

Can Large Language Models Uncover the Structure of Social Opinions?

Chongrui Ye*, Yuxiang Liu*, Haofei Yu, Jiayu Liu, Jiaxuan You
University of Illinois Urbana-Champaign

Correspondence: jiaxuan@illinois.edu

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 continuously evolving benchmark derived from multiple live prediction markets, designed to evaluate whether large language models (LLMs) can uncover the latent structure within evolving social opinions. OpinionBench employs a scalable and automated graph construction and labeling pipeline, which assigns edge labels 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 recover 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 collective expectations about uncertain future events—for example, presidential election outcomes, economic trends, or technological breakthroughs. These collective expectations emerge from the aggregation of individual beliefs, which are shaped by individuals’ education, experiences, and social context. These opinions reflect how groups collectively anticipate how the world will unfold and provide a basis for studying collective reasoning and societal decision-making. Social opinions exhibit two key structural features:

*Equal contribution. Order determined by coin flip; both authors reserve the right to list themselves as first author.

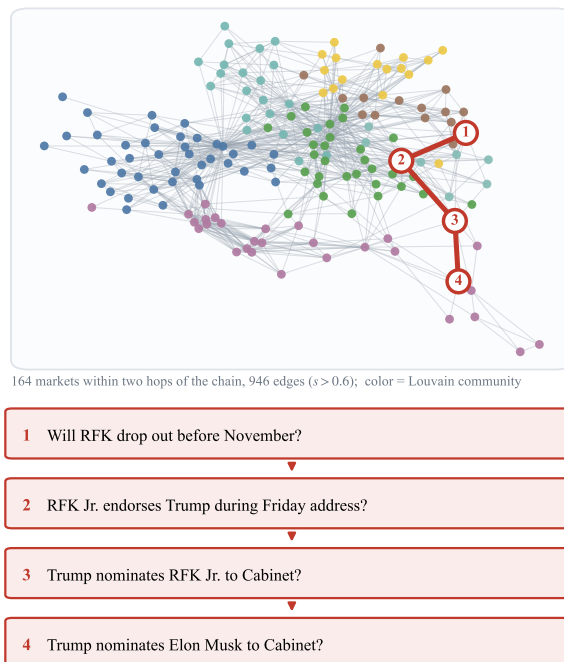


Figure 1: **A region of the opinion graph built from Polymarket data.** Nodes are social opinions, edges are ground-truth relations scoring above 0.6, and color marks a Louvain community. The four numbered markets form a chain whose two endpoints are not adjacent: they are related only through the intermediate events.

(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).

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. (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 latent 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 (Yu et al., 2026), a dataset compiled from Polymarket¹, a decentralized forecasting platform. To ensure our findings are robust to platform-specific biases (e.g., user demographics), we further conduct a cross-platform validation on Kalshi², a US-regulated prediction market (see Appendix §I for details). In the constructed graph, 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. The links models identify rest largely on shared context rather than explicit causal dependency, so what we measure is closer to recognizing relatedness than to causal inference; we analyze this in Section §9.

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 LENS (Xu et al., 2025) 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 prediction-market data across multiple platforms, 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 mod-

¹<https://polymarket.com/>

²<https://kalshi.com/>

eling 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 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 $v_e := (q_e, \{p_t^e\}_{t=1}^{T_e})$, 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 prediction markets. 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. We use *social opinion* throughout to refer to the crowd-aggregated expectation that this probability series tracks, rather than the individual sentiment or survey response the term often denotes. Individual trades may reflect hedging or liquidity needs rather than belief, but in actively traded mar-

kets the aggregated probability still serves as a collective estimate of how likely the event is. The volume and volatility filters in Section §4.1 are intended to retain only the markets where this holds.

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 $s : \mathcal{P} \rightarrow \mathbb{R}$, 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,

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

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 our primary dataset based on SWM (Yu et al., 2026), which is derived from Polymarket. To control for potential platform selection bias, we additionally curate a parallel dataset from Kalshi using the same construction pipeline (see Appendix §I). 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

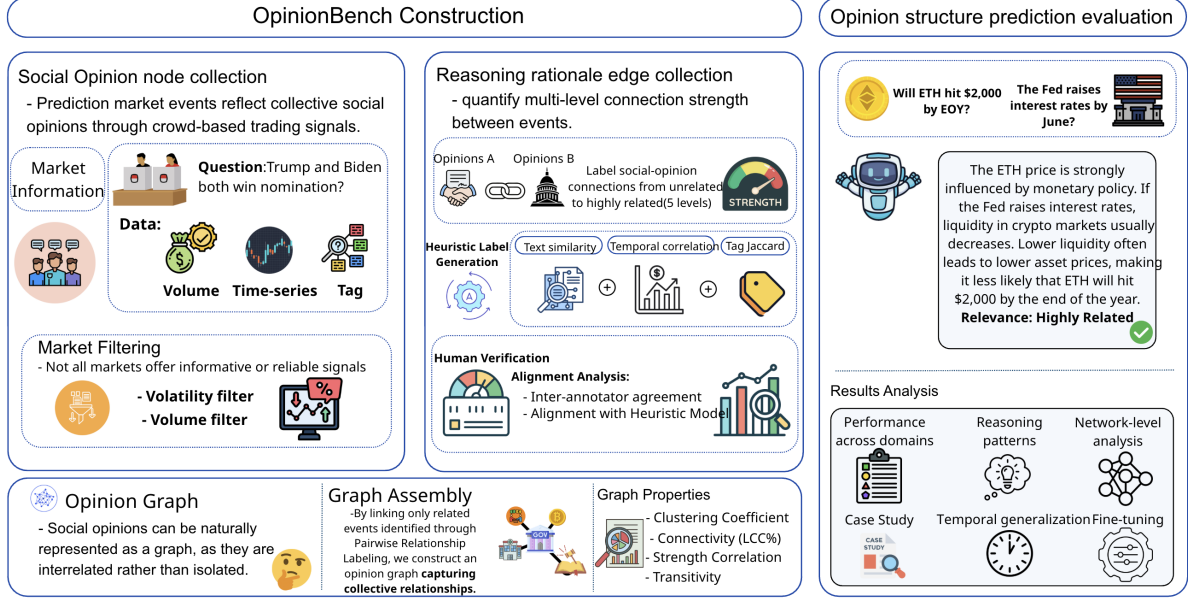


Figure 2: **Overview of the OPINIONBENCH pipeline.** (a) Social opinion node collection from prediction markets with market filtering for reliability. (b) Reasoning rationale edge collection quantifies multi-level connections (5 levels) via semantic, temporal, and tag similarities, verified by human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

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 and platform. 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 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 prediction market 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 $S(A, B) = \sum_{i=1}^4 w_i s_i(A, B)$, with weights tuned on a development set to align with human relevance judgments ($w_1=0.269$, $w_2=0.365$, $w_3=0.123$, $w_4=0.244$; see Appendix §C.2). We keep the combination linear so

that every edge label can be traced back to the features that produced it, and so that a feature can be removed and the labels rebuilt. Since s_4 is computed from the same event descriptions that models observe at inference time, we rebuild the ground truth with w_4 set to zero and re-evaluate; models retain their advantage over the baselines under the resulting labels (Appendix §H). 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 §L.

5 Predicting the Opinion Graph

We use LLMs to discover the opinion graph by jointly predicting a *rationale* and a *score* for each candidate edge. Given a pair of social opinions $(v_i, v_j) \in \mathcal{P}$, the model observes only their event descriptions q_i and q_j , and generates (i) a textual rationale explaining why the two opinions may be related, and (ii) a scalar score indicating the strength of this relationship. Formally, the LLM implements a mapping

$$\Phi_{\text{LLM}} : \mathcal{P} \rightarrow \mathcal{R} \times \mathbb{R}, \quad \Phi_{\text{LLM}}(v_i, v_j) = (r_{i,j}, s_{i,j}),$$

where $r_{i,j} \in \mathcal{R}$ denotes the generated textual rationale, and a score $s_{i,j} \in \mathbb{R}$ reflecting the predicted strength of their relationship. The predicted

opinion graph is reconstructed by thresholding the scores: edges with $s_{i,j} \geq \tau$ are retained to form $\hat{\mathcal{G}} = (\mathcal{V}, \hat{\mathcal{E}}_{\mathcal{G}})$ following Eq. 1. Unless otherwise stated, we set $\tau = 2$, corresponding to retaining edges with at least weak relevance. This joint rationale–score prediction enables interpretable edge inference and scalable graph recovery from pairwise evaluations.

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 model³ (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-4o (OpenAI et al., 2024), GPT-o3-mini (OpenAI, 2025), Qwen2 model family (Yang et al., 2024), LLaMA model family (Grattafiori et al., 2024) and DeepSeek-R1 (Guo et al., 2025) using a rationale-based classification setup. Given two social opinion titles and 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.

Evaluation metrics We evaluate model performance at both the *edge* and *graph* levels on the primary benchmark, and report additional cross-platform results in Appendix §I. Edge-level evaluation uses standard classification and ordinal agreement metrics (MSE, MAE, Accuracy, Macro-F1, and QWK), while graph-level evaluation assesses global structural fidelity via connectivity, clustering, degree alignment, and transitivity. Formal definitions of all metrics are provided in Appendix §D.

7 Experimental Results

LLMs generally achieve strong edge prediction performance across domains In Figure 3, we evaluate model performance across four domains, using both classification (Accuracy, F1, QWK) and regression (MSE, MAE) metrics. Overall, GPT-4o

³<https://huggingface.co/cross-encoder/nli-deberta-v3-base>

	Politics					Election					Crypto					Sports				
	MSE	MAE	Acc	F1	QWK	MSE	MAE	Acc	F1	QWK	MSE	MAE	Acc	F1	QWK	MSE	MAE	Acc	F1	QWK
DeepSeek-V3	.025	.129	.438	.310	.410	.033	.146	.413	.295	.409	.024	.123	.497	.343	.519	.088	.243	.168	.156	.344
GPT-4o	.023	.126	.432	.298	.409	.022	.111	.557	.294	.352	.025	.127	.468	.358	.523	.104	.256	.175	.143	.242
GPT-4o + CoT	.025	.121	.512	.262	.284	.035	.149	.403	.310	.415	.026	.127	.443	.269	.445	.074	.221	.227	.189	.368
o3-mini + CoT	.019	.115	.456	.241	.324	.023	.119	.545	.250	.418	.054	.164	.463	.353	.501	.120	.265	.218	.192	.362
o3-mini	.025	.132	.397	.272	.342	.028	.134	.455	.209	.275	.041	.155	.428	.336	.465	.093	.231	.284	.238	.480
DeepSeek-V3 + CoT	.024	.129	.401	.275	.370	.031	.145	.384	.293	.420	.029	.140	.396	.324	.489	.091	.248	.150	.131	.303
Llama3-70B	.038	.153	.385	.254	.408	.060	.197	.310	.247	.360	.042	.164	.383	.335	.500	.077	.231	.181	.179	.440
Qwen2-72B	.031	.143	.380	.229	.353	.043	.170	.323	.234	.413	.034	.149	.407	.307	.402	.098	.242	.200	.175	.248
Llama3-70B + CoT	.033	.152	.340	.259	.389	.047	.183	.266	.212	.340	.036	.161	.330	.328	.473	.084	.240	.163	.152	.392
Qwen2-72B + CoT	.034	.153	.320	.229	.275	.038	.164	.320	.274	.410	.039	.168	.287	.238	.355	.101	.248	.193	.163	.250
Qwen1.5-4B (LoRA)	.035	.130	.436	.181	.175	.068	.202	.306	.240	.387	.076	.194	.359	.333	.452	.129	.279	.285	.183	.147
DeepSeek-R1 + CoT	.051	.200	.196	.159	.116	.022	.119	.464	.189	.172	.035	.170	.232	.142	.071	.044	.163	.397	.262	.345
DeepSeek-R1	.083	.260	.094	.081	.079	.044	.183	.221	.114	.127	.044	.183	.221	.114	.127	.026	.131	.441	.460	.333
Time Overlap	.046	.169	.344	.191	.112	.064	.206	.257	.191	.138	.078	.227	.218	.134	.143	.088	.238	.216	.206	.419
Qwen1.5-4B (0-shot)	.084	.211	.296	.104	.031	.210	.365	.222	.116	.046	.120	.269	.243	.169	.180	.146	.303	.200	.129	.002

Figure 3: **Performance on opinion graph edge scoring across four domains: Politics, Election, Crypto, and Sports. Only the strongest baseline is included for comparison.** Each cell reports the raw metric value; shading encodes relative standing, computed per metric across all four domains so that colors are comparable between domains. Error metrics (MSE, MAE) are inverted before shading, so darker always indicates better performance. Models are sorted by their average standing across all domains and metrics.

+ 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.

The advantage does not come from textual similarity alone Because the heuristic score used to build edge labels includes a textual similarity term, part of the target signal is computable from the event descriptions themselves. We therefore rebuild the ground truth with that term removed and re-evaluate under otherwise identical settings. Absolute scores fall, but the margin over the strongest heuristic baseline stays wide in Politics, Election, and Crypto: GPT-4o still leads time overlap by 22.0 accuracy points in Crypto and by 0.245 QWK in Election. Full results are reported in Appendix §H.

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

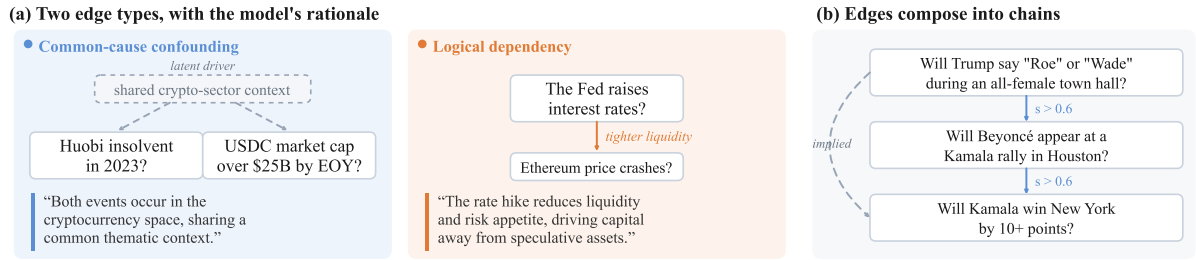


Figure 4: **Case study for edges between social opinions and the path they form.** (a) One edge of each reasoning type in Figure 5, with GPT-4o+CoT’s rationale. The dashed node is an unobserved shared driver, not a market. (b) A chain of consecutive high-relevance pairs (score > 0.6); the dashed edge is unscored and shows the transitivity discussed in Section §9.

Domain	LCC	Clustering	Corr.	Trans.
<i>LLM prediction</i>				
Politics	92.6→82.8	0.088→0.087	0.983	0.780*
Election	98.5→99.3	0.257→0.276	0.991	0.908*
Crypto	99.4→97.8	0.131→0.104	0.970	0.878*
Sports	91.6→45.0	0.077→0.034	0.954	0.900*
<i>Time overlap</i>				
Politics	100.0	0.702	0.749	0.693
Election	100.0	0.730	0.776	0.761
Crypto	92.3	0.525	0.547	0.547
Sports	33.3	0.833	0.671	1.000

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. $X \rightarrow Y$ represents the ground-truth one $\rightarrow$ the predicted one. A small difference between X and Y indicates the high fidelity for prediction. The lower block reports the same metrics for the strongest heuristic baseline, computed against the same ground-truth graphs. * indicates $p < 0.05$.

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 ($\approx 41\%$). *Politics* shows moderate degradation (83% vs. 93% GT connectivity), while *Sports* has lower connectivity (45%) but exceptionally high transitivity (90%), 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. The time overlap baseline does not reach this fidelity: it links nearly every pair of temporally co-

occurring events, which pushes clustering far above ground truth (0.53–0.83 against 0.077–0.257) and leaves edge weights only loosely aligned (0.55–0.78, against above 0.95 for the model), so its full connectivity in *Politics* and *Election* reflects density rather than structural agreement. We observe consistent qualitative trends when evaluating the same models on an independent prediction-market platform (Kalshi); full cross-platform results are reported in Appendix §I.

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 §F.

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 latent 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 (Q1) and investigate the reasoning strategies models use to make predictions (Q2). We then assess whether scaling and fine-tuning smaller models can improve efficiency without sacrificing accuracy (Q3), and evaluate the role of knowledge recency in generalizing to unseen events (Q4).

Q1: 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. Sports is better read as a boundary case than as a failure mode: the opinion graph formulation applies where events share institutional, temporal, or thematic structure, and we do not claim it extends to domains where they do not.

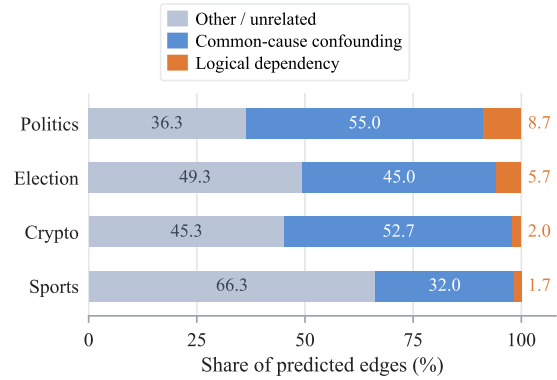


Figure 5: **Distribution of reasoning types across domains.** Each edge was labeled based on the explanation produced by the model. A majority of predictions are based on shared context (*common-cause confounding*) or links the model could not attribute to either category (*other/unrelated*), while only a small portion—1.7% to 8.7% depending on the domain—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.

Q2: 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*.

Q3: 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.

Q4: Does the knowledge cutoff matter?

To investigate reliance on pretraining knowledge, we compare performance on event pairs before and after the model’s knowledge cutoff. As shown in Table 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. This finding underscores the necessity of OPINION-BENCH’s design as an evolving benchmark. Unlike static datasets, its automated pipeline enables the systematic monitoring of this “temporal decay,” allowing researchers to quantitatively track how model performance degrades as real-world events drift further from the pretraining cutoff.

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. Importantly, we find that these conclusions hold consistently across prediction markets with different designs and participant populations, as validated through cross-platform

evaluation on Kalshi. Looking ahead, our framework can motivate future work for better studies on opinion dynamics and social simulation.

Limitations

We identify two primary limitations in our current evaluation framework. First, our ground-truth labels rely on a heuristic score $S(A, B)$ combining temporal synchrony, textual similarity, topical metadata similarity, and time alignment (Section §4.2). While this improves over pure correlation, it may still assign high scores to spurious pairs exhibiting coincident volatility. Consequently, these false positives can penalize models that correctly reject superficial links.

Second, our approach focuses on pairwise event relationships with overlapping active periods to capture dynamic co-movement. This design inherently limits the evaluation of delayed or sequential causal chains where events do not overlap in time, and does not explicitly model higher-order dependencies among multiple events. Although the construction pipeline is not specific to prediction markets and could in principle be applied to other sources with time-varying measures of social opinions, its validity in non-market settings remains to be established. Future work could address these gaps by exploring graph-based inference methods, lag-aware correlation measures, and broader geographic and non-market coverage.

Ethical Statement

This work analyzes aggregated, public event data from Polymarket and Kalshi. Because we utilize only market-level data without user-identifiable information, privacy risks are minimal. However, our dataset inherently reflects platform coverage biases, overrepresenting domains like politics and sports, which we encourage future work to diversify. Furthermore, our framework evaluates LLMs strictly on their reasoning regarding event relationships, not as outcome forecasters or decision-making systems. Since LLMs may exhibit unintended biases in socially charged topics, we strongly caution against deploying these models as authoritative predictors or decision-making tools in high-stakes financial or political environments.

References

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.
- DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, and 181 others. 2025. [Deepseek-v3 technical report](#). *Preprint*, arXiv:2412.19437.
- Songgaogun 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 Rzymiski. 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.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. [The llama 3 herd of models](#). *Preprint*, arXiv:2407.21783.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, and 175 others. 2025. [Deepseek-r1 incentivizes reasoning in llms through reinforcement learning](#). *Nature*, 645(8081):633–638.
- 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. What-sup: An event resolution approach for co-occurring events in social media. *Information Sciences*, 625:553–577.
- Seyed Mehran Kazemi, Rishab Goel, Kshitij Jain, Ivan Kobayez, 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 Adebisi. 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.
- OpenAI. 2025. [Openai o3-mini system card](#).
- OpenAI, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex

- Chow, Alex Kirillov, Alex Nichol, and 400 others. 2024. [Gpt-4o system card](#). *Preprint*, arXiv:2410.21276.
- 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 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](#).
- 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, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. [Chain-of-thought prompting elicits reasoning in large language models](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.
- Aaron Wheeler and Jeffrey D. Varner. 2024. [Marketsgpt: 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. 2025. [Lens: Large pre-trained transformer for exploring financial time series regularities](#). *Preprint*, arXiv:2408.10111.
- An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, and 43 others. 2024. [Qwen2 technical report](#). *Preprint*, arXiv:2407.10671.
- Haofei Yu, Yining Zhao, Guanyu Lin, and Jiaxuan You. 2026. Building social world models with large language models. *arXiv preprint arXiv:2606.11482*.
- 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.
 - **Model Usage:** Our study relies on querying several pretrained language models. We use **GPT-4o** and **GPT-o3-mini** via the OpenAI API,⁴ which are proprietary models licensed by OpenAI. We also evaluate open-weight models including **Meta-Llama-3 70B** (Grattafiori et al., 2024), **DeepSeek-V3** (DeepSeek-AI et al., 2025), **DeepSeek-R1** (Guo et al., 2025), and **Qwen2** (Yang et al., 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,⁵ 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`⁶ 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 (Yu et al., 2026). We also use a subset of Kalshi data. The final dataset will be released under the MIT License for academic use.

- **Source and Licensing:**
- **Use Consistency:** Our data usage is consistent

⁴<https://platform.openai.com/docs/models/gpt-4o>

⁵<https://huggingface.co/Qwen/Qwen1.5-4B>

⁶<https://huggingface.co/cross-encoder/nli-deberta-v3-base>

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 §L 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

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

i.e., at least α fraction of the windows have the standard deviation of the daily logit returns above γ . In all experiments, we fix $w = 7$, $\gamma = 0.0029$, and $\alpha = 0.6$. These values are selected based on empirical statistics of logit-return volatility and are kept fixed across all domains and platforms (Polymarket and Kalshi) to ensure consistency and avoid domain-specific tuning.

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:

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

Feature2: Tag Jaccard similarity. To estimate topical overlap, we use the Jaccard index over tag sets from prediction market 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 social opinion shifts in the two events occur in time. We convert this to a soft similarity score using a monotonic inverse transformation:

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

Feature 4: Textual similarity. We embed event descriptions using sentence-transformer models and compute cosine similarity between the resulting embeddings. This feature captures semantic proximity at both lexical and conceptual levels, complementing tag-based similarity with more nuanced language modeling:

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

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. The final normalized weights are: $w_{\text{sync}}=0.2688$, $w_{\text{jacc}}=0.3652$, $w_{\text{time}}=0.1228$, and $w_{\text{text}}=0.2435$.

D Evaluation Metrics

We report evaluation metrics at both the *edge* and *graph* levels to assess model performance from complementary perspectives. Edge-level metrics measure the accuracy of pairwise opinion correlation prediction, while graph-level metrics evaluate how well the reconstructed opinion graph preserves global structural properties of the ground-truth graph.

D.1 Edge-Level Metrics

Given a set of candidate opinion pairs with ground-truth ordinal labels in $\{1, 2, 3, 4, 5\}$ and corresponding model predictions, we report the following metrics:

Mean Squared Error (MSE). We map each ordinal label to its numeric value and compute

$$\text{MSE} = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2,$$

where y_i and $\hat{y}_i$ denote the ground-truth and predicted labels, respectively.

Mean Absolute Error (MAE). MAE measures the average absolute deviation between predictions and ground truth:

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}_i|.$$

Accuracy and Macro-F1. Accuracy reports the fraction of exactly matched labels. Macro-F1 computes the unweighted average of F1 scores across all five ordinal classes, treating each class equally regardless of frequency.

Quadratic Weighted Kappa (QWK). QWK measures ordinal agreement between predicted and ground-truth labels while penalizing larger disagreements more heavily. It is defined as

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

where O and E denote the observed and expected rating matrices, and $W_{ij} = \frac{(i-j)^2}{(K-1)^2}$ is the quadratic weight matrix for $K = 5$ ordinal classes. QWK ranges from -1 to 1 , with higher values indicating stronger ordinal agreement.

D.2 Graph-Level Metrics

To evaluate whether pairwise predictions recover the global structure of the opinion graph, we construct predicted graphs by thresholding edge scores and compare them with ground-truth graphs using the following metrics:

Largest Connected Component (LCC). LCC measures the fraction of nodes in the largest connected component:

$$\text{LCC} = \frac{|V_{\max}|}{|V|},$$

where $V_{\max}$ denotes the node set of the largest connected component.

Clustering Coefficient. The clustering coefficient captures local transitivity and triangle density:

$$C = \frac{1}{|V|} \sum_{i \in V} \frac{2T_i}{k_i(k_i - 1)},$$

where T_i is the number of triangles involving node i and k_i is its degree.

Strength Correlation. We compute the Pearson correlation between node strengths (weighted degrees) in the predicted and ground-truth graphs:

$$\rho = \text{corr}(s_i^{\text{GT}}, s_i^{\text{Pred}}).$$

Transitivity. Transitivity measures the global prevalence of closed triads:

$$\text{Trans} = \frac{3 \times \# \text{ triangles}}{\# \text{ connected triples}}.$$

Together, these metrics assess both local predictive accuracy and the extent to which LLM-based edge predictions preserve higher-order structural properties of the opinion graph.

E LLM Prompt

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

F Case Study Examples

F.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."

F.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?

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.

G Polymarket Full Results

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

H Ablation on Textual Similarity in Ground-Truth Construction

The heuristic score $S(A, B)$ includes a textual similarity term s_4 , so part of the signal that models are asked to recover is computed from the same event descriptions they observe at inference time. To separate this component from the temporal and meta-

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 <i>exactly one</i> 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 <i>not</i> 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 <i>exactly one</i> 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 <i>ONLY</i> one valid JSON object like: {"label": "0.4-0.6"}. Do <i>not</i> 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, <i>then</i> output the JSON object on the next line.

Table 3: Comparison of Original vs. CoT prompt templates

data signals, we build an alternative ground truth GT_{text} by setting $w_{\text{text}} = 0$ and renormalizing the remaining three weights. Node selection, candidate pairs, discretization boundaries, and prompts are all held fixed, so the two conditions differ only in how edge labels are defined.

Table 8 reports QWK and accuracy under both ground truths for GPT-4o, DeepSeek-V3, and the strongest heuristic baseline (time overlap). Removing s_4 lowers both LLMs in every domain, which is expected since the remaining signal is sparser and less recoverable from event text. The time overlap baseline is largely unaffected and improves slightly in three domains, consistent with it carrying no semantic information to begin with. The ordering between models and baseline is nevertheless preserved: in Politics, Election, and Crypto, both LLMs stay well above time overlap under GT_{text} , with GPT-4o exceeding it by 22.0 accuracy points in Crypto and by 0.245 QWK in Election. Model performance therefore does not reduce to recovering the embedding similarity used to construct the

labels.

Sports again separates from the other domains. Time overlap outperforms GPT-4o under both ground truths, matching our observation in Section §9 that sports events are episodic and actor-specific, so temporal co-occurrence carries more information than semantic relatedness. That this holds under both label definitions suggests it reflects the structure of the domain rather than an artifact of the scoring function.

I Cross-Platform Evaluation on Kalshi

Cross-platform robustness. To address concerns about potential single-source bias, we additionally evaluate all models on social opinion graphs constructed from *Kalshi*, a centralized, regulated prediction-market platform with different market design, participant composition, and resolution mechanisms from Polymarket. Tables 9–11 report full results across Politics, Election, and Crypto domains.

Despite these platform-level differences, we ob-

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 (Polymarket).

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
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 (Polymarket).

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
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 (Polymarket).

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
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 (Polymarket).

Domain	Time Overlap		GPT-4o		DeepSeek-V3	
	GT _{full}	GT _{¬text}	GT _{full}	GT _{¬text}	GT _{full}	GT _{¬text}
<i>QWK</i>						
Politics	0.1113	0.1501	0.4074	0.3427	0.3325	0.2435
Election	0.1172	0.1288	0.4196	0.3736	0.3933	0.3451
Crypto	0.1530	0.1404	0.5745	0.4843	0.4839	0.3971
Sports	0.3889	0.4159	0.3754	0.3643	0.5538	0.4637
<i>Accuracy (%)</i>						
Politics	31.67	34.67	40.67	38.00	43.67	38.33
Election	21.33	28.33	48.67	38.33	35.00	33.00
Crypto	18.33	21.33	46.67	43.33	48.33	40.67
Sports	20.67	22.33	21.33	18.67	20.67	18.33

Table 8: **Effect of removing textual similarity from ground-truth construction.** GT_{full} uses all four features; GT_{¬text} sets $w_{\text{text}} = 0$ and renormalizes the remaining weights. All other settings are identical across the two conditions, so the columns within each model are directly comparable.

serve consistent qualitative trends with the main results reported in the paper. LLM-based approaches continue to substantially outperform heuristic and neural baselines across all domains, and the relative ordering of models largely remains stable. In particular, stronger LLMs (e.g., GPT-4o, DeepSeek-V3, Qwen2-72B) consistently achieve higher QWK and Macro-F1 scores, indicating robust ordinal agreement in opinion correlation prediction.

These findings suggest that the opinion graph structure captured by OPINIONBENCH reflects general patterns of collective belief co-evolution rather than artifacts of a specific prediction-market platform, strengthening the external validity of our conclusions.

J 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
heuristic (vol. max→min)	0.0819	0.2491	0.1650	0.0570	0.0160
heuristic (vol. sim.)	0.0784	0.2358	0.2100	0.1099	0.0613
heuristic (time overlap)	0.0786	0.2263	0.2700	0.2260	0.2990
cross-encoder (nli-deberta-v3-base)	0.3262	0.5528	0.3150	0.1255	0.3221
GPT-4o	0.0428	0.1760	0.2800	0.1724	0.3684
GPT-4o + CoT	0.0373	0.1613	0.3300	0.2153	0.3584
GPT-o3-mini	0.0413	0.1640	0.3700	0.2694	0.4433
GPT-o3-mini + CoT	0.0644	0.2051	0.2950	0.2395	0.4250
Meta-Llama3-70B	0.0444	0.1760	0.3100	0.2165	0.3109
Meta-Llama3-70B + CoT	0.0323	0.1463	0.4200	0.2564	0.3669
DeepSeek-V3	0.0258	0.1292	0.4850	0.3138	0.4704
DeepSeek-V3 + CoT	0.0303	0.1415	0.4250	0.2632	0.4188
DeepSeek-R1	0.0542	0.1956	0.2750	0.1453	0.0871
DeepSeek-R1 + CoT	0.0344	0.1514	0.3950	0.2212	0.2391
Qwen2-72B	0.0433	0.1752	0.2950	0.2037	0.3346
Qwen2-72B + CoT	0.0300	0.1374	0.4650	0.3000	0.3426

Table 9: Performance on Politics domain (Kalshi).

Method	MSE	MAE	Accuracy	Macro-F1	QWK
heuristic (vol. max→min)	0.1095	0.3035	0.0350	0.0220	-0.0280
heuristic (vol. sim.)	0.0503	0.1855	0.2800	0.1228	0.1080
heuristic (time overlap)	0.0649	0.2179	0.2100	0.1632	0.2459
cross-encoder (nli-deberta-v3-base)	0.2679	0.5033	0.3550	0.1404	0.2429
GPT-4o	0.0761	0.2514	0.1200	0.0783	0.2235
GPT-4o + CoT	0.0702	0.2398	0.1100	0.1000	0.2145
GPT-o3-mini	0.0538	0.1823	0.3500	0.2013	0.3949
GPT-o3-mini + CoT	0.0688	0.2194	0.2400	0.1314	0.2757
Meta-Llama3-70B	0.0765	0.2526	0.1000	0.1167	0.2111
Meta-Llama3-70B + CoT	0.0360	0.1568	0.3500	0.2158	0.3793
DeepSeek-V3	0.0548	0.2069	0.2100	0.1257	0.2815
DeepSeek-V3 + CoT	0.0649	0.2301	0.1350	0.0893	0.2704
DeepSeek-R1	0.0247	0.1326	0.4000	0.2074	0.2335
DeepSeek-R1 + CoT	0.0224	0.1228	0.4650	0.2292	0.2906
Qwen2-72B	0.0197	0.1119	0.5300	0.3715	0.5056
Qwen2-72B + CoT	0.0309	0.1444	0.4050	0.2162	0.3915

Table 10: Performance on Elections domain (Kalshi).

Method	MSE	MAE	Accuracy	Macro-F1	QWK
heuristic (vol. max→min)	0.1082	0.2977	0.0650	0.0870	0.0540
heuristic (vol. sim.)	0.0532	0.1869	0.3200	0.1214	0.0445
heuristic (time overlap)	0.1185	0.2938	0.1700	0.1118	-0.0701
cross-encoder (nli-deberta-v3-base)	0.2543	0.4811	0.2260	0.0241	0.1332
GPT-4o	0.0369	0.1617	0.3300	0.2742	0.4290
GPT-4o + CoT	0.0550	0.2006	0.2000	0.1111	0.3396
GPT-o3-mini	0.0390	0.1350	0.5650	0.2744	0.3934
GPT-o3-mini + CoT	0.0724	0.1931	0.4350	0.4184	0.4884
Meta-Llama3-70B	0.0341	0.1472	0.4050	0.3454	0.4890
Meta-Llama3-70B + CoT	0.0735	0.2079	0.3000	0.1500	0.4230
DeepSeek-V3	0.0192	0.1093	0.5500	0.3961	0.4794
DeepSeek-V3 + CoT	0.0232	0.1233	0.5000	0.2750	0.5098
DeepSeek-R1	0.0448	0.1680	0.3500	0.1686	0.1486
DeepSeek-R1 + CoT	0.0686	0.2043	0.3000	0.1875	0.1222
Qwen2-72B	0.0581	0.2068	0.1900	0.1826	0.1789
Qwen2-72B + CoT	0.0236	0.1208	0.5200	0.3647	0.2882

Table 11: Performance on Crypto domain (Kalshi).

Table 12: **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?

K 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 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.

L Human Evaluation of Heuristic Scoring

L.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 de-

scribed 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. Annotators shared the same rubric and calibration examples, but rated every pair independently and without access to the heuristic scores, the underlying time series, or any model prediction.

L.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 12 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 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 13, 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

Table 13: **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

reflects substantial agreement across annotators and confirms the reliability of the human labels as a benchmark for model alignment.

L.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 14) indicates strong alignment between the scoring function and human reasoning.

Table 14: **Model-human alignment.** Pearson correlation between the heuristic score and human annotations.

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

M Detailed Performance Degradation After Cutoff

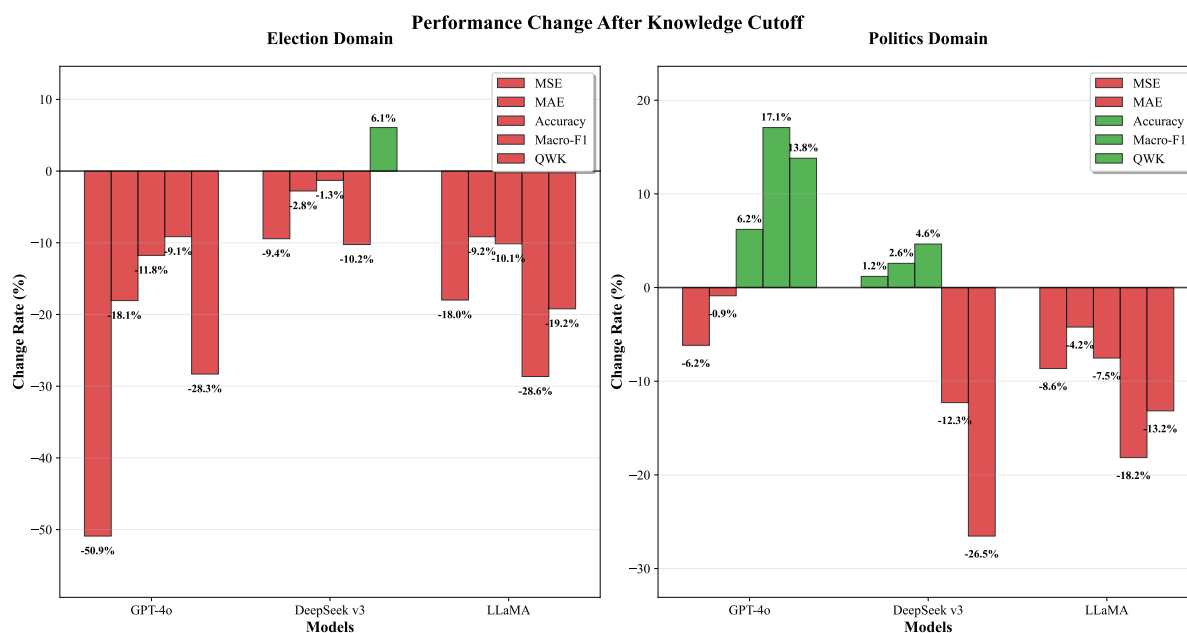


Figure 6: **Performance change after knowledge cutoff across domains and models.** Bars show the relative change in evaluation metrics on post-cutoff event pairs, compared to pre-cutoff ones. To ensure consistent directionality, we invert the sign for error metrics (MSE, MAE). Consequently, for all metrics in this figure, a positive bar indicates performance improvement, while a negative bar indicates degradation. GPT-4o shows a substantial decline across most metrics in the election domain, while performance remains more stable in the politics domain.