaccard similarity* s_2 compares Polymarket metadata tags, identifying topical overlap and shared discourse contexts. (3) *Minimum time gap* s_3 measures how closely in time the opinion shifts of two events occur, providing a soft measure of temporal proximity. (4) *Textual similarity* s_4 computes embedding-based semantic similarity between event descriptions, capturing lexical and conceptual relatedness beyond metadata. Full details of these feature definitions are provided in Appendix §C.2.

Edge label construction. The four feature scores are linearly combined into a single heuristic correlation score $S(A, B) = \sum_{i=1}^{4} \gamma_i s_i(A, B)$. The resulting scores are discretized into five ordinal levels (very weak, weak, medium, strong, very strong), which serve as ground-truth edge labels in our prediction task. By integrating heterogeneous information—temporal dynamics, topical metadata, and textual semantics—this framework produces high-quality, interpretable rationales for edge construction in the opinion graph. We also collect text-based rationale with state-of-the-art LLMs (GPT-4o) as part of the edge attribute.

Human verification. To assess the quality of the heuristic labels, we conduct a human annotation study on a representative subset of 200 event pairs, sampled uniformly across the five correlation levels. Three annotators independently rated each pair based on textual semantics and related news, without access to the underlying time series or model predictions. Inter-annotator agreement was strong, with pairwise Pearson correlations ranging from 0.739 to 0.840 and an intraclass correlation (ICC) of 0.777. Moreover, the aggregated human judgments were well aligned with the heuristic scores, yielding a Pearson correlation of 0.697. These results confirm that the scoring framework reflects intuitive assessments of social opinion correlation. Full details of the protocol and annotation examples are provided in Appendix §I.

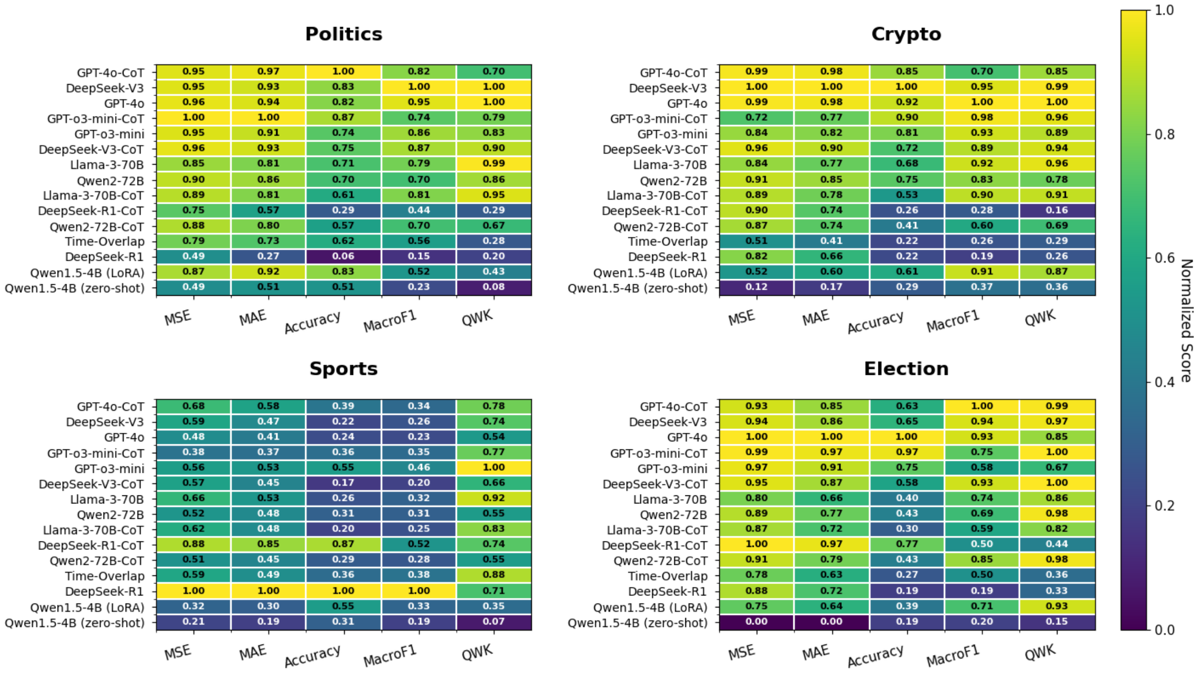


Figure 3: **Performance on opinion graph edge scoring across four domains: Politics, Crypto, Sports, and Election. Only the strongest baseline is included for comparison.** Each heatmap shows model performance across five metrics: MSE, MAE (lower is better before negation), and Accuracy, Macro-F1, QWK (higher is better). To make metrics comparable, error metrics are first negated so that higher values indicate better performance, and then 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.

5 Predicting the Opinion Graph

We use LLMs to perform opinion graph discovery by jointly predicting a *rationale* and a *score* for each candidate edge. Given a pair of social opinions $(v_i, v_j) \in \mathcal{P}$, represented by their event descriptions and opinion trajectories, the model is prompted to reason about their relationship and generate two outputs simultaneously: a textual explanation describing why the two opinions may be correlated, and a scalar score indicating the predicted strength of this relationship. Formally, for each node pair (v_i, v_j), the LLM implements a mapping $\Phi_{\text{LLM}} : \mathcal{P} \to \mathcal{R} \times \mathbb{R}$, where $\mathcal{R}$ denotes the space of textual rationales. The output

$$\Phi_{\text{LLM}}(v_i, v_j) = \big(r_{i,j}, , s_{i,j}\big) \qquad (2)$$

consists of a text-based rationale $r_{i,j} \in \mathcal{R}$ explaining the potential connection between v_i and v_j, and a score $s_{i,j} \in \mathbb{R}$ reflecting the predicted strength of their relationship.

The predicted opinion graph is then reconstructed by applying a threshold τ to the set of predicted scores. Specifically, we retain edges with $s_{i,j} \geq \tau$ to form the edge set following Eq. 1, which defines the predicted graph $\hat{\mathcal{G}} = (\mathcal{V}, \hat{\mathcal{E}}_{\mathcal{G}})$. The generated rationales provide interpretable justifications for each predicted edge, while the scores enable scalable graph recovery through pairwise evaluation. This joint rationale–score prediction framework allows LLMs to explicitly reason about social opinion relationships.

6 Experimental Settings

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 [2] (`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 models including GPT model family (Hurst et al., 2024), Qwen2 model family (Team et al., 2024), LLaMA model famility (Touvron et al., 2023) and DeepSeek-R1 (Guo et al., 2025) using a rationale-based classification setup. Given two social opinion titles and

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

descriptions, the LLM first generates a structured rationale explaining links between events, then selects relevance levels. For comparison, we also include a variant where models directly predict the relevance label without explicit reasoning.

Edge-level evaluation metrics. We evaluate performance at both the *edge* and *graph* levels. For edge-level metrics, we include MSE, MAE, Accuracy, Macro-F1, and Quadratic Weighted Kappa (QWK). All of them are aimed for evaluating the classification performance. Typically, QWK is a classical ranking correlation metric computed as

$$\mathrm{QWK} = 1 - \frac{\sum_{i,j} W_{ij} O_{ij}}{\sum_{i,j} W_{ij} E_{ij}}, \tag{3}$$

where O and E are the observed and expected rating matrices, and W is the quadratic weight matrix.

Graph-level evaluation metrics. Graph-level metrics include: (1) Largest Connected Component ratio (LCC), defined as LCC $= \frac{|V_{\max}|}{|V|}$; (2) Clustering coefficient, $C = \frac{1}{n}\sum_i \frac{2T_i}{k_i(k_i-1)}$, where T_i is the number of triangles through node i; (3) Strength correlation, $\rho = \mathrm{corr}\left(s_i^{\mathrm{GT}}, s_i^{\mathrm{Pred}}\right)$, where s_i is the weighted degree; (4) Transitivity, Trans $= \frac{3\times \text{\# triangles}}{\text{\# connected triples}}$. These metrics jointly assess the global structural fidelity.

7 Experimental Results

LLMs generally achieve strong edge prediction performance across domains. In Figure 3, we evaluate model performance across four domains—Politics, Election, Crypto, and Sports—using both classification (Accuracy, F1, QWK) and regression (MSE, MAE) metrics. Overall, GPT-4o + Chain-of-Thought (Wei et al., 2022) achieves the strongest and most stable results, outperforming smaller variants (e.g., GPT-o3-mini) and competitive open-source baselines (e.g., Meta-Llama-3-70B, Qwen2-72B) in most settings. While minor domain-specific variations exist, the trend highlights that LLMs can effectively infer event relationships from textual descriptions alone, without relying on metadata. Heuristic baselines based on time overlap perform markedly worse, further emphasizing the advantage of language-based reasoning.

Domain	LCC	Clustering	Corr.	Trans.
Politics	92.6→82.8	0.088→0.087	0.983	78.0%***
Election	98.5→99.3	0.257→0.276	0.991	90.8%***
Crypto	99.4→97.8	0.131→0.104	0.970	87.8%***
Sports	91.6→45.0	0.077→0.034	0.954	90.0%***

Table 1: **LLM-based predictions preserve network structure.** LCC represents the percentage of the largest connected component ratio. Clustering represents the local cohesion. Corr. represents edge-weight alignment. Trans. represents transitivity. X → Y represents the ground-truth one → the predicted one. A small difference between X and Y indicates the high fidelity for prediction. *** indicates $p < 0.05$.

Rationale generation helps most in domains requiring complex reasoning. CoT prompting provides the greatest benefit in domains like Politics and Election. In these settings, models like GPT-4o-CoT and GPT-o3-mini-CoT achieve notable gains in regression accuracy and ranking consistency, reflecting their ability to uncover indirect dependencies between events. However, these benefits are domain-specific: CoT improves calibration but not classification metrics in Politics, and primarily boosts regression in Election. In contrast, in Crypto and Sports—where relationships are largely surface-level—CoT often introduces unnecessary noise, leading to drops in Accuracy. Overall, rationale generation enhances performance when reasoning complexity is high, but can be detrimental when simple textual cues suffice.

Ground-truth opinion graphs show clear higher-order structure. Although the benchmark labels pairwise correlations, these links form cohesive multi-event networks. As shown in Table 1, across all domains, 92–99% of events belong to a single connected component. Clustering coefficients are 12–29× higher than random (Politics: 0.088 vs. 0.003; Election: 0.257 vs. 0.012), revealing abundant triangular motifs. Louvain detection yields modularity of 0.50–0.77, well above random partitions. These results show that pairwise correlations *naturally organize into structured, high-order opinion graphs*, allowing pairwise evaluation to probe broader belief dynamics.

LLM-predicted graphs preserve key structural patterns. We then assess whether model-predicted graphs retain the structural properties of ground truth. As shown in Table 1, in *Election* and *Crypto*, connectivity and clustering closely match ground truth (<2% difference), with identical modularity in Crypto (0.578). Strength correlations (weighted degrees) exceed 0.95 across all domains. Transitivity, reflecting logical consistency, remains high (76–92%) and far above random baselines (≈41%).

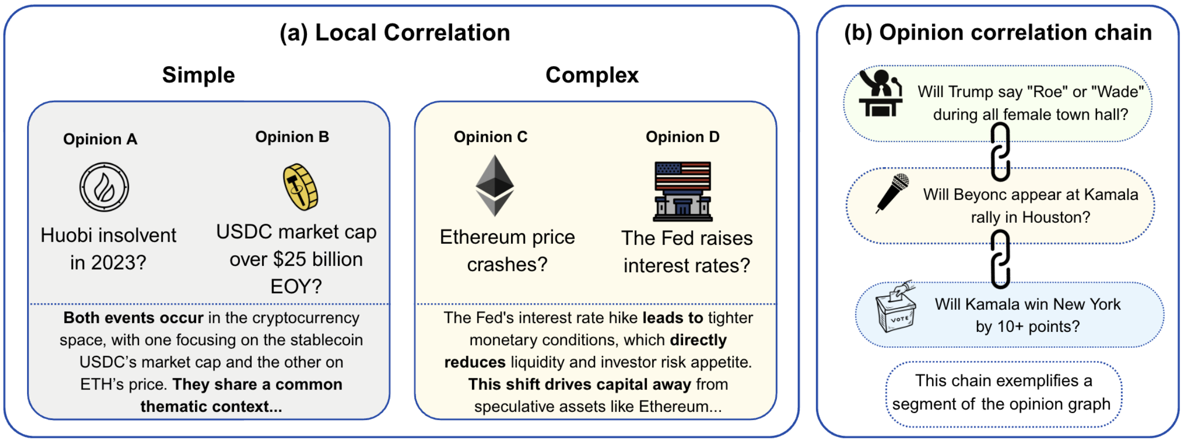

(a) Local Correlation
Simple
Complex
Opinion A
Opinion B
Huobi insolvent in 2023?
USDC market cap over $25 billion EOY?
Both events occur in the cryptocurrency space, with one focusing on the stablecoin USDC's market cap and the other on ETH's price. They share a common thematic context...
Opinion C
Opinion D
Ethereum price crashes?
The Fed raises interest rates?
The Fed's interest rate hike leads to tighter monetary conditions, which directly reduces liquidity and investor risk appetite. This shift drives capital away from speculative assets like Ethereum...
(b) Opinion correlation chain
Will Trump say "Roe" or "Wade" during all female town hall?
Will Beyonc appear at Kamala rally in Houston?
Will Kamala win New York by 10+ points?
This chain exemplifies a segment of the opinion graph

Figure 4: **Case study for edges between social opinions and the path they form**. (a) Pairwise relevance cases: GPT-4o+CoT explains both simple and complex social opinion pairs. (b) An example social opinion chain constructed from high-relevance event pairs (score > 0.6), with links inferred by GPT-4o using CoT reasoning.

Politics shows moderate degradation (83% vs. 93% GT connectivity), while *Sports* has lower connectivity (45%) but exceptionally high transitivity (98%), indicating locally consistent yet globally fragmented predictions. Overall, LLMs preserve network structure in semantically coherent domains and maintain local consistency even when global alignment weakens.

8 Case Study

Beyond statistical evaluation results on edge and graph prediction, we conduct microscopic case studies to examine how well LLM-based predictors reconstruct opinion graphs at two structural levels: *edges* and *paths*. This hierarchical perspective reveals how LLM-based rationale prediction can be used to make local predictions scale into global graph structure discovery. Additional examples are provided in Appendix §E.

Edge level: local correlations and reasoning rationales. Edges in the opinion graph can be supported by either simple or complex reasoning. Figure 4(a, left) shows a **simple** case, where two opinions share a clear topical overlap (e.g., both concern cryptocurrency but focus on different coins). The model's rationale is straightforward—it notes that they "share a common thematic context." In contrast, Figure 4(a, right) presents a **complex** case, where the connection depends on multi-step temporal or causal reasoning. For example, a Federal Reserve interest rate hike can tighten monetary conditions and redirect capital away from speculative assets like Ethereum, linking two seemingly distant opinions. Together, these simple topical links and complex causal chains form the backbone of the opinion graph structure.

Path level: chaining social opinions and the butterfly effect. Building on both simple and complex edges, Figure 4(b) shows how related opinions can form extended chains (paths in the opinion graph) through shared entities or causal dependencies. For example, a sequence linking Trump's campaign statements to rally participation and ultimately to electoral outcomes illustrates how opinions evolve across interconnected contexts. Simple edges link central events to related ones, while complex edges propagate these connections across domains, creating reasoning paths that reflect the *butterfly effect*—where local signals spread through institutional or topical structures to shape broader expectations. These case studies demonstrate that LLMs can uncover such latent chains and generate coherent, evolving rationales that connect multiple social opinions, ultimately revealing the hidden structure underlying collective beliefs.

9 Discussion

To understand when and why LLMs succeed at uncovering social opinion structures, we focus on four key factors. We examine how performance varies across domains with different semantic structures (RQ1) and investigate the reasoning strategies models use to make predictions (RQ2). We then assess whether scaling and fine-tuning smaller models can improve efficiency without sacrificing accuracy (RQ3), and evaluate the role of knowledge recency in generalizing to unseen events (RQ4).

RQ1: What domain are LLMs good at?

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.

RQ2: How do LLMs reason for prediction? To better understand what types of relationships LLMs rely on when judging social opinion correlation, we analyze their 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 *causal inference*.

RQ3: Does training on LLMs improve performance for edge prediction? As shown in Figure 3, fine-tuning significantly boosts the performance of smaller models for opinion graph prediction. We fine-tuned Qwen1.5-4B on 500 social opinion pairs and evaluated it on the same test set. The fine-tuned model shows substantial gains over its zero-shot version and becomes competitive with much larger models, achieving performance comparable to Meta-Llama3-70B in certain domains. These results suggest that smaller, specialized models can serve as efficient and effective alternatives to large general-purpose LLMs.

RQ4: Does the knowledge cutoff matter? 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 2, 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 social opinion correlation often depends on up-to-date world knowledge learned during pretraining.

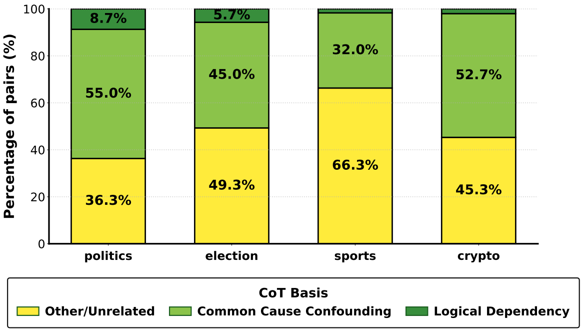


Figure 5: **Distribution of reasoning types across domains**. Each edge was labeled based on the explanation produced by the model. A majority of predictions are based on shared context (*confounding*) or loose narrative links (*CoT basis*), while only a small portion exhibit explicit logical or causal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

Model	Before↓	After↓	Δ (%)
Heuristic (Time-OL)	0.0639	–	–
LLaMA	0.0556	0.0656	+18.0
DeepSeek v3	0.0318	0.0348	+9.4
GPT-4o	0.0167	0.0252	**+50.9**

Table 2: **Results on temporal generalization**. We report MSE before/after the knowledge cutoff (Election). All models exhibit worse MSE performance on post-cutoff event pairs, highlighting challenges in temporal generalization. GPT-4o shows the highest increase. Time-OL represents time-overlapping.

10 Conclusion

In summary, social opinions form richly structured networks and make graph-based representations a natural framework for understanding collective belief dynamics. By leveraging high-quality opinion graphs and applying LLMs to edge prediction tasks, we show that LLMs not only excel at predicting pairwise relations but also recover higher-order structural patterns that closely mirror ground-truth networks. These findings demonstrate that LLMs implicitly capture the latent organization of social opinions, enabling scalable analysis of emerging belief dynamics. Looking ahead, our framework can motivate future work for better studies on opinion dynamics and social simulation.

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 and time alignment (see Section §4.2). 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.

Pairwise relation assumption. Our framework estimates the strength of social opinion correlations using pairwise relationships between events. While this design enables interpretable and scalable analysis, it does not explicitly capture higher-order dependencies among multiple events. Future work could explore multi-event or graph-based inference methods to model collective reasoning patterns that go beyond pairwise interactions.

Temporal overlap assumption. Our approach focuses on social opinion 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.

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.

Santo Fortunato. 2010. Community detection in graphs. *Physics Reports*, 486(3–5):75–174.

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.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Rainer Hegselmann and Ulrich Krause. 2002. Opinion dynamics and bounded confidence: Models, analysis and simulation. *Journal of Artificial Societies and Social Simulation*, 5(3).

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.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Seyed Mehran Kazemi, Rishab Goel, Kshitij Jain, Ivan Kobyzev, Akshay Sethi, Peter Forsyth, and Pascal Poupart. 2020. Representation learning for dynamic graphs: A survey. *Journal of Machine Learning Research*, 21(70):1–73.

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. 2024. Political-llm: Large language models in political science. *Preprint*, arXiv:2412.06864.

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.

Qwen Team. 2024. Introducing qwen1.5.

Qwen Team and 1 others. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2:3.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Haoyu Wang, Hongming Zhang, Kaiqiang Song, Dong Yu, and Dan Roth. 2024. 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.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

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.

Xiaojin Zhu, Zoubin Ghahramani, and John Lafferty. 2003. Semi-supervised learning using gaussian fields and harmonic functions. ICML'03, page 912–919. AAAI Press.

A Artifact Details

A.1 Artifact Information

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

- **Code:** A complete implementation of the pairwise social opinion correlation scoring pipeline, including preprocessing, model inference (with and without CoT 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 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 social opinion chains, visualizing social opinion 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. In addition, we fine-tune **Qwen1.5-4B** (Team, 2024) (LoRA variant) using the Hugging Face Transformers library,[4] which is an open-weight model under the Apache 2.0 License. The fine-tuning was performed on a single NVIDIA A100 GPU for approximately 10 minutes, with no large-scale computational resources required. For comparison, we include a **cross-encoder baseline** using `nli-deberta-v3-base`[5] 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.

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

[4] https://huggingface.co/Qwen/Qwen1.5-4B

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

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:**
- **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 social opinion 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 §I 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, social opinion 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 social opinion 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 Computational Resources.

The only locally fine-tuned model was **Qwen1.5-4B (LoRA)**, trained on a single NVIDIA A100 GPU (80GB) for approximately 10 minutes, corresponding to a total compute budget of ∼0.17 GPU-hours. The LoRA adapters introduce fewer than 1% of the base model's parameters. All other models were accessed via the OpenAI and Together.ai inference APIs, requiring no additional training. All runs were executed deterministically with fixed random seeds and single-threaded decoding to ensure reproducibility.

C Dataset Construction Details

C.1 Volatility filter details

We include the volatility filter here. 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

$$\frac{1}{T-w+1}\sum_{t=w}^{T}\mathbf{1}\left[\sigma_t^{(w)} \geq \gamma\right] \;\geq\; \alpha, \qquad (4)$$

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

C.2 Edge construction feature details

Feature1: Change-point synchrony. We identify time points where an event's social opinion 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 social opinions are likely to be correlated:

$$s_1(A,B) = \frac{1}{|T_A|}\sum_{t\in T_A} \nVdash \left[\exists\, t' \in T_B,\ |t-t'| < \delta\right]. \qquad (5)$$

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:

$$s_2(A, B) = \frac{|\mathcal{K}_A \cap \mathcal{K}_B|}{|\mathcal{K}_A \cup \mathcal{K}_B|}. \tag{6}$$

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 social opinion shifts in the two events occur in time. We convert this to a soft similarity score using a monotonic inverse transformation:

$$s_3(A, B) = \frac{1}{1 + \min_{t \in T_A,\, t' \in T_B} \frac{|t-t'|}{\tau}}. \tag{7}$$

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:

$$s_4(A, B) = 1 - \cos\left(\mathbf{e}_A, \mathbf{e}_B\right). \tag{8}$$

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:

$$S(A, B) = \sum_{i=1}^{4} w_i \cdot s_i(A, B) \tag{9}$$

where w_i is tuned to make the prediction highly aligned with a small set of human annotation results. The final normalized weights are: w_{sync}=0.2688, w_{jacc}=0.3652, w_{time}=0.1228, and w_{text}=0.2435.

D LLM Prompt

The prompt of the zero shot and CoT version are in Table 3

E Case Study Examples

E.1 Simple

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."

E.2 Complex

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?
Event 2: Will Kristi Noem win the U.S. 2024 Republican VP nomination?

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 3: Comparison of Original vs. CoT prompt templates

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 4–7.

F Full Benchmark Results

We include the full benchmark results across all domains and metrics in Tables 4–7.

G 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
Qwen1.5-4B (zero-shot)	0.0839	0.2109	0.2960	0.1038	0.0311
Qwen1.5-4B (fine-tuned, LoRA)	0.0351	0.1300	0.4362	0.1813	0.1752

Table 4: **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
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
Qwen1.5-4B (zero-shot)	0.1203	0.2693	0.2433	0.1691	0.1800
Qwen1.5-4B (fine-tuned, LoRA)	0.0760	0.1942	0.3592	0.3326	0.4519

Table 5: **Performance on Cryptocurrency domain.** Evaluation across selected metrics.
**Note: the "tag overlap" method was used as a feature in the creation of the ground-truth labels (see Section 5.2* and is therefore not a benchmark baseline.

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
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
Qwen1.5-4B (zero-shot)	0.1463	0.3033	0.2000	0.1290	0.0018
Qwen1.5-4B (test metrics)	0.1286	0.2788	0.2850	0.1834	0.1468

Table 6: **Performance on Sports domain.** Evaluation across selected metrics.
**Note: the "tag overlap" method was used as a feature in the creation of the ground-truth labels (see Section 5.2* and is therefore not a benchmark baseline.

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
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
Qwen1.5-4B (zero-shot)	0.2099	0.3650	0.2217	0.1162	0.0458
Qwen1.5-4B (fine-tuned, LoRA)	0.0681	0.2017	0.3058	0.2401	0.3873

Table 7: **Performance on Election domain.** Evaluation across selected metrics.
**Note: the "tag overlap" method was used as a feature in the creation of the ground-truth labels (see Section 5.2* and is therefore not a benchmark baseline.

Table 8: **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?

H Heuristic Selection Methods

To provide interpretable baselines for social opinion 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 social opinions.

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 social opinion coactivation.

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.

I Human Evaluation of Heuristic Scoring

I.1 Setup

Objective and Sampling. To assess whether our heuristic scoring function aligns with human intuition, we conducted an annotation study over 200 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 §4.2. This stratified sampling ensured that the full range of social opinion 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

I.2 Annotation Protocol

Rubric Development and Scoring Process. Prior to annotation, the three annotators collaboratively developed a shared rubric to define five levels of social opinion 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 200 event pairs on a continuous scale from 0.0 to 1.0 using the agreed rubric. Table 8 summarizes the scoring

Table 9: **Inter-annotator agreement.** Pearson correlation coefficients between annotators.

	Annotator A	Annotator B	Annotator C
Annotator A	1.000	0.840	0.739
Annotator B	0.840	1.000	0.794
Annotator C	0.739	0.794	1.000

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 social opinion 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 9, pairwise Pearson scores range from 0.739 to 0.840, indicating strong linear consistency among annotators. The highest alignment is observed between Annotators A and B (0.840), while A and C show slightly lower but still robust agreement (0.739). To complement this, we compute ICC(2,1) under a two-way random effects model, yielding a value of 0.777. This reflects substantial agreement across annotators and confirms the reliability of the human labels as a benchmark for model alignment.

I.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 $\rho = 0.697$ (Table 10) indicates strong alignment between the scoring function and human reasoning.

Table 10: **Model-human alignment.** Pearson correlation between the heuristic score and human annotations.

Method	Pearson Correlation
Heuristic score $S(A, B)$	0.697

J Detailed Performance Degradation After Cutoff

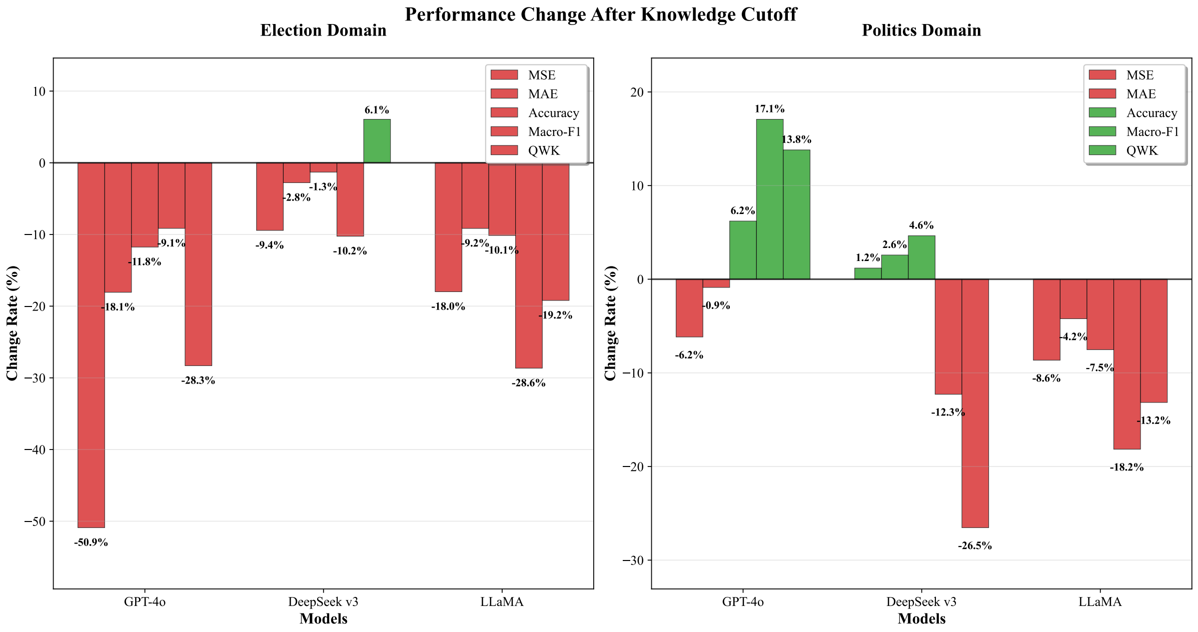
K Demo Interface Overview

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 7), 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.

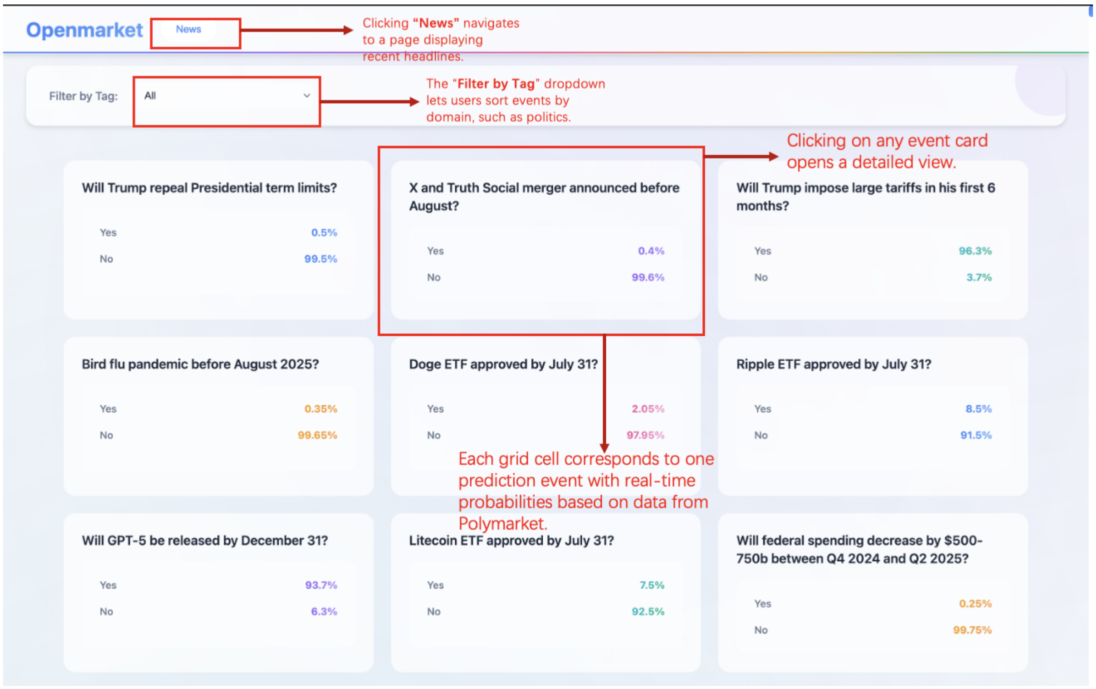
News Integration. The “News” section (Figure 8) presents a chronological list of recent headlines. Clicking on any headline redirects users to the original article. Users can also expand or collapse a card by clicking the dropdown triangle on the right. When expanded, the card reveals any prediction events automatically identified as semantically or causally related to the article, bridging news and social opinion markets.

Detailed Event View. When clicking on a grid cell, users are taken to a dedicated page for that prediction question (Figure 9). 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 social opinion.



Note: For MSE and MAE, values are sign-inverted for consistent interpretation. Positive values indicate performance improvement, negative values indicate degradation.

Figure 6: **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.



Clicking **"News"** navigates to a page displaying recent headlines.

Filter by Tag: All

The "**Filter by Tag**" dropdown lets users sort events by domain, such as politics.

Clicking on any event card opens a detailed view.

Will Trump repeal Presidential term limits?
Yes 0.5%
No 99.5%

X and Truth Social merger announced before August?
Yes 0.4%
No 99.6%

Will Trump impose large tariffs in his first 6 months?
Yes 96.3%
No 3.7%

Bird flu pandemic before August 2025?
Yes 0.35%
No 99.65%

Doge ETF approved by July 31?
Yes 2.05%
No 97.95%

Ripple ETF approved by July 31?
Yes 8.5%
No 91.5%

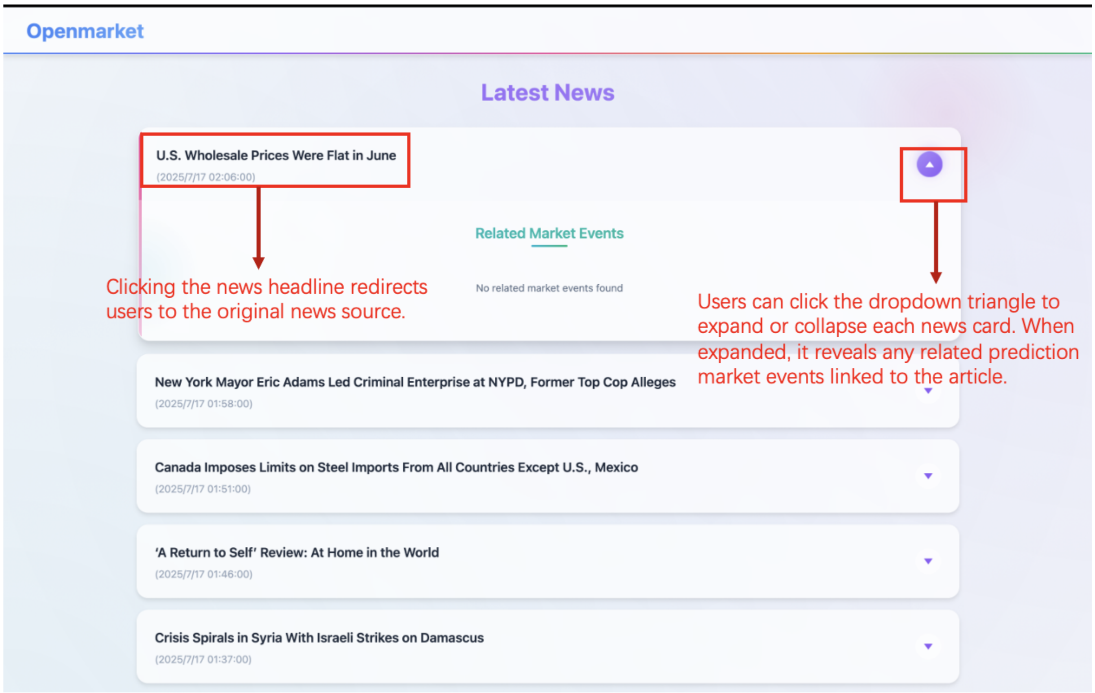
Each grid cell corresponds to one prediction event with real-time probabilities based on data from Polymarket.

Will GPT-5 be released by December 31?
Yes 93.7%
No 6.3%

Litecoin ETF approved by July 31?
Yes 7.5%
No 92.5%

Will federal spending decrease by $500-750b between Q4 2024 and Q2 2025?
Yes 0.25%
No 99.75%

Figure 7: Main interface with real-time prediction events. Cards show current market probabilities and are filterable by topic.



Openmarket

Latest News

U.S. Wholesale Prices Were Flat in June
(2025/7/17 02:06:00)

Related Market Events

No related market events found

Clicking the news headline redirects users to the original news source.

Users can click the dropdown triangle to expand or collapse each news card. When expanded, it reveals any related prediction market events linked to the article.

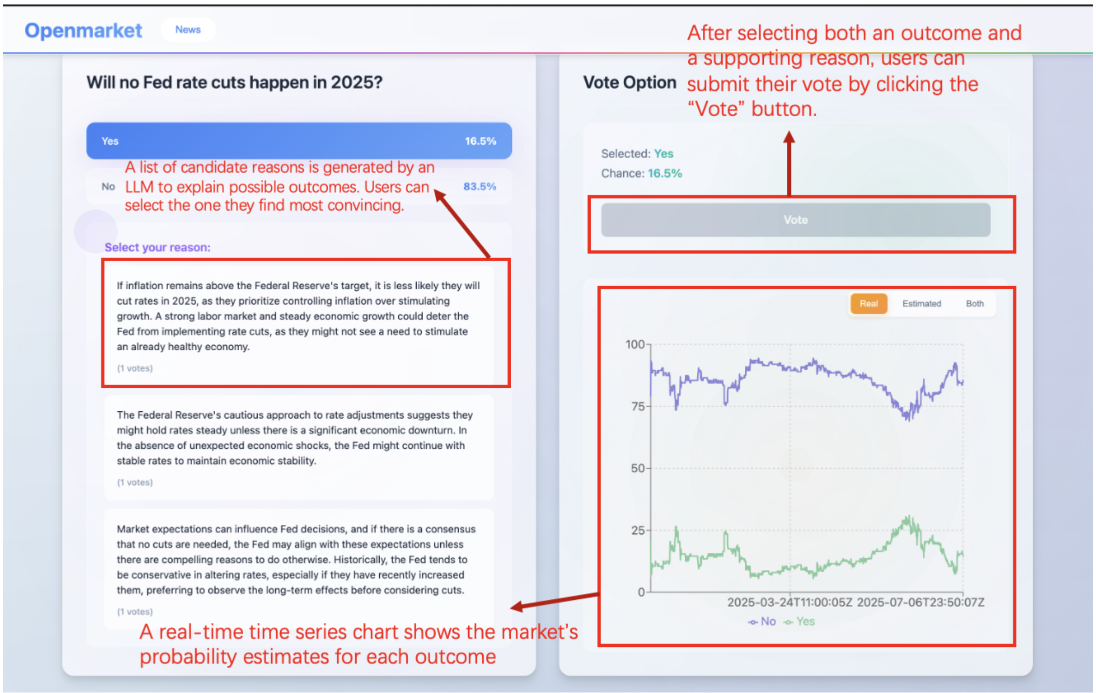
New York Mayor Eric Adams Led Criminal Enterprise at NYPD, Former Top Cop Alleges
(2025/7/17 01:58:00)

Canada Imposes Limits on Steel Imports From All Countries Except U.S., Mexico
(2025/7/17 01:51:00)

'A Return to Self' Review: At Home in the World
(2025/7/17 01:46:00)

Crisis Spirals in Syria With Israeli Strikes on Damascus
(2025/7/17 01:37:00)

Figure 8: News page interface. Each news item links to the source and may surface relevant market events.



Openmarket

Will no Fed rate cuts happen in 2025?

Yes 16.5%

No 83.5%

A list of candidate reasons is generated by an LLM to explain possible outcomes. Users can select the one they find most convincing.

Select your reason:

If inflation remains above the Federal Reserve's target, it is less likely they will cut rates in 2025, as they prioritize controlling inflation over stimulating growth. A strong labor market and steady economic growth could deter the Fed from implementing rate cuts, as they might not see a need to stimulate an already healthy economy.
(1 votes)

The Federal Reserve's cautious approach to rate adjustments suggests they might hold rates steady unless there is a significant economic downturn. In the absence of unexpected economic shocks, the Fed might continue with stable rates to maintain economic stability.
(1 votes)

Market expectations can influence Fed decisions, and if there is a consensus that no cuts are needed, the Fed may align with these expectations unless there are compelling reasons to do otherwise. Historically, the Fed tends to be conservative in altering rates, especially if they have recently increased them, preferring to observe the long-term effects before considering cuts.
(1 votes)

A real-time time series chart shows the market's probability estimates for each outcome

Vote Option

After selecting both an outcome and a supporting reason, users can submit their vote by clicking the "Vote" button.

Selected: Yes
Chance: 16.5%

Figure 9: Detailed view of a prediction event. Users select an outcome and reason, then submit their vote.